THE UNIVERSITY OF CHICAGO

SPARSITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
MICOL FEDERICA TRESOLDI

CHICAGO, ILLINOIS
AUGUST 2021

A mia mamma, ai miei nonni, a Nico e, ovviamente, a C.G.J..

Contents

List of Figures

List of Tables

Acknowledgements

The premise is that I do not really believe that few words of mine can actually convey what I would like to say. At the same time, I can see there is a value in acknowledging the people who have accompanied me in this long and quite incredible journey.

It goes without saying that the first person I would like to thank, for whom I have an infinite gratitude, is my advisor, Professor Peter McCullagh. Working with him has been the most precious gift I have been given by my Ph.D., here at the University of Chicago. Professor McCullagh's kindness and generosity are certainly out of the ordinary, going well beyond any usual dimension of time and ideas. His invaluable help and constant support have been absolutely crucial for me to arrive at this point. The love for critical thinking and the hunger for unexplored, and therefore exciting problems, are perhaps the two things I will bring with me forever, wherever I will go.

Besides my advisor, I would like to express my gratitude to Professor Rina Foygel Barber, Professor Veronika Ročková, and Professor Nicholas Polson, for serving on my thesis committee and being for me such bright examples of statisticians.

I am also very grateful to the Statistics Department, for giving me the opportunity to experience during these years, the extremely privileged, even if hard at most times, research life.

Outside of school, here in Chicago, I would like to thank M. and C. for their true friendship and concrete help. And, of course, R. and P. Watt, for their unique gentleness, their

continuous support, and above all, their time with me. Back home, I want to thank my mother, for letting me go this far away, and always being there for me. And there is no need to say how much grateful I am to N., without whom I simply could not have done anything.

Abstract

Despite its generic title, this thesis is about a specific notion of sparsity, the one introduced by McCullagh and Polson (2018) [51]. In that paper, the intuitive idea that sparsity, in a statistical framework, refers to those "phenomena that are mostly negligible or seldom appreciably large", has, for the first time, been given a mathematical definition. In studying this definition of statistical sparsity as a limiting property of a sequence of probability distributions, research has proceeded along different lines, which nevertheless intersect at all times. In all cases, our work has been driven by both theoretical and practical motivations.

The notion of negligibility, for instance, is developed from the necessity of describing the behavior of a sparse distribution in a region around zero, a necessity which is commonly encountered in applied work. At the same time, doing this in a mathematical way, allows us to define very clearly what is the perimeter within which this notion is informative, and can be used. Another main direction of research we pursue, aims at extending the definition of sparsity to distributions which are defined on $\mathbb{R}^d$, $d > 1$. Within this framework, we consider two scenarios: in the first one, the d-dimensional measure is a product of d one-dimensional sparse measures; in the second one, instead, the d-dimensional measure is rotationally invariant with respect to the inner product imposed on $\mathbb{R}^d$, and sparsity is driven by the radial component. For both cases, we develop some theory as well as present how this theory can be in fact applied in the context of various statistical problems.

Introduction

Despite its generic title, this thesis is about a specific notion of sparsity, the one introduced by McCullagh and Polson (2018) [51]. In that paper, the intuitive idea that sparsity, in a statistical framework, refers to those "phenomena that are mostly negligible or seldom appreciably large", has, for the first time, been given a mathematical definition.

In studying this definition of statistical sparsity as a limiting property of a sequence of probability distributions, research has proceeded along different lines, which, nevertheless, intersect at all times. Therefore, only for the sake of organization of exposition, we structure this thesis in three parts, but these should be considered separately only up to a limited extent.

The aim of this introduction is three-fold: first, to give the reader some background on the notion of sparsity introduced by McCullagh and Polson (2018) [51], this being the starting point of everything thereafter presented; second, to highlight some of the critical points that this initial definition presents under some circumstances, which in fact act as motivations for our work; last, to provide the reader with a roadmap, which, in our best intentions, would serve as a compass in the reading of the thesis, allowing the reader to see how certain ideas are shared by different parts of the work.

McC&P's definition of sparsity

McCullagh and Polson (2018) [51] (McC&P henceforth) first introduced the definition of statistical sparsity as a limiting property of a sequence of probability distributions, $\{P_\nu\}_\nu$, defined on the real line, and indexed by ν. This latter is called sparsity parameter since, as $\nu \to 0$, the sequence of measures P_ν is assumed to converge weakly to the Dirac delta measure at zero, δ_0. In other words, if $X \sim P_\nu$, then in the limit, X is equal to zero with probability one. However, sparsity as defined in McC&P, does not concern the limit itself, but rather it concerns the behavior of the sequence P_ν as the limit is approached. Put differently, for a family of distributions P_ν to have a sparse limit according to McC&P, one needs to be able to describe how P_ν approaches δ_0 in terms of expectations of certain functions, which will shortly be defined in a more precise way. For this description to be possible, two objects need to exist: a rate parameter ρ, which is a function of the sparsity parameter, and an exceedance measure H. These two objects characterize the behavior of the sequence P_ν in approaching the Dirac delta limit at zero, providing a description of how fast the probability concentrates around the origin and, at the same time, capturing its behavior in the tails. So, in a sense, such probabilistic definition of the intuitive idea of sparsity, endows one with an asymptotic approximation, which is driven by the sparsity parameter going to zero and does not have any requirement in terms of the sample size on hand. As an approximation device, sparsity can be used for inferential purposes in a sparse signal-plus-noise setting, to derive for instance, the sparse approximation to the marginal distribution of the observation and to the conditional distribution of the signal, given the observation.

All main sparsity models that have been proposed in the statistical literature on sparse-signal detection are indeed sparse according to the probabilistic limit definition of McC&P. In fact, for all of the following families of distributions, one can find their characteristic sparsity pair (ρ, H): the two-component atom-and-slab mixtures, very frequently used in the literature (see, for example, Jonhstone and Silverman, 2004 [45] and Efron and Tibshirani, 2001 [30]); the spike-and-slab mixtures proposed by George and McCulloch (1993) [40] and

Ročková and George (2018) [60]; the low-index gamma model by Griffin and Brown (2013) [14], as well as all scale families with polynomial tails, such as the scaled Cauchy and scaled horseshoe (Carvalho et al., 2010 [16]).

As a matter of fact, looking at the sparsity pair (ρ, H) for each of these families, one realizes that, if for each sparse family P_ν, there is only one exceedance measure, the reverse is not true. In other words, different sparse families can have the same exceedance measure. This is the case, for instance, of the scaled Cauchy family and scaled horseshoe family; or of the spike-and-slab and atom-and-slab families, whenever they share the same slab distribution and the spike distribution converges to the Dirac measure fast enough. Therefore, besides serving as a classification tool which declares a family of distributions to be either sparse or not sparse, the sparsity definition also establishes equivalence relations among different families. This equivalence is to be interpreted in terms of the asymptotic approximations that one is able to derive exclusively based on the pair (ρ, H).

We now report the formal definitions of exceedance measure, Lévy integrable functions, and sparse sequence of distributions on the real line, which are given in McC&P, and to which we will very often refer throughout all work.

Definition 0.0.1. A nonnegative measure H on the real line excluding the origin is termed an exceedance measure if $\int_{\mathbb{R}\setminus\{0\}} \min(x^2, 1)\, H(dx) < \infty$. An exceedance measure is called a unit exceedance measure if $\int_{\mathbb{R}\setminus\{0\}} (1 - e^{-x^2/2})\, H(dx) = 1$.

Definition 0.0.2. The space $\mathcal{W}^{\#}$ of Lévy-integrable functions consists of bounded and continuous functions $w(x)$ on the real line such that $x^{-2}w(x)$ is also bounded and continuous. Lévy-integrability implies $\int_{\mathbb{R}\setminus\{0\}} w(x)\, H(dx) < \infty$ for every $w \in \mathcal{W}^{\#}$ and every exceedance measure H.

Definition 0.0.3. A sequence of probability distributions $\{P_\nu\}$ is said to have a sparse limit

with rate ρ_ν if there exists a unit exceedance measure H such that

$$\lim_{\nu \to 0} \rho_\nu^{-1} \int_{\mathbb{R}} w(x)\, P_{\nu,d}(dx) = \int_{\mathbb{R}\setminus\{0\}} w(x)\, H(dx)\,, \tag{1}$$

for every $w \in \mathcal{W}^{\#}$. Otherwise, if the limit is zero for every w, the sequence is said to be sparse with rate $o(\rho_\nu)$.

From Definition 0.0.3, sparsity can be used as a mathematical tool to derive approximations of certain functionals of P_ν, which are expectations of Lévy integrable functions, as the sparsity parameter goes to zero. Sometimes, we will refer to this as $\mathcal{W}^{\#}$-convergence.

Going back to the original intuitive idea of sparsity as a characteristic of phenomena that are rarely appreciable, one may want to translate this $\mathcal{W}^{\#}$-convergence definition into a threshold-exceedance definition. Given a positive threshold $\epsilon > 0$, the probability that the sparse signal, in absolute value, is above the threshold, being an integral of a discontinuous function, cannot be approximated directly from Definition 0.0.3. Nevertheless, the hard-threshold function $\chi_{(\epsilon,\infty)}(|x|)$ can be approximated with arbitrary accuracy by a sequence of soft-threshold functions, of the kind $w_\epsilon(x) = 1 - e^{-x^2/2\epsilon^2}$, which indeed belong to the class $\mathcal{W}^{\#}$. Therefore, applying (1) to $w_\epsilon(x)$, one has that

$$\int w_\epsilon(x)\, P_\nu(dx) = \rho \int w_\epsilon(x)\, H(dx) + o(\rho)\,,$$

which, in terms of the hard-threshold event $\epsilon^+ = \{|X| > \epsilon\}$, can be written as

$$P_\nu(\epsilon^+) = \rho H(\epsilon^+) + o(\rho)\,. \tag{2}$$

This last equation leads to interpreting the product ρH as the rarity of threshold exceedances, where ρ captures the velocity in approaching the zero limit, while H gauges the tail behavior

of the sparse measure, i.e., the exceedance probability. If H is atomless, then (1) implies (2), but (2) can hold even with some measure H, that is not a Lévy measure. An example of this case is the family $P_\nu(dx) = (1-\nu)\delta_0(dx) + \nu N(0, 1/\nu)$, which satisfies (2) with $\rho = \nu$ and $H(dx) = \delta_{\pm\infty}(dx)$, even if this latter is not a Lévy measure or a measure on $\mathbb{R}$. Moreover, (2) can hold with H being a Lévy measure, even if (1) does not hold. For instance, the family $P_\nu(dx) = (1-\nu)e^{-|x|/\sqrt{\nu}}/2\sqrt{\nu}\,dx + \nu/\pi(x^2+1)\,dx$ satisfies (2) with $\rho H(dx) = \nu/\pi(x^2+1)\,dx$, but, as $\nu \to 0$,

$$\begin{aligned}\int (1-e^{-x^2/2})P_\nu(dx) &\sim \int (1-e^{-x^2/2})\frac{e^{-|x|/\sqrt{\nu}}dx}{2\sqrt{\nu}} + \nu\int (1-e^{-x^2/2})\frac{dx}{\pi(x^2+1)} \\ &\sim \nu\cdot\int (1-e^{-x^2/2})\left(\frac{1}{x^2}\delta_0(dx) + \frac{dx}{\pi(x^2+1)}\right),\end{aligned}$$

and the measure appearing in the last integral is not an exceedance measure. However, in most cases (1) and (2) are equivalent, and this is true also for spike-and-slab measures as long as the variance of the spike distribution, as a function of ν, is of order greater than the sparsity rate.

Some examples

To give the reader an idea of how one can find the sparsity pair (ρ, H) for a given family P_ν, we now present a couple of examples, out of the many more that can be found in McC&P. We start with the scaled Cauchy family, whose density is $\nu/(\pi(x^2+\nu^2))$. Computing the ϵ-exceedance probability $P_\nu(\epsilon^+)$, one obtains $2/\pi(\pi/2 - \arctan(\epsilon/\nu))$, which, for ν going to zero, behaves like $2/\pi\cdot\nu/\epsilon$. Therefore, if $H(dx) = 1/\sqrt{2\pi}\,|x|^{-2}dx$, then $H(\epsilon^+) = \sqrt{2/\pi}\cdot 1/\epsilon$, so that (2) holds with $\rho = \sqrt{2/\pi}\,\nu$. Before presenting the next example, we make a couple of remarks: first, the constant $1/\sqrt{2\pi}$ is chosen to make H a unitary exceedance measure; second, one could have guessed the exceedance measure for scaled Cauchy by looking at the tail behavior of the unscaled Cauchy density $1/(\pi(x^2+1))$ as $x \to \infty$.

This is indeed a property of any sparse family whose sparsity is driven by its scale

parameter going to zero, so that $P_\nu(dx) = \nu^{-1}p(x/\nu)dx$. In this case, if there is a definite sparsity rate, i.e., P_ν does not have exponential tails, then the exceedance density is an inverse-power function $h(x) \propto x^{-\alpha-1}$ reflecting the tail behavior of $p(x)$ at infinity. For this reason, for $\alpha \in (0, 2)$, the class of unit inverse-power measures

$$H(dx) = \frac{\alpha 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)} \frac{dx}{|x|^{\alpha+1}}\,, \tag{3}$$

is the class of exceedance measures which we will mostly use in applications.

The second example we report is instead the atom-and-slab mixture, such as that proposed by Mitchell and Beauchamp (1988) [52] or that by Johnstone and Silverman (2004) [45], only to mention a few references. In all of these cases, $P_\nu(dx) = (1-\nu)\delta_0(dx)+\nu F(dx)$, where F is some probability distribution on $\mathbb{R}$, usually symmetric around zero. Once again from (2), it is easy to see that $\rho H(dx) = \nu F(dx)$, so that, for this family, the exceedance measure is finite and the sparsity rate is $\rho = \nu \int (1 - e^{-x^2/2})F(dx)$. Notice that, if we replace the Dirac delta measure with a spike distribution having variance of order greater than ν, such as $N(0, \nu^2)$ (George and McCulloch, 1993 [40]), or scaled Laplace $e^{-|x|/\nu}/(2\nu)\,dx$ (Ročková and George, 2018 [60]), then we still have the same pair $\rho H(dx) = \nu F(dx)$, so that all these families are first-order equivalent in the sparse limit.

Atom at zero and negligibility

The tail behavior determines the probability, under the sparse measure P_ν, of exceeding a given positive threshold, but it is not enough to tell us what probability mass, the measure P_ν gives to the atom at zero or how it is distributed near zero. In fact, consider two sparse families: the scaled Cauchy

$$P^1_\nu(dx) = \frac{\nu}{\pi(x^2+\nu^2)}\,dx\,,$$

and the mixture

$$P_\nu^2(dx) = (1-\sqrt{\nu})\delta_0(dx) + \sqrt{\nu}\,\frac{\sqrt{\nu}}{\pi(x^2+\nu)}\,dx\,.$$

These two sparse measures share the same first-order sparsity pair given by $\rho = \nu\sqrt{2/\pi}$ and $H(dx) = 1/\sqrt{2\pi}\,|x|^{-2}dx$. Therefore the sparse approximations to the expectation of any function $w \in \mathcal{W}^{\#}$ with respect to P_ν^1 and P_ν^2, are exactly the same, insofar they are solely based on the pair (ρ, H). Yet, the two measures give very different probability mass to the atom at zero, as $P_\nu^1(X=0) = 0$ while $P_\nu^2(X=0) = 1-\sqrt{\nu}$.

This fact, in turn, highlights that Definition 0.0.3 is not enough to directly approximate the probability of every event in the sigma algebra of $X \sim P_\nu$: for all those events A whose closure contains the atom at zero, we cannot write $\lim_{\nu\to 0}\rho^{-1}P_\nu(A) = H(A)$. Nevertheless, in most cases, we can still approximate $P_\nu(A)$ by some, more or less artificial, modification of the $\mathcal{W}^{\#}$-function approximating $\chi_A(x)$, so to be able to apply the $\mathcal{W}^{\#}$-integral definition 0.0.3.

This brings us to talk about one of the critical parts of the sparsity theory developed in McC&P, the non-identifiability of the atom at zero. In fact, in many statistical applications, it is often the case that the interest lies in establishing whether a signal is active or not, so that one can identify the presumably few active, out of a large number of signals. And, although in the literature, there is not a universal consensus nor a formal mathematical definition of what constitutes signal activity, fairly often, in both theoretical and applied work, the dichotomy of signal non-activity/activity refers to the events that the signal be zero or not zero (see for instance, Donoho *et al.*, 1992 [24] and Efron, 2007 [27]).

Yet, as illustrated above, the pair (ρ, H) is not sufficient to approximate the probability mass of P_ν at zero. Indeed, this zero-non-identifiability issue strictly relates to the perimeter within which the sparsity definition in (1) applies and can be used: the limit-approaching behavior of P_ν is described by looking at the expectations, with respect to P_ν, of bounded and continuous functions $w(x)$, for which the function $x^{-2}w(x)$ is also bounded and contin-

uous. Instead, the indicator function $\chi_{x=0}$ is a discontinuous function at the limit point and such discontinuity at the limit opens the door to different answers from the sparse measures, even when their limiting behavior in approaching the Dirac delta limit is the same.

The way we propose to circumvent the zero-non-identifiability issue is to take a slightly different perspective in looking at the problem, and adopt a strategy that is similar in spirit to the "limit-approaching" standpoint, from which the sparsity theory by McC&P has been formulated in first place. The idea is to look at the atom $\{0\}$ as the limit point of a sequence of intervals $[-\epsilon_\nu, \epsilon_\nu]$, where $\epsilon_\nu \to 0$ as $\nu \to 0$, and to describe, as the limit takes place, the approaching behavior of the moving sequence of measures P_ν over the moving region $[-\epsilon_\nu, \epsilon_\nu]$. So, instead of asking for the probability at the limit point $P_\nu(X = 0)$, which requires the expectation of a discontinuous function at zero, we ask for the probability that the signal is in a region converging to the limit point, $P_\nu(|X| \leq \epsilon_\nu)$, which by contrast, can be approximated by the expectation of a bounded and continuous function,

$$\int e^{-x^2/2\epsilon_\nu^2} \, P_\nu(dx) \, .$$

With some conditions on the speed of convergence to zero of the threshold sequence ϵ_ν, we can use the sparsity integral definition in (1) to obtain the sparse-negligibility approximation

$$\int e^{-x^2/2\epsilon_\nu^2} \, P_\nu(dx) \sim 1 - \rho \int (1 - e^{-x^2/2\epsilon_\nu^2}) \, H(dx) \, ,$$

where this approximation holds with an error of order $\rho \int (1 - e^{-x^2/2\epsilon_\nu^2}) \, H(dx)$.

All this leads us to introduce a mathematical definition of signal negligibility. In what follows, we use the notation $F(w_z) = \int (1 - e^{-x^2/2z^2}) \, F(dx)$, where F is some non-zero measure.

Definition 0.0.4. Let $\{P_\nu\}_\nu$ be a sparse sequence of symmetric distributions on $\mathbb{R}$, and let

$\{\epsilon_\nu\}_\nu$ be a sequence of strictly positive thresholds, $\epsilon_\nu > 0$. We say that ϵ_ν is a negligibility sequence for P_ν if, as $\nu \to 0$,

1. $\epsilon_\nu \to 0$,

2. $P_\nu(w_{\epsilon_\nu}) \to 0$,

3. $P_\nu(w_{\epsilon_\nu}) = \rho H(w_{\epsilon_\nu}) + o\left(\rho H(w_{\epsilon_\nu})\right)$.

Given $X \sim P_\nu$, we say that X is negligible if $|X| \leq \epsilon_\nu$.

With this limiting definition of signal negligibility, we can derive another integral approximation for the sparse measure P_ν, alternative to (1), which is valid up to an error larger than the usual $o(\rho)$, but has a Dirac delta measure component in it. This last component makes in fact negligibility equivalent to being zero in terms of integrals of bounded and continuous functions.

For the case when H is the inverse-power measure in (3) proportional to $|x|^{-\alpha-1}dx$, this integral approximation is

$$\int w(x)\, P_\nu(dx) = (1 - \rho\epsilon_\nu^{-\alpha}) \int w(x)\, \delta_0(dx) + \rho\epsilon_\nu^{-\alpha} \int w(x)\, \tilde{H}(dx) + o(\rho\epsilon_\nu^{-\alpha})\,, \tag{4}$$

where w is a bounded and continuous function, while $\tilde{H}$ is the weighted exceedance measure proportional to $(1 - e^{-x^2/2\epsilon_\nu^2})|x|^{-\alpha-1}dx$.

The integral approximation we develop within the negligibility-sparsity theory can be naturally seen as an alternative to many other Bayesian approaches developed for the so called two-groups model (Efron, 2007 [27]). Yet, despite the general structure resemblance, there are two main differences. The first most obvious discrepancy concerns the target event, which leads to the mixture distribution for the signal. In our framework, this event is the

signal negligibility as defined in Definition 0.0.4, whereas, in the two-group model, is the event of the signal being absolutely zero. The second difference follows from the first one to the extent that we obtain an atomic mixture (i.e., a Dirac delta measure component) for the signal distribution, only as an asymptotic approximation, driven by the sparsity limit, and true solely in terms of integrals of bounded and continuous functions.

Signal plus noise

The signal-plus-noise model is perhaps the most basic statistical setting in which sparsity can be seen arise naturally. The model considered for instance by Johnstone and Silverman (2004) [45], assumes that the observation Y is the sum of two independent components: a sparse signal μ and a standard Gaussian noise η,

$$Y = \mu + \eta\,. \tag{5}$$

In the statistical sparsity framework of McC&P, μ has a sparse distribution P_ν, symmetric around the origin, with sparsity rate ρ and unit exceedance measure H. Then, for each $y \in \mathbb{R}$, the marginal of Y at y, being a functional of the symmetric P_ν, can be approximated in the following way

$$\begin{aligned} m_\nu(y) &= \phi(y) \int e^{yx} e^{-x^2/2}\, P_\nu(dx) \\ &= \phi(y) \left(\int (\cosh(yx) - 1) e^{-x^2/2}\, P_\nu(dx) + 1 - \int (1 - e^{-x^2/2})\, P_\nu(dx) \right) \\ &= \phi(y) \left(\rho \int (\cosh(yx) - 1) e^{-x^2/2}\, H(dx) + 1 - \rho \int (1 - e^{-x^2/2})\, H(dx) \right) + o(\rho) \\ &= \phi(y) \left(\rho \int (\cosh(yx) - 1) e^{-x^2/2}\, H(dx) + 1 - \rho \right) + o(\rho)\,, \end{aligned} \tag{6}$$

where the last equality follows from the normalization chosen for the unit exceedance measure, $\int (1 - e^{-x^2/2})\, H(dx) = 1$. The integral appearing in last expression is a transform of

the exceedance measure H and, as a function of y, is called zeta function,

$$\zeta(y) = \int (\cosh(yx) - 1) e^{-x^2/2} \, H(dx) \, . \tag{7}$$

This function is analytic at the origin, and satisfies $\zeta(0) = 0$. It is also positive and finite, and this allows its product with the Gaussian density $\phi(y)\zeta(y)$, to be a probability density function. This in turn, means that the sparse approximation to the marginal density of Y can be written as a mixture of two components,

$$\rho\phi(y)\zeta(y) + (1-\rho)\phi(y) + o(\rho) \, .$$

The integrand of the zeta function is called zeta measure

$$\zeta(dx; y) = (\cosh(yx) - 1) e^{-x^2/2} H(dx) \, , \tag{8}$$

and it appears in the sparse approximation to the signal conditional distribution, given the observation $Y = y$,

$$\begin{aligned} P_\nu(dx \mid y) &= \frac{e^{yx} e^{-x^2/2} P_\nu(dx)}{\mathbb{P}(Y = y)} \\ &= \frac{e^{yx}}{\cosh(yx)} \frac{(\cosh(yx) - 1) e^{-x^2/2} P_\nu(dx) + e^{-x^2/2} P_\nu(dx)}{\mathbb{P}(Y = y)} \\ &= \frac{e^{yx}}{\cosh(yx)} \frac{\rho\zeta(dx; y) + e^{-x^2/2} P_\nu(dx)}{\rho\zeta(y) + 1 - \rho} + o(\rho) \, . \end{aligned} \tag{9}$$

This approximation is to be interpreted in terms of $\mathcal{W}^\#$-integrals. Thus, if one wants to compute the sparse approximation to $P_\nu(\epsilon^+ \mid |Y| = y)$, then, as for the unconditional probability of signal activity, we compute the expectation of $w_\epsilon(x) = 1 - e^{-x^2/2\epsilon^2}$,

$$P_\nu(\epsilon^+ \mid |Y| = y) = \frac{\rho \int w_\epsilon(x) \, \zeta(dx; y) + \rho \int e^{-x^2/2} \, H(dx)}{\rho\zeta(y) + 1 - \rho} + o(\rho) \, . \tag{10}$$

Now, as $\nu \to 0$ and $y \neq 0$, this approximation just converges to zero. In order not to get the trivial limit, one needs to let $y \to \infty$ in such a way that $\rho\zeta(y)$ converges to a strictly positive constant $\lambda > 0$. However, for the double limit,

$$\lim_{\nu \to 0} \lim_{y \to \infty} \int (\cosh(yx) - 1) e^{-x^2/2} \, P_\nu(dx) = \lim_{\nu \to 0} \lim_{y \to \infty} \rho\zeta(y) \,,$$

to be true for any sparse family P_ν having (ρ, H) as sparsity pair, H cannot have Gaussian nor sub-Gaussian tails. We derive this extra requirement in Section 8.5.2 of Chapter 8, and refer to it as the double limit condition (DLC). If DLC holds, as it does for inverse-power measures, then, as $\rho\zeta(y) \to \lambda$, the conditional probability of signal activity in (10) can be approximated by

$$\frac{\rho\zeta(y)}{1 + \rho\zeta(y)} + o(1) \,.$$

Symmetry and vector sparsity

The sparse signal-plus-noise model we presented in the previous section, is derived for a one-dimensional signal, whose distribution is symmetric around zero. Now suppose that, instead of one-dimensional, the signal is d-dimensional, with $d > 1$.

If the d components of the random vector are mutually independent, then we can regard the random vector as a collection of d independent signals, each of which has a sparse distribution in the sense of Definition 0.0.3; so in this case, we talk about component-wise sparsity. For this kind of product of sparse measures, except for some further connection with Lévy processes, the theory is really the same as for the univariate case, and the notion of signal negligibility presented above can become particularly useful in some applications, such as multiple testing, wavelet regression and Gaussian graphical models.

On the other hand, if this independence assumption is not very compelling, but one still has exchangeability of the d components of the random vector, then the sparse measure $P_{\nu,d}$

on $\mathbb{R}^d$ can be taken to be invariant under rotations. This means that, for all Borel sets A,

$$P_{\nu,d}(A) = P_{\nu,d}(\sigma A)$$

for every transformation $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ which is orthogonal with respect to the inner product chosen on $\mathbb{R}^d$, defining the metric $d(x, y) = \|x - y\| = \langle x-y, x-y\rangle^{1/2}$. In fact, since rotations are isometries, they preserve the distance between any two points $x, y \in \mathbb{R}^d$, $d(x, y) = d(\sigma x, \sigma y)$. So, if the inner product on $\mathbb{R}^d$ is defined by some positive definite matrix $A \in \mathbb{R}^{d\times d}$, $\langle x, y\rangle = x'Ay$, then σ must be such that $\sigma' A\sigma = A$ and $\det(\sigma) = \pm 1$.

Another way of seeing this is that if all directions are to be equally likely, then it is possible to factorize the sparse measure $P_{\nu,d}$ into two components,

$$\Gamma(d\,\tilde{x})P_\nu^R(d\,\|x\|)\,,$$

a spectral measure Γ for the direction vector $\tilde{x} = x/\,\|x\|$ on the unit sphere $\mathcal{S}^d = \{z : \|z\|^2 = 1\}$, and a radial measure P_ν^R for the radius of the vector $\|x\|$. Rotational invariance requires Γ to be the uniform measure on $\mathcal{S}^d$; so in order to have $P_{\nu,d}$ converging to the delta measure at the origin, as $\nu \to 0$, it is necessary that the radial measure P_ν^R converges to the Dirac delta measure at zero. This can be easily achieved by assuming that P_ν^R is two times the positive part of some sparse measure P_ν on $\mathbb{R}$.

Following this approach, the sparsity of the random vector is driven by the sparsity of its radius, and the scalar notion of symmetry on $\mathbb{R}$ is generalized to spherical symmetry (rotational invariance) on $\mathbb{R}^d$. In this case, we talk about vector sparsity.

The flexibility of choosing the metric that one wants to be preserved under rotations, allows one to frame the linear regression problem $Y = X\beta + \eta$, $\eta \sim N(0, \sigma^2 I_n)$, in such a way that the coefficient $\beta \in \mathbb{R}^d$ has a sparse distribution which is rotationally invariant with respect to the Fisher-information metric, given by $\|\beta\|^2 = \beta' X'X\beta$. With this assumption,

the marginal distribution of Y only depends on the covariate matrix X through its projection matrix $P_{\mathcal{X}} = X(X'X)^{-1}X'$, so the choice of basis for X is immaterial. Moreover, the sparse approximation to the Bayes factor for the exceedance event $\{\|\beta\| > \epsilon\}$ is

$$\zeta_d(\|P_{\mathcal{X}} y\| / \sigma),$$

where ζ_d is the generalization of the zeta function to the d-dimensional case. Then it is quite natural to start looking at this expression, or its generalization to the case when σ^2 is estimated by the residual mean square, as a sparse counterpart to the ANOVA F-test.

Tour around thesis

As mentioned at the beginning of this introduction, the thesis is structured in three parts. Here we try to give an insight of how we grouped certain topics together, even though, as already pointed out, the reader should not think of them as separate.

Part I further investigates the univariate notion of sparsity as a probabilistic limit and as an approximation device. The intention of this first part is two-fold: on the one hand, we show the potential use of this theoretical definition in very practical ways. For instance, in Chapter 2, we use the sparsity rate estimated from the data, as a likelihood-based criterion to choose the sparsity scale of a large covariance matrix. On the other hand, we explore how far one can push the limit definition of sparsity in terms of finding higher-order terms, $\rho_k H_k$, in the integral expansion of the sparse measure P_ν.

Both second and third parts, instead, aim at extending the univariate notion of sparsity to d-dimensional distributions, for $d > 1$, and present some examples of how these extensions can be used in certain statistical problems.

In Part II, we investigate d-dimensional measures that are the product of d univariate sparse measures, and refer to this case as component-wise sparsity. We show how this kind of sparsity assumption, together with the negligibility idea, can be used for different purposes: (i) constructing a multiple testing procedure to declare negligible and non negligible signals; (ii) obtaining a thresholding estimator such as the conditional median, to smooth the estimated function in wavelet regression; (iii) obtaining a thresholding estimator based on the conditional probability of non-negligibility, for the recovery of the graphical structure in a Gaussian graphical model.

Nevertheless, Part II could very well be considered as a development of the univariate sparsity theory, insofar the notion of signal negligibility, and its related integral sparse approximation, really refer to any univariate sparse measure P_ν. Thus, one should look at component-wise sparsity as only one of the possible contexts in which the negligibility notion and its machinery can be used.

In Part III instead, we study those sparse measures which are invariant under rotations about the origin, and introduce the idea of vector sparsity. This latter refers to random vectors whose sparsity is induced by the univariate sparsity of their radial part. Within this context, we introduce the d-dimensional $\cosh_d$ function and extend the signal-plus-noise theory to the case when the signal is vector sparse. We also apply this notion of vector sparsity to the linear regression setting, assuming that the coefficient vector β has a vector sparse distribution which is rotationally invariant with respect to the Fisher-information metric. This in turn allows us to derive a sparsity analog to the F-ratios, classically employed in analysis of variance.

Part I

Univariate sparsity

Chapter 1

Sparsity for comparing two means

1.1 Introduction

In our intentions, this chapter serves as a bridge between some of the ideas presented in a somewhat condensed way, in the introduction, and the rest of the thesis, where these ideas are further explored and developed in different contexts. To this end, we illustrate how one can use the sparse signal-plus-noise model of Equation (5) to analyze, in probabilistic terms, the difference between two unknown quantities, which are indeed believed to be very close to each other.

Motivated by this, we first derive the formulas for the scale-sparse signal-plus-noise model, in the case when the error variance is unknown, and is estimated using a chi-square statistic, independent of the signal-noise convolution. This leads us to introduce a zeta function for t-statistics, the t-zeta function on k degrees of freedom. After studying some asymptotic properties of this function, we use the sparse approximations to give a sparsity-based analysis of the 'Hubble constant dilemma', which has been at the center of a debate inflaming the cosmological community for the last decade.

1.2 Comparing two means

Suppose that we are interested in the difference between two unknown quantities, μ_1 and μ_2, which are expected to be very close to each other, and perhaps equal with some positive probability. Then we can look at the difference between these two unknown quantities, as a signal having a sparse distribution according to Definition 0.3. More specifically,

$$\mu_1 - \mu_2 \sim P_\nu \,,$$

where P_ν is a scale sparse distribution, which is symmetric around zero and has sparsity rate ρ, and inverse-power exceedance measure H. Further, suppose that, for each of the two unknown quantities, we have some Gaussian observations having mean μ_i. So, for $i = 1, 2$, we observe $(X_{i,1}, \ldots, X_{i,n_i})$,

$$X_{i,j} \mid \mu_1, \mu_2 \sim N(\mu_i, \sigma^2) \quad \text{for} \quad j = 1, \ldots, n_i \,,$$

where σ^2 is the unknown variance of the noise component. Then the difference between the two sample means,

$$\bar{X}_1 - \bar{X}_2 = \frac{1}{n_1} \sum_j X_{1,j} - \frac{1}{n_2} \sum_j X_{2,j} \,,$$

is such that

$$\bar{X}_1 - \bar{X}_2 \mid \mu_1, \mu_2 \sim N\left(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)\right) \,,$$

so one can write

$$\bar{X}_1 - \bar{X}_2 = (\mu_1 - \mu_2) + (\eta_1 - \eta_2) = \mu + \eta \,, \tag{1.1}$$

where $\mu \sim P_\nu$ is scale sparse with sparsity pair (ρ, H), and is independent of the noise $\eta \sim N(0, \sigma^2(1/n_1 + 1/n_2))$.

Now, if σ^2 was known, after rescaling, this could be easily framed as the sparse signal-

plus-noise problem presented in the introduction, for which the sparse approximations of all the important functionals have already been developed in McC&P. However, if σ^2 is not known, then one has to estimate it and take the variability of its estimator into account when deriving the sparse approximations. In the next section, we extend the sparsity analysis of the classical signal-plus-noise model to the case when the variance of the noise part is unknown.

1.3 Signal-plus-noise estimating the error variance

Consider the signal-plus-noise model

$$Y = \mu + \eta\,,$$

where $\mu \sim P_\nu$ is scale sparse with rate ρ and inverse-power exceedance measure H, while $\eta \sim N(0, \sigma^2)$, and σ^2 is unknown. Moreover, assume μ and η to be independent. Then the scaled variable

$$\tilde{Y} = \frac{Y}{\sigma}\,,$$

is the sum of the two independent components: a scaled sparse signal $\mu/\sigma \sim P_{\nu,\sigma}$ and a standard Normal $\eta/\sigma \sim N(0,1)$. Now, because P_ν is assumed to be a sparse distribution whose sparsity parameter coincides with its scale parameter, then $P_{\nu,\sigma}$ is also scale sparse with sparsity-scale parameter given by $\tilde{\nu} = \nu/\sigma$. This means that $P_{\nu,\sigma} = P_{\tilde{\nu}}$ is sparse with exceedance measure H and sparsity rate $\tilde{\rho} = \tilde{\nu}^{\,\alpha} = \rho\sigma^{-\alpha}$. Indeed, H has density function which is homogeneous of order $\alpha - 1$, so that

$$H(d(\sigma x)) = h(\sigma\, x)d(\sigma x) = \sigma^{-\alpha}h(x)dx = \sigma^{-\alpha}H(dx)\,.$$

This leads us to write the sparse approximation to the marginal density of $\tilde{Y}$ at $\tilde{y}$, as

$$m_{\nu,\sigma}(\tilde{y}) = \phi(\tilde{y})\,(\tilde{\rho}\zeta(\tilde{y}) + 1 - \tilde{\rho}) + o(\tilde{\rho})\,.$$

Now, suppose that s^2 is an estimator of σ^2 such that $ks^2/\sigma^2 \sim \chi^2_k$, for some $k \geq 1$. If Y and s^2 are independent, then letting

$$T = \frac{Y}{s}\,,$$

the sparse approximation to the marginal density of T at t is

$$(1-\tilde{\rho})t_k(t) + \tilde{\rho}\, t_k(t)\zeta_k^T(t) + o(\tilde{\rho})\,.$$

Here t_k denotes the density function of Student's t distribution with k degrees of freedom, while

$$\zeta_k^T(t) = \sum_{r=1}^{\infty} \left(\frac{t^2}{t^2+k}\right)^r \frac{2^{2r}}{(2r)!} \frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} \frac{\alpha\,\Gamma(r-\alpha/2)}{2\,\Gamma(1-\alpha/2)}\,, \tag{1.2}$$

denotes the zeta function for the t-statistic on k degrees of freedom, associated with the α-inverse-power measure H. See the appendix for the derivation.

Notice that, because $\psi(z) = \phi(z)\zeta(z)$ is a probability density function, also the product $\psi_k(t) = t_k(t)\,\zeta_k^T(t)$ is a probability density function as

$$\begin{aligned}\int t_k(t)\,\zeta_k^T(t)dt &= \int\int \phi(t\sqrt{u})\zeta(t\sqrt{u})\chi^2_k(ku)k\sqrt{u}\,du\,dt \\ &= \int\int \phi(z)\zeta(z)dz\sqrt{u}^{-1}\,\chi^2_k(ku)k\sqrt{u}\,du = \int \chi^2_k(ku)k\,du = 1\,.\end{aligned}$$

In Figure 1.1, we compare the ψ_k function, depicted by the light blue solid curve, with ψ, depicted by the black dashed curve, for different degrees of freedom k. We can see that both densities are bimodal and symmetric around the origin. Yet, for small values of k, the ψ_k density has heavier tails than ψ, while, as k gets larger, the two densities get closer and closer. Indeed, as $k \to \infty$, the t-zeta function in (1.2) converges to the ordinary zeta function, which in fact, when the exceedance measure is the α-inverse-power measure, can

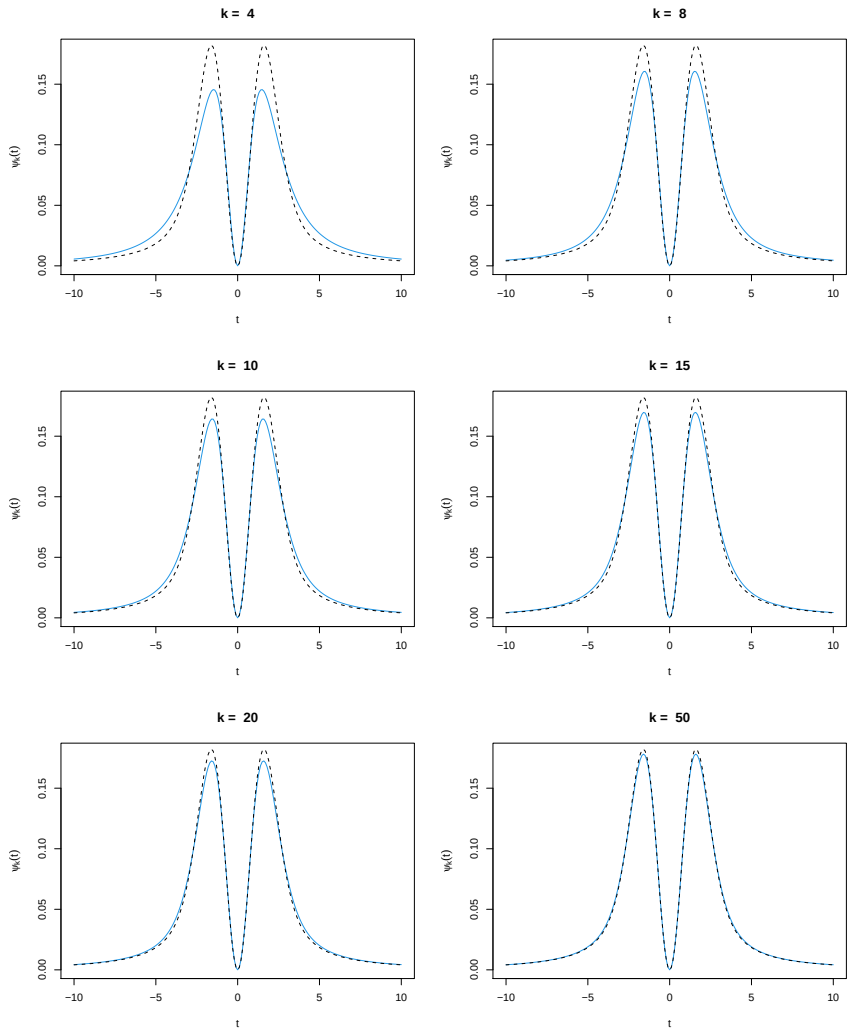


Figure 1.1: Comparison of the product $\psi_k(t) = t_k(t)\,\zeta_k^T(t)$ (solid blue line) versus $\psi(t) = \phi(t)\,\zeta(t)$ (dashed black line), for different values of the degrees of freedom k.

be written as a convergent power series

$$\zeta(t) = \sum_{r=1}^{\infty} t^{2r} \frac{2^r}{(2r)!} \frac{\alpha\,\Gamma(r-\alpha/2)}{2\,\Gamma(1-\alpha/2)}\,. \tag{1.3}$$

See the appendix for a proof of this convergence. Therefore, as expected, $k \to \infty$ implies that the marginal density of $T = Y/s$,

$$m_\nu^T(t) = t_k(t)\left((1-\tilde{\rho}) + \tilde{\rho}\,\zeta_k^T(t)\right) + o(\tilde{\rho})\,,$$

converges to the marginal density of $\tilde{Y} = Y/\sigma$,

$$m_\nu(t) = \phi(t)\left((1-\tilde{\rho}) + \tilde{\rho}\,\zeta(t)\right) + o(\tilde{\rho})\,.$$

This convergence is shown in Figure 1.2. The top panel shows the convergence $\zeta_k^T(t) \to \zeta(t)$, plotted on the log scale, while the bottom panel shows $t_k(t) \to \phi(t)$, again on the log scale.

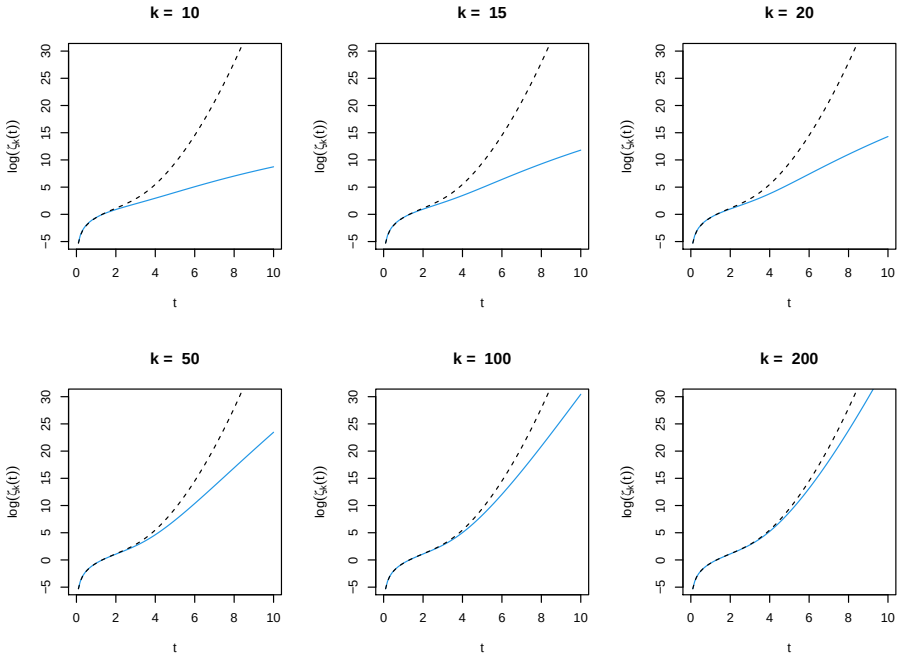
On the other hand, for fixed k, if we let the argument $t \to \infty$, then $t^{2r}/(t^2+k)^r \to 1$ so that the t-zeta function converges to

$$\sum_{r=1}^{\infty} \frac{2^{2r}}{(2r)!} \frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} \frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\,.$$

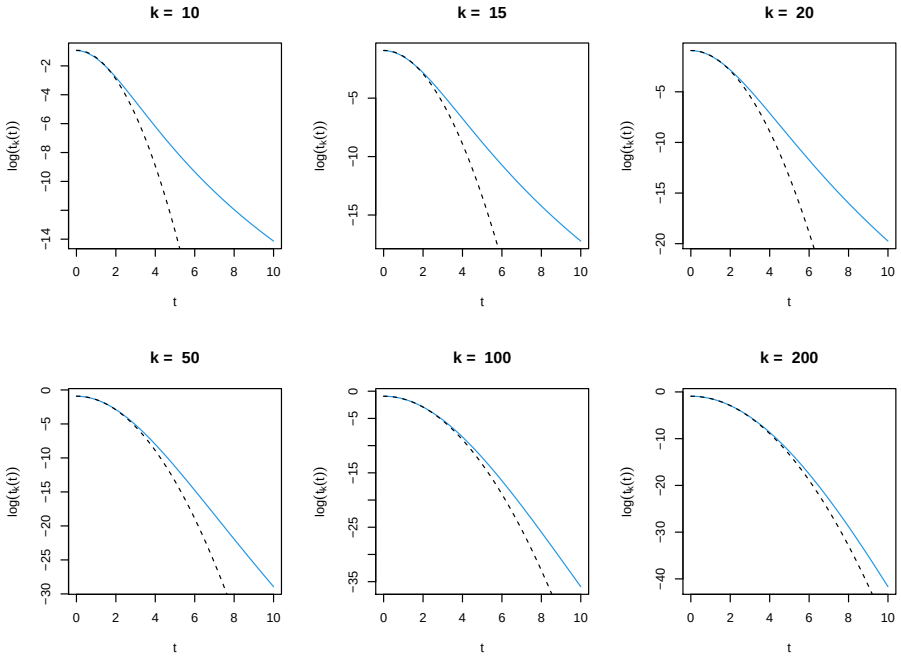
This series is divergent so that in fact, $\zeta_k^T(t) \to \infty$ as $t \to \infty$. Nevertheless, using the Laplace approximation, we can gauge the behavior of $\zeta_k^T(t)$ for large t,

$$(t^2+k)^{\frac{k+1}{2}} |t|^{-\alpha-1} \frac{\sqrt{\pi}\,\Gamma\left(\frac{k-\alpha}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right) k^{k/2-\alpha/2}} \frac{\alpha}{2\,\Gamma(1-\alpha/2)}\,. \tag{1.4}$$

See the appendix for all details. This asymptotic approximation is shown in Figure 1.3. The Laplace approximation in (1.4), the dashed red curve, is quite close to the exact ζ_k^T, the light



(a) Convergence of the t-zeta function to the ordinary zeta function as the degrees of freedom $k \to \infty$. The blue solid line depicts $\log(\zeta_k^T(t))$ while the black dashed line depicts $\log(\zeta(t))$.



(b) Convergence of the t of Student density with k degrees of freedom to the Normal density as $k \to \infty$. The blue solid line depicts $\log(t_k(t))$ while the black dashed line depicts $\log(\phi(t))$.

Figure 1.2: Convergence of the marginal density of T to the marginal density of $\tilde{Y}$.

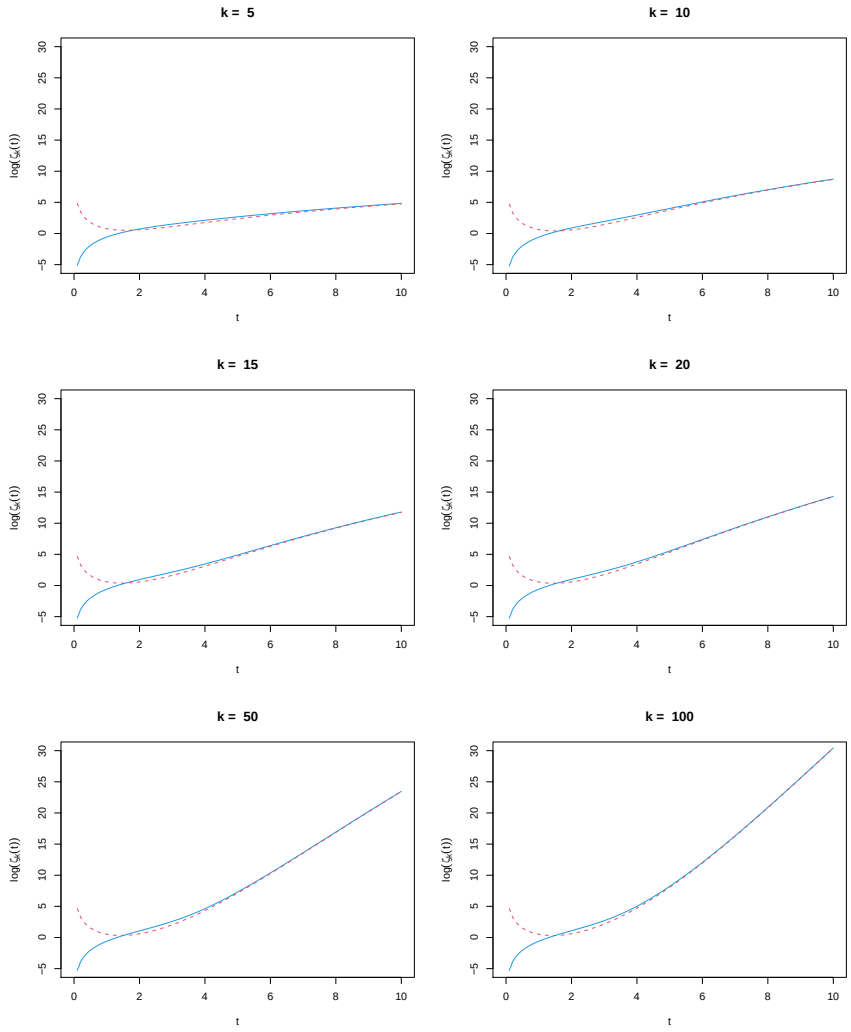


Figure 1.3: Asymptotic behavior of $\log(\zeta_k^T(t))$ as $t \to \infty$, light blue solid line. The red dashed line depicts the Laplace approximation in (1.4), also on the log scale.

blue solid curve, even for moderate values of t. Both functions are plotted on the log scale.

1.3.1 Signal conditional distribution with estimated error variance

From univariate sparsity theory, assuming $\sigma = 1$, the sparse approximation to the symmetrized conditional distribution of the signal, given $|Y| = y$, is

$$P_\nu(dx \mid |Y|) = \frac{1-\rho}{1-\rho+\rho\zeta(y)} \frac{e^{-x^2/2} P_\nu(dx)}{1-\rho} + \frac{\rho\zeta(y)}{1-\rho+\rho\zeta(y)} \frac{\zeta(dx;y)}{\zeta(y)} + o(\rho)\,.$$

Analogously, one can derive the first-order sparse approximation of the symmetrized conditional distribution of the scaled signal μ/σ given $|T| = t$

$$P_{\nu,\sigma}(\mu/\sigma \in dx \mid |T|) = \frac{1-\tilde{\rho}}{1-\tilde{\rho}+\tilde{\rho}\zeta_k^T(t)} \frac{e^{-x^2/2} P_{\nu,\sigma}(dx)}{1-\tilde{\rho}} + \frac{\tilde{\rho}\zeta_k^T(t)}{1-\tilde{\rho}+\tilde{\rho}\zeta_k^T(t)} \frac{\zeta_k^T(dx;t)}{\zeta_k^T(t)} + o(\tilde{\rho}),$$

where $\zeta_k^T(du;t)$ denotes the t-zeta measure defined as

$$\zeta_k^T(du;t) = \sum_{r=1}^{\infty} \left(\frac{t^2}{t^2+k}\right)^r \frac{2^r}{(2r)!} \frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} u^{2r} e^{-u^2/2} H(du).$$

See the appendix for details. Then the conditional probability of the scaled signal activity, $|\mu|/\sigma > \epsilon$, for any $\epsilon > 0$, can be approximated by

$$\begin{aligned} P_\nu(|\mu|/\sigma > \epsilon \mid |T| = t) &\approx \int (1 - e^{-x^2/2\epsilon^2})\, P_{\nu,\sigma}(dx \mid |T| = t) \\ &= \frac{\int (1 - e^{-x^2/2\epsilon^2}) \left(\tilde{\rho} e^{-x^2/2} H(dx) + \tilde{\rho}\zeta_k^T(dx;t)\right)}{1-\tilde{\rho}+\tilde{\rho}\zeta_k^T(t)} + o(\tilde{\rho}). \end{aligned} \tag{1.5}$$

As $\tilde{\rho} \to 0$ and $t \neq 0$, this expression converges to zero. In order to get a non trivial limit, we need to let $t \to \infty$ so that $\zeta_k^T(t) \to \infty$ in such a way $\tilde{\rho}\zeta_k^T(t) \to \lambda$, for $\lambda > 0$. Under this double limit regime, the dominating term in the numerator of (1.5), is $\tilde{\rho}\zeta_k^T(t)$ since, as $t \to \infty$,

$$\int (1 - e^{-x^2/2\epsilon^2})\, \zeta_k^T(dx;t) \sim \zeta_k^T(t),$$

when H has inverse-power tails. Therefore, as $\tilde{\rho}\zeta_k^T(t) \to \lambda$,

$$P_\nu(|\mu|/\sigma > \epsilon \mid |T| = t) = \frac{\tilde{\rho}\zeta_k^T(t)}{1+\tilde{\rho}\zeta_k^T(t)} + o(1).$$

Since this ratio does not depend on the threshold ϵ, if we instead consider ϵ/σ then, under the double limit regime, the conditional probability of activity for the original unscaled signal is

still

$$P_\nu(|\mu| > \epsilon \mid |T| = t) = \frac{\tilde{\rho}\zeta_k^T(t)}{1 + \tilde{\rho}\zeta_k^T(t)} + o(1)\,. \tag{1.6}$$

Table 1.1 reports the values of t such that $\tilde{\rho}\zeta_k^T(t) = 1$, corresponding to different degrees of freedom k, for decreasing values of the rate $\tilde{\rho}$. The column with $k = \infty$ corresponds to the values t such that $\tilde{\rho}\zeta(t) = 1$. We can observe that, for small values of the degrees of freedom, the value of t necessary for $\tilde{\rho}\zeta_k^T(t) = 1$ to hold, needs to be larger. In other words, when fewer degrees of freedom are available to estimate σ^2, we need to observe more extreme values of the statistic $|T|$ for getting $1/2$ as the conditional probability that the signal is larger than ϵ.

$\tilde{\rho}$	$k = 5$	$k = 8$	$k = 10$	$k = 15$	$k = 20$	$k = \infty$
10%	4.31	3.56	3.36	3.12	3.01	2.72
5%	5.62	4.35	4.03	3.66	3.50	3.07
1%	9.35	6.25	5.55	4.80	4.48	3.71
0.5%	11.36	7.12	6.22	5.26	4.87	3.94
0.1%	17.45	9.39	7.86	6.34	5.74	4.40

Table 1.1: Values of t such that $\tilde{\rho}\zeta_k^T(t) = 1$, for different values of $\tilde{\rho}$ and k.

It then follows that the Bayes Factor for the event $|\mu| > \epsilon$,

$$BF(|\mu| > \epsilon) = \frac{\text{Odds}(|\mu| > \epsilon \mid |T| = t)}{\text{Odds}(|\mu| > \epsilon)}\,,$$

under the double limit regime, behaves like

$$\frac{\tilde{\rho}\zeta_k^T(t) + o(1)}{\rho H(\epsilon^+)/(1 - \rho H(\epsilon^+)) + o(\rho)} = \frac{\sigma^{-\alpha}\zeta_k^T(t)}{H(\epsilon^+)} + o(1)\,,$$

where $H(\epsilon^+) = 2\int_\epsilon^\infty H(dx)$. This last expression does not depend on the sparsity rate ρ and can be estimated by

$$\frac{s^{-\alpha}\,\zeta_k^T(t)}{H(\epsilon^+)} + o(1)\,.$$

If we choose ϵ to be the standard activity threshold for which $H(\epsilon^+) = 1$, then for the

inverse-power exceedance measure, this threshold is

$$\epsilon = \left(\frac{2^{\alpha/2}}{\Gamma(1-\alpha/2)} \right)^{1/\alpha} .$$

For this choice of threshold activity, under the double limit $\tilde{\rho}\zeta_k^T(t) \to \lambda$, the Bayes factor for the event that the original sparse signal is above the threshold, $|\mu| > \epsilon$, can be estimated by

$$s^{-\alpha}\, \zeta_k^T(t)\,. \tag{1.7}$$

1.4 Comparing two means estimating the error variance

We now go back to the original problem of investigating the difference between two means μ_1 and μ_2, where, for each mean μ_i, the observations are independent

$$X_{i,j} \mid \mu_i \sim N(\mu_i, \sigma^2) \quad \text{for} \quad j = 1, \ldots, n_i\,.$$

Then,

$$\bar{X}_1 - \bar{X}_2 = (\mu_1 - \mu_2) + (\eta_1 - \eta_2) = \mu + \eta\,, \tag{1.8}$$

where $\mu \sim P_\nu$ is scale sparse with first-order pair (ρ, H), and is independent of $\eta \sim N(0, \sigma^2(1/n_1 + 1/n_2))$. To estimate σ^2, one can use the pooled estimator

$$s^2_{\text{pool}} = \frac{n_1 - 1}{n_1 + n_2 - 2}\, s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\, s_2^2\,, \tag{1.9}$$

where, for each $i = 1, 2$, $s_i^2 = \sum_j (X_{i,j} - \bar{X}_i)^2/(n_i - 1)$ is the sample variance, for which $(n_i - 1)s_i^2/\sigma^2 \sim \chi^2_{n_i-1}$. Then the variance of η can be then estimated by

$$s^2 = s^2_{\text{pool}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) ,$$

which is a linear combination of two independent mean squares, s_1^2 and s_2^2, on $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively. Therefore we can use the Welch-Satterthwaite approximation (Satterthwaite, 1946 [64]) to compute the degrees of freedom

$$k = \frac{(s^2)^2}{\sum_{i=1,2} \left(s_i^2 \frac{n_i - 1}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)^2 \frac{1}{n_i - 1}} , \tag{1.10}$$

for which approximately $s^2 \sim \chi^2_k$.

Since, for each of the two samples $i = 1, 2$, $\bar{X}_i$ and s_i^2 are independent conditionally on μ_i, they are also unconditionally independent as

$$\mathbb{P}(\bar{X}_i, s_i^2) = \int \mathbb{P}(\bar{X}_i, s_i^2 \mid \mu_i) \mathbb{P}(\mu_i) \, d\mu_i = \int \mathbb{P}(\bar{X}_i \mid \mu_i) \mathbb{P}(s_i^2) \mathbb{P}(\mu_i) \, d\mu_i = \mathbb{P}(\bar{X}_i) \mathbb{P}(s_i^2) \, .$$

So it follows that $\bar{X}_1 - \bar{X}_2$ and s^2 are independent. Thus, letting

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\text{pool}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} , \tag{1.11}$$

following the previous section, for any positive threshold ϵ, given the observed value t, we can compute the conditional probability that the signal $\mu = \mu_1 - \mu_2$ is above that threshold. Moreover, assuming the double limit regime $\tilde{\rho} \, \zeta_k(t) \to \lambda > 0$, the Bayes Factor for the event

$|\mu_1 - \mu_2| > \epsilon$ can be estimated by

$$\frac{\left(s^2_{\text{pool}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)^{-\alpha/2}\,\zeta_k(t)}{H(\epsilon^+)}\,.$$

1.5 The Hubble constant debate

In this section, we provide an example of how the difference between two means can be investigated using the theory presented in the previous section. The Hubble constant, the subject of our example, has been at the center of a heated debate among cosmologists for the last decade or so. In 1929 Edwin Hubble discovered that the universe is expanding, and, just as two marked points on an expanding balloon diverge at a rate proportional to their current separation, so too galaxies move away from Earth at a rate proportional to their distance from Earth. In formulae, if v is the galaxy recessional velocity and D is its distance to the Earth, the Hubble constant H_0 is the proportionality constant relating speed and distance:

$$v = H_0\, D\,.$$

Therefore, after computing the recessional velocity of a galaxy, the value of H_0 can be inferred by estimating its distance from the Earth. This last task can be carried out using two different methods. The local universe method, proposed by Hubble himself, is based on determining the distance D of far-away objects using the so called cosmic distance ladder (CDL). This method relies upon identifying standard candles, objects that shine with an intrinsic brightness, and then use parallax to convert this brightness to distance from the Earth. The idea of CDL is to start by measuring brightness of nearby pulsating stars, for instance the Cepheids in the Milky Way, and then, using geometry to calibrate the luminosity, move out in the universe, measuring brightness of exploding stars, like the Type 1A Supernovae, present in much farther-away galaxies.

The second method, on the contrary, looks back in time and uses the cosmic microwave background (CMB), the electromagnetic radiation leftover from the Big Bang, to establish a unit of distance at the early stage of the cosmos, when the universe was 380,000 years old. These measurements can then be fast forward and used to predict the Hubble constant following the ΛCDM model. This latter is the standard cosmology model which aims at describing all the visible matter and energy, together with dark energy (Λ) and cold dark matter (CDM), showing how they evolve according to Albert Einstein's theory of gravity.

The debate about the Hubble constant began a few years ago when, despite the increased accuracy of the measurements in both methods, the predictions for H_0 delivered by the local universe and the CMB methods, started to get farther and farther apart. When cosmologists from the Planck group (Aghanim et al., 2020 [2]), among others, used the data from the early universe to predict the expansion rate, they found it to be 67.4 ± 0.5. Yet, when adopting the local universe method and using the data from the current stage of the universe, several different teams of cosmologists found much higher estimates for H_0. For instance, the estimate found by Professor Riess's team was 73.5 ± 1.4, roughly 4.1σ above the other prediction.

Now, if in fact the two predictions were indeed not due to statistical and measurement errors, then it would mean that something is missing in the ΛCDM cosmology model since the universe would now be expanding at a faster rate than that predicted for the early-stage universe. However, among cosmologists, there is not a consensus in regarding this difference as real or not. Indeed, some of the most recent measurements delivering higher estimates for H_0, such as those of Professor Riess and his team (Riess et al., 2019 [59]), are put into question by Professor Wendy Freedman, who, despite having pioneered the usage of the Cepheids as standard candles, now casts doubts on their reliability. Freedman's team, in their paper Freedman et al. (2019) [37], report a much lower estimate, 69.8 ± 1.9, obtained

using the "tip of the red giant branch" (TRGB) stars. These are stars that, being about to die, gradually grow brighter until they reach a characteristic peak brightness, which is always the same. This stability makes them good standard candles, besides the fact that, as old stars, they can be observed in the clean peripheries of galaxies, and not the dusty and crowded regions where the Cepheids are usually observed.

The debate is far from being closed and new developments are expected when new data from the Gaia space telescope will enable much more precise calibrations of the Cepheids and TRGB stars. For a gentle introduction to this fascinating dilemma, we refer to Verde et al. (2019) [71] and to two articles published in the Quanta Magazine, Wolchover (2019) [77] and Wolchover (2020) [78].

1.5.1 Sparsity analysis for the Hubble constant

What we propose in this section is to look at the hypothetical difference between the early and late Hubble constants, i.e., the difference between the cosmic expansion rate at the early stage of the universe μ_1, and the present rate μ_2, as a sparse signal. As such, the difference $\mu_2 - \mu_1$ can be thought of being zero with some positive probability and being different from zero with a probability distribution having polynomial tails. In formulae,

$$\mu_2 - \mu_1 \sim P_\nu \,,$$

where P_ν is a scale sparse distribution with rate ρ and H is the inverse-power exceedance measure. Then, we look at the discrepancies between measurements obtained from different methods, as corrupted observations of the sparse signal, as in (1.8).

The first comparison is between the CMB - Planck measurement and the CDL - Cepheids measurement, and the data reported in Riess (2020) [58] are:

- CMB - Planck: $\bar{x}_1 = 67.4$, $\text{SE}_1 = \frac{s_1}{\sqrt{n_1}} = 0.5$,
- CDL - Cepheids: $\bar{x}_2 = 73.5$, $\text{SE}_2 = \frac{s_2}{\sqrt{n_2}} = 1.4$.

For the second comparison instead, the CMB - Planck measurement versus the CDL - tip of the red giant branch (TRGB) stars measurement, the data reported in Riess (2020) [58] are:

- CMB - Planck: $\bar{x}_1 = 67.4$, $\text{SE}_1 = \frac{s_1}{\sqrt{n_1}} = 0.5$,
- CDL - TRGB stars: $\bar{x}_2 = 69.8$, $\text{SE}_2 = \frac{s_2}{\sqrt{n_2}} = 1.9$.

We do not know how the accuracy measures were determined, so we ignore on how many degrees of freedom these standard errors were obtained. For this reason, we will present our sparsity study hypothesizing different combinations of values for n_1 and n_2, so to account for different scenarios for the precision of the measurements.

Assuming equality of the variances for the Gaussian noises corrupting $\bar{X}_i$, $i = 1, 2$, we use the pooled estimator for σ^2 in (1.9) and let T as in (1.11). Table 1.2 reports the values for $t = (\bar{x}_2 - \bar{x}_1)/s$ and $s = s_{\text{pool}}\sqrt{(1/n_1 + 1/n_2)}$, shown in brackets, for different combinations of (n_1, n_2). As expected, the t-values comparing the Planck estimate to the Cepheids estimate (Table 1.2a) are much larger than those comparing the Planck estimate to the estimate using the TRGB stars as standard candles (Table 1.2b). This reflects both the larger difference of $\bar{x}_2$ from $\bar{x}_1$ for the Cepheids comparison, and the smaller estimated variance s. One side note is the following. Looking at the first column in Table 1.2a, we can observe that the relationship between s^2 and n_1, for fixed n_2, is not monotone. In fact,

$$s^2 = \left(\frac{n_1 - 1}{n_1 + n_2 - 2} n_1 \text{SE}_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} n_2 \text{SE}_2^2 \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right),$$

so keeping n_2 constant, as n_1 grows, the relative weights push s_{pool} towards the smaller s_1^2 rather than the bigger s_2^2. Yet, at the same time $s_1^2 = n_1 0.5^2$ also increases. On the other hand, if we fix n_1 and let n_2 increase, then both effects push in the same direction, leading

to a bigger estimate of s^2.

(a) Planck versus Cepheids

	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 15$
$n_1 = 4$	4.10	3.40	2.97	2.67	2.20
	(1.49)	(1.79)	(2.06)	(2.29)	(2.78)
$n_1 = 6$	4.80	4.10	3.61	3.26	2.69
	(1.27)	(1.49)	(1.69)	(1.87)	(2.27)
$n_1 = 8$	5.14	4.59	4.10	3.73	3.10
	(1.19)	(1.33)	(1.49)	(1.64)	(1.97)
$n_1 = 10$	5.27	4.90	4.47	4.10	3.44
	(1.16)	(1.24)	(1.36)	(1.49)	(1.77)
$n_1 = 15$	5.13	5.22	5.01	4.74	4.10
	(1.19)	(1.17)	(1.22)	(1.29)	(1.49)

(b) Planck versus TRGB stars

	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 15$
$n_1 = 4$	1.22	1.00	0.87	0.78	0.64
	(1.96)	(2.41)	(2.77)	(3.09)	(3.77)
$n_1 = 6$	1.48	1.22	1.06	0.95	0.78
	(1.63)	(1.96)	(2.26)	(2.52)	(3.07)
$n_1 = 8$	1.64	1.39	1.22	1.10	0.90
	(1.47)	(1.72)	(1.96)	(2.18)	(2.65)
$n_1 = 10$	1.73	1.52	1.35	1.22	1.01
	(1.39)	(1.58	(1.78)	(1.96)	(2.38)
$n_1 = 15$	1.80	1.71	1.57	1.45	1.22
	(1.34)	(1.41)	(1.52)	(1.65)	(1.96)

Table 1.2: Values of t and s (in brackets), for different combinations of n_1 and n_2.

In terms of sparsity, we assume that the sparsity rate for $\mu_2 - \mu_1$ is 0.05 and let the exceedance measure be the inverse square measure, i.e., $\alpha = 1$, so that the standard-activity threshold is $\epsilon = \sqrt{2/\pi} = 0.8$. For each different estimate of s corresponding to different combinations of (n_1, n_2), $\tilde{\rho}$ is computed as $s^{-\alpha} \cdot 0.05 = s^{-1} \cdot 0.05$.

In Table 1.3 we report $\tilde{\rho}\,\zeta_k(t)/(1 + \tilde{\rho}\,\zeta_k(t))$, the sparse approximation to the conditional probability that the difference between the two means is larger some positive threshold $\epsilon > 0$.

The values are multiplied by 10^2, so they can be read in terms of percentages. Here the degrees of freedom k are found using the Welch-Satterthwaite approximation in (1.10). Looking at the posterior probabilities that $|\mu_2 - \mu_1| > \epsilon$, for $\epsilon > 0$, the two methods give very different answers. Using the TRGB stars (Table 1.3b), the conditional probability ranges from as low as 0.16% to roughly 4%, whereas, using the Cepheids stars (Table 1.3a), the range of probabilities goes from 3.2% to as high as nearly 82%. Overall, these figures seem to be telling two stories hard to reconcile. However, some caution should be used when looking at the results for the TRGB stars method. In fact, the t-values for this method are very low and hardly justify the double limit regime under which the expression in (1.6) is derived for the conditional probability.

(a) Planck versus Cepheids

	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 15$
$n_1 = 4$	13.99	10.90	8.07	6.01	3.23
$n_1 = 6$	24.89	20.78	14.33	11.08	5.95
$n_1 = 8$	43.56	33.18	23.97	18.90	9.73
$n_1 = 10$	59.51	45.70	35.61	29.19	15.31
$n_1 = 15$	81.80	76.55	67.99	59.15	36.14

(b) Planck versus TRGB stars

	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 15$
$n_2 = 4$	1.46	0.75	0.46	0.32	0.16
$n_2 = 6$	2.28	1.21	0.75	0.52	0.26
$n_2 = 8$	2.97	1.66	1.05	0.73	0.37
$n_2 = 10$	3.45	2.09	1.35	0.95	0.48
$n_2 = 15$	3.94	2.87	2.02	1.48	0.78

Table 1.3: Conditional probabilities for $|\mu_2 - \mu_1| > \epsilon$, for any $\epsilon > 0$, multiplied by 10^2.

Table 1.4 shows the sparse approximation to the Bayes factor for the event $|\mu_2 - \mu_1| > \epsilon$, $\epsilon = 0.8$, estimated by $s^{-1}\,\zeta_k(t)$ as in (1.7). Recall that the unconditional probability for this event is exactly $\rho = 0.05$, irrespective of any sample size. We also report in brackets, the value of $\zeta_k(t)$ to give an idea of the effect of n_1 and n_2 on the t-zeta function. Looking

at these figures, the divergence between the two measurement approaches appears to be even more remarkable. In fact, even if we were to assume that all three methods had the same accuracy, with the Cepheids measurement we can get a Bayes factor over 30, while the corresponding Bayes factor with the TRGB stars measurement, does not even reach 0.5. And this divergence is mainly driven by the values of $\zeta_k(t)$. Yet, once more, one should be cautious in interpreting the Bayes factors in Table 1.4b, as the double limit regime is not taking place. All the same, we can regard these figures as upper bounds of the exact values, so the conclusion from TRGB stars is even more on the negligibility side.

After seeing this analysis, the tension arisen among cosmologists, should not come as a surprise. Depending on which kind of stars are used as standard candles to construct the cosmic distance ladder, the evidence in favor of a discrepancy between the hypothesized early and late Hubble constants, varies dramatically. With Freedman team's choice for the TRGB stars, it seems there is not much evidence for hypothesizing two different Hubble constants. On the contrary, following Riess's approach, one can reach a totally opposite conclusion. Yet, for this last method, the accuracy of the measurements plays a crucial role in determining how strong the evidence is in favor of the accelerating expansion hypothesis. Indeed, if the Planck measurement is much more precise than the Cepheids measurement, then the posterior probability of a difference in the constants is around 90%; but if the relative precision is inverted, then the same probability goes down to 5.6%.

(a) Planck versus Cepheids

	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 15$
$n_1 = 4$	4.60	3.79	2.87	2.16	1.18
	(6.84)	(6.80)	(5.90)	(4.95)	(3.29)
$n_1 = 6$	10.3	9.09	6.20	4.82	2.62
	(13.1)	(13.5)	(10.5)	(9.03)	(5.94)
$n_1 = 8$	25.2	18.4	12.6	9.83	4.92
	(29.9)	(24.5)	(18.8)	(16.1)	(9.70)
$n_1 = 10$	49.7	32.6	23.3	18.4	8.86
	(57.6)	(40.6)	(31.8)	(27.4)	(15.7)
$n_1 = 15$	160	135	97.1	70.9	31.0
	(190)	(158)	(118)	(91.4)	(46.1)

(b) Planck versus TRGB stars

	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 15$
$n_1 = 4$	0.42	0.23	0.15	0.11	0.06
	(0.82)	(0.56)	(0.42)	(0.33)	(0.22)
$n_1 = 6$	0.72	0.42	0.28	0.20	0.11
	(1.18)	(0.83)	(0.63)	(0.51)	(0.33)
$n_1 = 8$	1.00	0.63	0.43	0.31	0.17
	(1.46)	(1.08)	(0.84)	(0.68)	(0.45)
$n_1 = 10$	1.21	0.83	0.58	0.43	0.24
	(1.67)	(1.30)	(1.03)	(0.84)	(0.56)
$n_1 = 15$	1.46	1.22	0.94	0.73	0.43
	(1.95)	(1.72)	(1.44)	(1.21)	(0.84)

Table 1.4: Bayes factor for $|\mu_2 - \mu_1| > 0.8$ and $\zeta_k(t)$ shown in brackets.

1.6 Appendix

1. Here we derive the sparse approximation to the marginal density of $T = Y/s$. Denote by $\tilde{Y} = Y/\sigma$ and $W = s^2/\sigma^2$, then

$$
\begin{aligned}
&\mathbb{P}(T \in dt) = \\
&\mathbb{P}\left(\frac{\tilde{Y}}{\sqrt{W}} \in dt\right) = \\
&\int \mathbb{P}(\tilde{Y} \in d(t\sqrt{u}), W \in d(ku))\, du = \\
&\int m_{\nu,\sigma}(t\sqrt{u})\chi^2_k(ku)k\sqrt{u}\, du = \\
&\int \phi(t\sqrt{u})(1-\tilde{\rho}+\tilde{\rho}\zeta(t\sqrt{u})+o(\tilde{\rho}))\chi^2_k(ku)k\sqrt{u}\, du = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\int \phi(t\sqrt{u})\zeta(t\sqrt{u}))\chi^2_k(ku)k\sqrt{u}\, du + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\int \phi(t\sqrt{u})\int \sum_{r=1}^{\infty}\frac{t^{2r}u^r x^{2r}}{(2r)!}e^{-x^2/2}\, H(dx)\ \chi^2_k(ku)k\sqrt{u}\, du + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\int\int \phi(t\sqrt{u})\sum_{r=1}^{\infty}\frac{t^{2r}u^r x^{2r}}{(2r)!}\chi^2_k(ku)k\sqrt{u}\, du\ e^{-x^2/2}\, H(dx) + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\int \sum_{r=1}^{\infty}\frac{t^{2r}x^{2r}}{(2r)!}\left(\int \phi(t\sqrt{u})u^r\chi^2_k(ku)k\sqrt{u}\, du\right)\ e^{-x^2/2}\, H(dx) + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\int \sum_{r=1}^{\infty}\frac{t^{2r}x^{2r}}{(2r)!}\left(t_k(t)\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{2^r}{(t^2+k)^r}\right)\ e^{-x^2/2}\, H(dx) + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\, t_k(t)\sum_{r=1}^{\infty}\frac{t^{2r}x^{2r}}{(2r)!}\left(\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{2^r}{(t^2+k)^r}\right)\int x^{2r}e^{-x^2/2}\, H(dx) + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\, t_k(t)\sum_{r=1}^{\infty}\frac{t^{2r}}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{2^r}{(t^2+k)^r}\int x^{2(r-\alpha/2)-1}e^{-x^2/2}K_\alpha\, dx + o(\tilde{\rho}) = \\
&(1-\tilde{\rho})t_k(t) + \tilde{\rho}\, t_k(t)\sum_{r=1}^{\infty}\frac{t^{2r}}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{2^{2r}}{(t^2+k)^r}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1/\alpha/2)} + o(\tilde{\rho}) = \\
&t_k(t)\left((1-\tilde{\rho}) + \tilde{\rho}\,\zeta_k^T(t)\right) + o(\tilde{\rho})\,.
\end{aligned}
$$

2. Here we show that the t-zeta function on k degrees of freedom converges to the ordinary zeta function as $k \to \infty$. Indeed, as $k \to \infty$, the ratio

$$\frac{\Gamma(r + \frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} \sim \frac{k^r}{2^r}\,,$$

so that

$$\left(\frac{t^2}{t^2+k}\right)^r \frac{\Gamma(r + \frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} \sim \left(\frac{k}{t^2}\right)^{-r} \frac{k^r}{2^r} \sim \frac{t^{2r}}{2^r}\,.$$

Now, write the t-zeta function as

$$\zeta_k^T(t) = \sum_{r=1}^{\infty} \left(\frac{t^2}{t^2+k}\right)^r \frac{\Gamma(r + \frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} \frac{2^{2r}}{(2r)!} \frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} = \sum_{r=1}^{\infty} f_k(r)\,,$$

where $f_k(r)$ is an decreasing sequence of non-negative functions as, for all $k \geq 1$, $f_{k+1}(r) \leq f_k(r)$ for all $r \geq 1$. This sequence has a limit function,

$$\lim_{k\to\infty} f_k(r) = t^{2r} \frac{2^r}{(2r)!} \frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} = f(r)\,,$$

which is summable since

$$\sum_{r=1}^{\infty} f(r) = \sum_{r=1}^{\infty} t^{2r} \frac{2^r}{(2r)!} \frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} = \zeta(t)\,.$$

Because the convergence is monotone and $\sum_{r=1}^{\infty} f_1(r) < \infty$, we can conclude that

$$\lim_{k\to\infty} \zeta_k^T(t) = \sum_{r=1}^{\infty} \lim_{k\to\infty} f_k(r) = \zeta(t)\,.$$

3. The series

$$\sum_{r=1}^{\infty} \frac{2^r}{(2r)!} \frac{\Gamma(r + \frac{k+1}{2})}{\Gamma(\frac{k+1}{2})} \int x^{2r} e^{-x^2/2}\, H(dx)\,,$$

diverges since its r^{th} term

$$\begin{aligned} s_r &= \frac{2^r}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\int x^{2r}e^{-x^2/2}\,H(dx) \\ &= \frac{2^r}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{\Gamma(r-\alpha/2)}{2^{-r+\alpha/2}}\frac{\alpha\,2^{\alpha/2-1}}{\Gamma(1-\alpha/2)} \\ &= \frac{2^{2r}}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{\Gamma(r-\alpha/2)}{\Gamma(1-\alpha/2)}\frac{\alpha}{2} \\ &= s_{r-1}\,\frac{2^2}{(2r)(2r-1)}\left(r-1+\frac{k+1}{2}\right)(r-1-\alpha/2) \\ &= s_{r-1}\,b_r\,, \end{aligned}$$

where $b_r = \frac{2^2}{(2r)(2r-1)}\left(r-1+\frac{k+1}{2}\right)(r-1-\alpha/2) = \frac{(2r-1+k)(r-1-\alpha/2)}{(2r-1)r} \to 1$, as $r\to\infty$. Therefore, we have that $s_r \to s_{r-1}$ as $r\to\infty$, so that $\sum_{r=1}^{\infty} s_r = \infty$. Similarly to above, we can write the t-zeta function as

$$\zeta_k^T(t) = \sum_{r=1}^{\infty}\left(\frac{t^2}{t^2+k}\right)^r \frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\frac{2^{2r}}{(2r)!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} = \sum_{r=1}^{\infty} f_t(r)$$

where the sequence of non-negative functions $f_t(r)$ is such that $f_t(r) \le f_{t+1}(r)$. So by monotone convergence theorem, we conclude

$$\lim_{t\to\infty}\zeta_k^T(t) = \sum_{r=1}^{\infty}\lim_{t\to\infty} f_t(r) = \infty\,.$$

4. Using the Laplace approximation, we can investigate the behavior of the t-zeta function as $t\to\infty$. In fact,

$$\begin{aligned} \zeta_k^T(t) &= \frac{1}{t_k(t)}\int \phi(t\sqrt{u})\zeta(t\sqrt{u})\chi_k^2(ku)k\sqrt{u}\;du \\ &= \frac{1}{t_k(t)}\int\int \phi(t\sqrt{u})(\cosh(t\sqrt{u}\,x)-1)e^{-x^2/2}H(dx)\,\chi_k^2(ku)k\sqrt{u}\;du \\ &\sim \frac{1}{t_k(t)}\int\int \phi(t\sqrt{u})e^{t\sqrt{u}\,x}e^{-x^2/2}H(dx)\,\chi_k^2(ku)k\sqrt{u}\;du \\ &\sim \frac{K_\alpha}{t_k(t)}\int\int \phi(t\sqrt{u}-x)|x|^{-\alpha-1}dx\,\chi_k^2(ku)k\sqrt{u}\;du \end{aligned}$$

Now, the function $f(x) = \phi(t\sqrt{u} - x)|x|^{-\alpha-1}$ is maximized at approximately $x_0 = t\sqrt{u}$ with second derivative roughly given by 1. So for the inner integral

$$\int \phi(t\sqrt{u} - x)|x|^{-\alpha-1}dx\,,$$

the Laplace approximation gives

$$\frac{1}{\sqrt{2\pi}}|t\sqrt{u}|^{-\alpha-1}\sqrt{2\pi} = |t\sqrt{u}|^{-\alpha-1}\,.$$

Then, substituting this expression for the inner integral, one has that

$$\begin{aligned}
\zeta_k^T(t) &\sim \frac{K_\alpha}{t_k(t)}\int |t\sqrt{u}|^{-\alpha-1}\,\chi_k^2(ku)k\sqrt{u}\;du \\
&\sim \frac{K_\alpha}{t_k(t)}|t|^{-\alpha-1}\int u^{-\frac{\alpha+1}{2}}\,\chi_k^2(ku)k\sqrt{u}\;du \\
&\sim (t^2+k)^{\frac{k+1}{2}}\,\frac{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}{k^{\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right)}\,|t|^{-\alpha-1}\frac{\alpha\,2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}\int u^{-\frac{\alpha}{2}}\,k\frac{(1/2)^{k/2}}{\Gamma(k/2)}(ku)^{k/2-1}e^{-ku/2}\;du \\
&\sim (t^2+k)^{\frac{k+1}{2}}\,\frac{\sqrt{\pi}\,\Gamma\left(\frac{k}{2}\right)}{k^{\frac{k}{2}}\Gamma\left(\frac{k+1}{2}\right)}\,|t|^{-\alpha-1}\frac{\alpha\,2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}\,k^{k/2}\frac{(1/2)^{k/2}}{\Gamma(k/2)}\int u^{-\frac{\alpha}{2}+k/2-1}e^{-ku/2}\;du \\
&\sim (t^2+k)^{\frac{k+1}{2}}\,\frac{\sqrt{\pi}\,\Gamma\left(\frac{k}{2}\right)}{k^{\frac{k}{2}}\Gamma\left(\frac{k+1}{2}\right)}\,|t|^{-\alpha-1}\frac{\alpha\,2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}\,(k/2)^{k/2}\frac{1}{\Gamma(k/2)}\frac{\Gamma(\frac{k-\alpha}{2})}{(k/2)^{-\frac{\alpha+k}{2}}} \\
&\sim (t^2+k)^{\frac{k+1}{2}}\,\frac{\sqrt{\pi}}{k^{\frac{k}{2}}\Gamma\left(\frac{k+1}{2}\right)}\,|t|^{-\alpha-1}\frac{\alpha\,2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}\,\frac{\Gamma(\frac{k-\alpha}{2})}{(k/2)^{-\frac{\alpha}{2}}} \\
&\sim (t^2+k)^{\frac{k+1}{2}}\,\frac{\sqrt{\pi}}{k^{\frac{k}{2}-\frac{\alpha}{2}}\Gamma\left(\frac{k+1}{2}\right)}\,|t|^{-\alpha-1}\frac{\alpha\,2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}\,\frac{\Gamma(\frac{k-\alpha}{2})}{2^{\frac{\alpha}{2}}} \\
&\sim (t^2+k)^{\frac{k+1}{2}}|t|^{-\alpha-1}\,\frac{\sqrt{\pi}\,\Gamma\left(\frac{k-\alpha}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)k^{k/2-\alpha/2}}\frac{\alpha}{2\,\Gamma(1-\alpha/2)}.
\end{aligned}$$

5. Here we derive the sparse approximation to the symmetrized conditional distribution of μ/σ given $T = Y/s$. Let $Z = \eta/\sigma \sim N(0,1)$ and $W = ks^2/\sigma^2 \sim \chi_k^2$. The joint distribu-

tion of $(\mu/\sigma, T)$ is

$$
\begin{aligned}
&\mathbb{P}(\mu/\sigma \in dx,\, T \in dt) = \\
&\mathbb{P}\left(\mu/\sigma \in dx,\, \frac{\mu/\sigma + \eta/\sigma}{\sqrt{s^2/\sigma^2}} \in dt\right) = \\
&\int \mathbb{P}(\mu/\sigma \in dx,\, \frac{\mu/\sigma + \eta/\sigma}{\sqrt{W/k}} \in dt,\, W \in du)\, du = \\
&\int \mathbb{P}(\mu/\sigma \in dx,\, Z \in \sqrt{u/k}\, dt - x,\, W \in du)\, du = \\
&\int \mathbb{P}(\mu/\sigma \in dx)\, \mathbb{P}(Z \in \sqrt{u/k}\, dt - x)\, \mathbb{P}(W \in du)\, du = \\
&\int P_{\nu,\sigma}(dx) \sqrt{u/k} \phi(\sqrt{u/k}\, t - x) \chi^2(u)\, du = \\
&\int P_{\nu,\sigma}(dx) \sqrt{u/k} \phi(\sqrt{u/k}\, t - x) \frac{(1/2)^{k/2}}{\Gamma(k/2)} u^{k/2-1} e^{-u/2}\, du = \\
&\int P_{\nu,\sigma}(dx) \sqrt{v} \phi(\sqrt{v}\, t - x) \frac{(k/2)^{k/2}}{\Gamma(k/2)} v^{k/2-1} e^{-kv/2}\, dv = \\
&P_{\nu,\sigma}(dx) e^{-x^2/2} \frac{(k/2)^{k/2}}{\sqrt{2\pi}\Gamma(k/2)} \int e^{\sqrt{v}\, tx} v^{k/2+1/2-1} e^{-kv/2} e^{-vt^2/2}\, dv = \\
&P_{\nu,\sigma}(dx) e^{-x^2/2} t_k(t) \left(\frac{\Gamma(\frac{k+1}{2}) 2^{\frac{k+1}{2}}}{(t^2+k)^{\frac{k+1}{2}}}\right)^{-1} \int e^{\sqrt{v}\, tx} v^{k/2+1/2-1} e^{-kv/2} e^{-vt^2/2}\, dv\,.
\end{aligned}
$$

Because we are interested in the symmetrized distribution, we can substitute $\cosh(\sqrt{v}\, tx) = \frac{e^{\sqrt{v}\, tx} + e^{\sqrt{v}\,(-tx)}}{2}$ to $e^{\sqrt{v}\, tx}$, so the integral appearing in the last expression is

$$
\begin{aligned}
&\int \cosh(\sqrt{v}\, tx) v^{k/2+1/2-1} e^{-kv/2} e^{-vt^2/2}\, dv = \\
&\int (\cosh(\sqrt{v}\, tx) - 1) v^{k/2+1/2-1} e^{-kv/2} e^{-vt^2/2}\, dv + \int v^{k/2+1/2-1} e^{-kv/2} e^{-vt^2/2}\, dv = \\
&\sum_{r=1}^{\infty} \frac{(tx)^{2r}}{(2r)!} \int v^{r+k/2+1/2-1} e^{-v(t^2+k)/2}\, dv + \int v^{k/2+1/2-1} e^{-v(t^2+k)/2}\, dv = \\
&\sum_{r=1}^{\infty} \frac{(tx)^{2r}}{(2r)!} \frac{\Gamma(r + \frac{k+1}{2}) 2^{r+\frac{k+1}{2}}}{(t^2+k)^{r+\frac{k+1}{2}}} + \frac{\Gamma(\frac{k+1}{2}) 2^{\frac{k+1}{2}}}{(t^2+k)^{\frac{k+1}{2}}} = \\
&\frac{\Gamma(\frac{k+1}{2}) 2^{\frac{k+1}{2}}}{(t^2+k)^{\frac{k+1}{2}}} \left(\sum_{r=1}^{\infty} \frac{(tx)^{2r}}{(2r)!} \frac{\Gamma(r + \frac{k+1}{2}) 2^r}{\Gamma(\frac{k+1}{2})(t^2+k)^r} + 1\right)\,.
\end{aligned}
$$

Plugging this expression back in the former derivation, one obtains

$$\begin{aligned}
&\mathbb{P}(\mu/\sigma \in dx,\, |T| \in dt) = \\
&P_{\nu,\sigma}(dx)e^{-x^2/2}t_k(t)\left(\frac{\Gamma(\frac{k+1}{2})2^{\frac{k+1}{2}}}{(t^2+k)^{\frac{k+1}{2}}}\right)^{-1}\int \cosh(\sqrt{v}\,tx)v^{k/2+1/2-1}e^{-kv/2}e^{-vt^2/2}\,dv = \\
&P_{\nu,\sigma}(dx)e^{-x^2/2}t_k(t)\left(\frac{\Gamma(\frac{k+1}{2})2^{\frac{k+1}{2}}}{(t^2+k)^{\frac{k+1}{2}}}\right)^{-1}\frac{\Gamma(\frac{k+1}{2})2^{\frac{k+1}{2}}}{(t^2+k)^{\frac{k+1}{2}}}\left(\sum_{r=1}^{\infty}\frac{(tx)^{2r}}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})2^r}{\Gamma(\frac{k+1}{2})(t^2+k)^r}+1\right) = \\
&P_{\nu,\sigma}(dx)e^{-x^2/2}t_k(t)\left(\sum_{r=1}^{\infty}\frac{(tx)^{2r}}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})2^r}{\Gamma(\frac{k+1}{2})(t^2+k)^r}+1\right).
\end{aligned}$$

This last measure, as $\nu \to 0$, in terms of $\mathcal{W}^{\#}$-integrals, is equivalent to

$$t_k(t)\left(e^{-x^2/2}P_{\nu,\sigma}(dx)+\tilde{\rho}e^{-x^2/2}H(dx)\sum_{r=1}^{\infty}\frac{(tx)^{2r}}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})2^r}{\Gamma(\frac{k+1}{2})(t^2+k)^r}\right).$$

Then to obtain the symmetrized conditional distribution, it suffices to divide by the sparse approximation to the marginal of $|T|$:

$$\frac{e^{-x^2/2}P_{\nu,\sigma}(dx)+\tilde{\rho}\,\sum_{r=1}^{\infty}\frac{t^{2r}}{(t^2+k)^r}\frac{2^r}{(2r)!}\frac{\Gamma(r+\frac{k+1}{2})}{\Gamma(\frac{k+1}{2})}\,x^{2r}e^{-x^2/2}H(dx)}{1-\tilde{\rho}+\tilde{\rho}\zeta_k^T(t)},$$

so that one can write

$$P_{\nu,\sigma}(\mu/\sigma \in dx \mid |T|) = \frac{1-\tilde{\rho}}{1-\tilde{\rho}+\tilde{\rho}\zeta_k^T(t)}\frac{e^{-x^2/2}P_{\nu,\sigma}(dx)}{1-\tilde{\rho}}+\frac{\tilde{\rho}\zeta_k^T(t)}{1-\tilde{\rho}+\tilde{\rho}\zeta_k^T(t)}\frac{\zeta_k^T(dx;t)}{\zeta_k^T(t)}+o(\tilde{\rho}).$$

Chapter 2

Sparsity scales for large covariance matrices

2.1 Introduction

Given n independent copies of a random p-vector $(X_1, \ldots, X_p)$, the estimation of the covariance matrix Σ and of its inverse, the precision matrix Ω, is central to multivariate analysis, being a useful device to summarize the linear relationships between the variables. Estimating Σ or Ω represents a crucial step in many statistical methods, such as linear discriminant analysis and principal component analysis, among many others.

However, in high-dimensional settings, the estimation of Σ and Ω becomes very challenging. Indeed, when the dimension p is larger than the sample size n, the sample covariance matrix $S_n = \frac{1}{n}XX'$ is not invertible and, despite being entry-wise consistent, is not consistent in other metrics, such as the spectral norm, which are more useful in practical work. Many results in random matrix theory, starting from the classical Marčenko-Pastur law, illustrate the dramatic change of behaviour of S_n under the 'large p, large n' asymptotic assumption. For a thorough introduction to these ideas, see for instance El Karoui (2008) [32]. In order to progress with inference and obtain a better estimator for Σ, some sparsity assumptions

need to be made.

Within the immense literature on estimation of sparse covariance matrices when p is growing at the same rate or even faster than n, the novelty of Battey (2019) [8] is to not assume that Σ is necessarily sparse on the original scale or on the inverse scale. What is assumed is that there exists a transformation of Σ such that the transformed matrix is in fact sparse. Provided that one can find such transformation, the paper investigates under what conditions it is possible to exploit the sparsity on the transformed scale to deduce spectral-norm convergence results on the original scale.

Yet, in that paper, little detail is given on how to identify such 'sparsity scale', which is in fact assumed to have been already determined. This chapter tackles the problem of how to possibly identify a sparsity-inducing transformation. The idea is to exploit the sparsity framework described in the introduction to provide a heuristic (data-dependent) guidance for the choice of the sparsity scale. So differently from Battey (2019) [8], where sparsity is an assumption made on deterministic matrices, we consider a matrix to be sparse if its off-diagonal entries can be seen as random sparse signals. The sparsity rate, estimated from the data under this assumption, is then used as a likelihood-based criterion for comparing the level of sparsity induced by different matrix transformations.

2.2 Brief review of Battey (2019)

In this section, we briefly review some of the main ideas presented in Battey (2019) [8] (B. henceforth). For clarity, we adopt the same notation as that paper. Let $\mathcal{V}_p^+$ be the space of symmetric positive-definite $p \times p$ matrices and $\mathcal{V}_p$ a general space of $p \times p$ matrices. The Frobenius, component-wise and spectral norm of a matrix A are defined as $||A||_{\mathrm{F}}^2 = \sum_{u,v} |a_{uv}|^2$, $||A||_{\max} = \max_{u,v} |a_{u,v}|$ and $||A||_{\mathrm{op}} = |\lambda_1|$ respectively, where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of A, ordered in absolute magnitude. The covariance matrix of the random vector $(X_1, \ldots, X_p)$, having zero mean, is denoted by $\Sigma^\star$, and $\lambda_1^\star, \ldots \lambda_p^\star$ denote its ordered

eigenvalues. Moreover, the symbol $\asymp$ means equality in order of magnitude, while writing $Z = O_{\Pr}(t_{n,p})$ means that

$$\mathbb{P}(Z \leq C_0 t_{n,p}) \geq 1 - \epsilon_{n,p}\,,$$

where C_0 is some positive constant, and $\epsilon_{n,p}$ is a deterministic sequence converging to zero as $n, p \to \infty$ such that $t_{n,p} \to 0$.

The motivation of the paper is to find conditions under which an estimator $\tilde{\Sigma}_n \in \mathcal{V}_p$ is consistent for $\Sigma^\star \in \mathcal{V}_p^+$ in spectral norm, under the double asymptotic regime $p, n \to \infty$, where p grows at least as fast as n. This means that the $p \times p$ covariance matrix $\Sigma^\star$ is also converging to an operator $\Sigma^\star_\infty : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$.

We now outline the necessary steps to identify such conditions, exploiting a sparsity-inducing transformation f to apply to $\Sigma^\star$. Before proceeding though, we would like to emphasize that in B., the population covariance matrix $\Sigma^\star$, and any transformation of it, are considered fixed unknown parameters. So when using the term convergence in this section, we refer to convergence in probability of some sample estimator to some unknown deterministic matrix. Moreover, a symmetric positive-definite matrix is defined to be sparse if all of its rows belong to a sufficiently small l_q-ball around zero; more formally, if it belongs to the class of matrices

$$\mathcal{F}(s(p), q) = \{A \in \mathcal{V}_p^+ : \max_u |A_{uu}| \leq C\,,\ \max_u \sum_{v=1}^{p} |A_{uv}|^q \leq s(p)\}\,, \tag{2.1}$$

where $s(p)/p \to 0$ as $p \to \infty$ and $q \in [0, 1]$ while C is just some positive constant.

The idea of B. hinges upon the fact that, if the transformed matrix $F^\star = f(\Sigma^\star) \in \mathcal{V}_p$ is sparse in the sense of (2.1), then on this sparsity scale, provided that one can find a pilot

estimator $\hat{F}_n^P \in \mathcal{V}_p$ converging to $F^\star$ component-wise, then the hard-thresholding estimator

$$\mathcal{T}(\hat{F}_n^P) = \hat{F}_{n,uv}^P \mathbf{1}(|\hat{F}_{n,uv}^P| > \tau)\,, \tag{2.2}$$

converges to $F^\star$ in spectral norm. In fact, Bickel and Levina (2008) [12] showed that, for data $(X_1, \ldots, X_p)$ drawn from a distribution with Gaussian or sub-Gaussian tails, thresholding the sample covariance matrix leads to an estimator which is consistent for the covariance matrix in spectral norm, when both dimensions of the matrices, n and p, go to infinity in a regime such that $\log(p)/n \to 0$. Avella-Medina et al. (2018) [4] generalized this result to less stringent requirements on the tails of the distribution of the data, by considering some robust pilot estimators other than the sample covariance matrix.

There are two main steps necessary to convert this spectral norm convergence on the sparsity scale $f(\Sigma^\star)$, to the spectral norm convergence on the original scale $\Sigma^\star$. The first step concerns under what conditions, when $p > n$, given a positive-definite estimator $\hat{\Sigma}_n^P$ converging to $\Sigma^\star$ component-wise, the transformed estimator $\hat{F}_n^P = f(\hat{\Sigma}_n^P)$ also converges to the transformed matrix $F^\star = f(\Sigma^\star)$ component-wise. The second step, on the other hand, has to do with the opposite direction of the transformation and aims at finding conditions under which the spectral-norm convergence on the transformed scale, $||\mathcal{T}(\hat{F}_n^P) - F^\star||_{\text{op}} \to 0$, implies the same kind of convergence on the original scale, $||f^{-1}\{\mathcal{T}(\hat{F}_n^P)\} - f^{-1}(F^\star)||_{\text{op}} \to 0$, as $n, p \to \infty$.

The conditions for the first step are given in Theorem 1 of B., which we restate here for completeness.

Theorem 1: Suppose that as $p \to \infty$, the sequence of smallest eigenvalue $\lambda_p^\star$ of $\Sigma^\star$ is bounded away from zero. Let $\hat{\Sigma}_n^P$ be a symmetric matrix with ordered eigenvalues $\lambda_1, \ldots, \lambda_p$, where as

$p \to \infty$, the sequence of smallest eigenvalue λ_p is bounded away from zero. Let $\hat{F}_n^P = f(\hat{\Sigma}_n^P)$. Then, as $n, p \to \infty$, $\hat{F}_n^P$ is component-wise consistent for $F^\star$, if $\hat{\Sigma}_n^P$ is component-wise consistent for $\Sigma^\star$, and

$$\mathrm{Res}\Big\{\frac{f(z)}{(z-\hat{\lambda}_k)(z-\lambda_\nu^\star)}, \hat{\lambda}_k\Big\} + \mathrm{Res}\Big\{\frac{f(z)}{(z-\hat{\lambda}_k)(z-\lambda_\nu^\star)}, \lambda_\nu^\star\Big\}, \;\; z \in \mathbb{C}\,,$$

is bounded in absolute value for all $1 \leq k, \nu \leq p$ as $n, p \to \infty$, where, for a function g of a complex variable, $\mathrm{Res}(g, a)$ denotes the residue of g at a. If this condition is satisfied, then $||\hat{F}_n^P - F^\star||_{\max} \asymp ||\hat{\Sigma}_n^P - \Sigma^\star||_{\max}$.

So now, suppose that the pilot estimator $\hat{\Sigma}_n^P$ of $\Sigma^\star$ is component-wise consistent, so that $||\hat{\Sigma}_n^P - \Sigma^\star||_{\max} = O_{\Pr}(r_{n,p})$, where $r_{n,p} \to 0$ as $n, p \to \infty$. Usually this latter condition implies that $\log(p)/n \to 0$. Further suppose that a transformation f satisfying the assumptions in Theorem 1 has been found and is such that the transformed matrix $F^\star = f(\Sigma^\star)$ belongs to the class of matrices in (2.1). So we have that the transformed pilot estimator $\hat{F}_n^P = f(\hat{\Sigma}_n^P)$ is component-wise consistent for the sparse $F^\star$, with $||\hat{F}_n^P - F^\star||_{\max} = O_{\Pr}(r_{n,p})$. Then, Corollary 3 in B. states that as $n, p \to \infty$, the thresholded estimator $\mathcal{T}(\hat{F}_n^P)$, is consistent for $F^\star$ in spectral norm, $||\mathcal{T}(\hat{F}_n^P) - F^\star||_{\mathrm{op}} = O_{\Pr}(s(p) r_{n,p}^{1-q})$. Here $\mathcal{T}(\hat{F}_n^P)$ is defined as in (2.2), with the threshold τ chosen in such a way that $\tau \asymp r_{n,p}$, where $r_{n,p}$ is the rate characterizing the component-wise convergence of the pilot estimator on the original scale $||\hat{\Sigma}_n^P - \Sigma^\star||_{\max} = O_{\Pr}(r_{n,p})$.

At this point, however, it is not an easy task to establish general conditions under which the spectral-norm convergence on the transformed scale implies the same kind of convergence on the original scale. In the case where f^{-1} has a bounded first derivative then, for some

constant positive c, one has

$$||f^{-1}(F^\star) - f^{-1}\{\mathcal{T}(\hat{F}_n^P)\}||_{\text{op}} \leq c\,||(f^{-1})'||_\infty ||F^\star - \mathcal{T}(\hat{F}_n^P)||_{\text{op}}\,,$$

where $||g||_\infty = \sup_x |g(x)|$ and g' denotes the first derivative of g. Yet, in general the inverse of f does not have a bounded first derivative and *ad hoc* justifications need to be provided.

For instance, Theorem 2 in B. considers the log transform and it implies that, if the sequences of maximum and minimum eigenvalues of $L^\star = \log \Sigma^\star$ are bounded as $p \to \infty$, then, letting $\tilde{L} = \mathcal{T}(\log \hat{\Sigma}_n^P)$, one has that both $\tilde{\Sigma} = \exp(\tilde{L})$ and $\tilde{\Omega} = \exp(-\tilde{L})$, converge in spectral norm to $\Sigma^\star$ and $(\Sigma^\star)^{-1}$ respectively, as $n, p \to \infty$.

2.3 Estimating the sparsity scale

In the introduction of B. it is suggested that a likelihood-based criterion could be used to estimate the sparsity scale associated to a given transformation. However this idea is not further developed in the paper, and the estimation theory summarized in the previous section is established under the assumption that a suitable sparsity scale has already been determined. In this section we propose a heuristic strategy for finding such sparsity-inducing transformation f. In doing so, we take a completely different approach in which the unknown matrix $F^\star$ has random off-diagonal entries having a sparse distribution P_ν. Within this framework, given a pilot estimator $\hat{\Sigma}_n^P$ of $\Sigma^\star$, we consider some parametric family of scalar-valued transformations, which can be applied to the eigenvalues of $\hat{\Sigma}_n^P$ to obtain a family of transformed pilot estimators, $\hat{F}_n^P = f(\hat{\Sigma}_n^P)$. Given this family of transformed matrices, we then construct a likelihood-based criterion to select one scalar-valued transformation, out of the parametric family, depending on the estimated level of sparsity characterizing each transformed matrix $F^\star = f(\Sigma^\star)$.

More precisely, first we define a transformation f to be sparsity-inducing if the off-diagonal entries of the transformed matrix $F^\star = f(\Sigma^\star)$ can be considered to be independent random variables with a common sparse distribution P_ν. Then the level of sparsity induced by f is defined to be the sparsity rate ρ characterizing the convergence of P_ν to the Dirac delta measure at zero.

Second, we assume that the transformation of the observed pilot estimator $\hat{\Sigma}_n^P$, $\hat{F}_n^P = f(\hat{\Sigma}_n^P)$, is such that each off-diagonal entry $\hat{F}_{n,uv}^P$ is the scaled convolution of the corresponding entry $F_{uv}^\star$ with an independent Gaussian noise η_{uv}. In formulae,

$$\hat{F}_{n,uv}^P = \sigma(F_{uv}^\star + \eta_{uv})\,,$$

where σ is an unknown scale parameter, which allows us to obtain the same likelihood estimates for the sparsity parameters, under any arbitrary change of scale of the observed entries of $\hat{F}_n^P$. We discuss this fact more at the end of this section.

Then, following the signal-plus-noise model presented in the introduction, the sparse approximation to the marginal density of each off-diagonal entry $\hat{F}_{n,uv}^P$ at f is

$$m_\nu(f) = \tfrac{1}{\sigma}\phi(f/\sigma)(1-\rho+\rho\zeta(f/\sigma)) + o(\rho)\,.$$

Here we consider the exceedance measure of P_ν to be an inverse-power measure indexed by $\alpha \in (0,2)$. In this case, denoting by $\{\hat{f}_{n,uv}^P\,, u<v\}$, the observed off-diagonal entries of $\hat{F}_n^P$, the sparse log likelihood for the triplet (ρ,σ,α) can be written as

$$\log L(\rho,\sigma,\alpha;\hat{f}_{n,uv}^P\,, u<v) = \sum_{u<v}\log\left(\tfrac{1}{\sigma}\phi(\hat{f}_{n,uv}^P/\sigma)\right) + \sum_{u<v}\log\left(1-\rho+\rho\zeta_\alpha(\hat{f}_{n,uv}^P/\sigma)\right), \quad (2.3)$$

where the subscript in ζ_α aims at highlighting the dependence of the zeta function on the α

parameter of the inverse-power exceedance measure. Maximizing (2.3) gives the maximum likelihood estimate $(\hat{\rho}, \hat{\sigma}, \hat{\alpha})$ associated to the transformation f. The estimated level of sparsity induced by f is $\hat{\rho}$ and this can be used to compare and select the scale of 'maximal sparsity', i.e., the transformation f having the smallest estimated sparsity rate.

As in B., we define a transformation acting on a square matrix A by specifying a scalar transformation to apply to each of its eigenvalues, while holding its eigenvectors fixed. For example, the log transform would lead to a transformed matrix having the same eigenvectors of A and the eigenvalues on the log scale. In general, given a one-dimensional transformation f and the spectral decomposition $A = Q\Lambda Q'$, the corresponding matrix transformation simply is

$$f : A \mapsto Q f(\Lambda) Q' \,,$$

where $f(\Lambda)$ is the diagonal matrix of the transformed eigenvalues $f(\lambda_i)$, $i = 1, \ldots, p$.

As far as the selection of the transformation is concerned, we only consider parametric families of transformations, such as the family of power transformations

$$f_\beta : y \mapsto y^\beta \,, \quad \beta \in \mathbb{R} \,, \tag{2.4}$$

and the Box-Cox family

$$f_\beta : y \mapsto (y^\beta - 1)/\beta \,, \quad \beta \in \mathbb{R} \,, \tag{2.5}$$

where, in both cases, the limit $\beta \to 0$ corresponds to the log transform. Both families in (2.4) and (2.5) are parametrized by $\beta \in \mathbb{R}$. In practice however, we only consider a finite grid of possible values for β over a prespecified finite range $\mathcal{R}$.

Notice that with both families of transformations, the estimation of the sparsity param-

eters ρ and α is invariant under scalar multiplication of the observed matrix. In fact, if $A = Q\Lambda Q'$ is multiplied by $r > 0$, then $A_r = Qr\Lambda Q'$ and $f_\beta(A_r) = Q\left(r^\beta f_\beta(\Lambda) + c\right)Q' = r^\beta Q f_\beta(\Lambda)Q' + cI$, where I is the identity matrix and $c = 0$ for the power transform while $c = (r^\beta - 1)/\beta$ for the Box-Cox transform. Since in (2.3) we are estimating (ρ, σ, α) only from the transformed off-diagonal entries, the estimation based on $\{f_\beta(r\hat{\Sigma}_n^P)\}_{u>v}$ is equivalent to that based on $\{r^\beta f_\beta(\hat{\Sigma}_n^P)\}_{u>v}$, so the only parameter that gets affected by the scaling operation is indeed the scale parameter σ. Here $\{A\}_{u>v}$ denotes the set of the lower-diagonal entries of A. It is easy to show that the scale parameter estimated from $\{f_\beta(r\hat{\Sigma}_n^P)\}_{u>v}$ is r^β times the scale parameter estimated from $\{f_\beta(\hat{\Sigma}_n^P)\}_{u>v}$.

2.4 Simulation study

To illustrate how the heuristic method described in the previous section works, we mimic the simulation study presented in Section 4.1 of B., where $\Sigma^\star \in \mathcal{V}_p^+$ is constructed to be sparse, in the sense of (2.1), on the logarithmic scale. More precisely, $\Sigma^\star = \exp(L^\star)$ and the matrix $L^\star$ is constructed as

$$L^\star = \sum_{m=1}^{|\mathcal{B}|} \alpha_m^\star B_m \,,$$

with matrices $B_m \in \mathcal{B}$. Here $\mathcal{B}$ denotes the natural symmetrized basis for the space of symmetric matrices which is the union of

$$\mathcal{B}_1 = \{B \,:\, B = e_j e_j',\, j \in [p]\}\,,$$

and

$$\mathcal{B}_2 = \{B \,:\, B = e_j e_k' + e_k e_j',\, j,k \in [p],\, j \neq k\}\,,$$

where e_j denotes the j^{th} vector of the canonical basis of $\mathbb{R}^p$, with $p = 200$. The coefficient vector $\alpha^\star$ has a support of non-zero entries which is randomly sampled from the index set $[p(p+1)/2]$ and has size $K = 60$. Given the non-zero support, half of these components are

uniformly drawn from $[0.5, 1]$ and the other half from $[-1, -0.5]$. The resulting $L^\star$ belongs to the class $\mathcal{F}(s(p), q)$ with $q = 0$.

For $n = 20, 60, 100, 140$, we simulate n independent copies of a p-dimensional random vector having distribution $N_p(0, \Sigma^\star)$, and we construct the pilot estimator $\hat{\Sigma}_n^P$ as

$$\hat{\Sigma}_n^P = (1 - \delta)S_n + \delta \hat{D}_n \,,$$

which is a convex combination of the sample covariance matrix $S_n = \frac{1}{n} XX'$ and $\hat{D}_n = \text{diag}(S_n)$. Here $\delta = ((\log p)/n)^{1/2}$. This pilot estimator of $\Sigma^\star$, proposed in Proposition 2 of B., is guaranteed to have its smallest eigenvalue pulled away from zero.

Following Section 2.3, we consider the two families of transformations (2.4) and (2.5) with values of β ranging from -1 to 1 and, after applying each transformation f_β to $\hat{\Sigma}_n^P$, giving $\hat{F}_{n,\beta}^P = f_\beta(\hat{\Sigma}_n^P)$, we obtain the maximum likelihood estimate of the triplet (ρ, σ, α), corresponding to each transformation f_β. The estimation results are reported from Table 2.1 to Table 2.4. The minimum of $\hat{\rho}$ is not always reached at $\beta = 0$ so, if we were to choose the transformation $f_{\beta^\star}$ based on the estimated sparsity rate, we would not necessarily select the log transform. In fact, since the sparsity assumptions of Section 2.3 on $f(\Sigma^\star)$ are somewhat more general than the sparsity structure imposed on $\log \Sigma^\star$ in its construction, there is no compelling reason why we should expect the log transform to be selected by our likelihood procedure. Yet, in a similar fashion of how the Box-Cox transformation is chosen in the linear regression setting, given the results of Theorem 2 in B., one might still prefer to opt for the log transform, if the difference in the estimated sparsity rates is not dramatic. Figure 2.1 shows the estimated pair $(\hat{\rho}, \hat{\sigma})$ for the two different families, black curves for the power transform and red curves for the Box-Cox transform. The Box-Cox transformation appears to correct for the discontinuity at zero, which instead can be easily detected in the

curves corresponding to the power transform.

Table 2.5 reports the performance of five different estimators of $\Sigma^\star$, averaged over 100 Monte Carlo replications. The performance of each estimator matrix E, is assessed in terms of relative errors in both spectral and Frobenius norm. The estimators considered are: the sample covariance matrix S_n, its thresholded version $\mathcal{T}(S_n)$, the pilot estimator $\hat{\Sigma}_n^P$ and two transformed, thresholded and transformed-back versions of $\hat{\Sigma}_n^P$. Column 4 refers to $\hat{\Sigma}_{\log}^P = \exp\{\mathcal{T}(\log(\hat{\Sigma}_n^P))\}$, while column 5 reports results for $\hat{\Sigma}_{\beta^\star}^P = f^{-1}\{\mathcal{T}(f(\hat{\Sigma}_n^P))\}$, with $f = f_{\beta^\star}$ corresponding to the Box-Cox transformation with smallest estimated sparsity rate. The hard-thresholding procedure is carried out following (2.2), and choosing $\tau = c\,((\log p)/n)^{1/2}$, with $c = 1$ on both original (column 2) and transformed scales (column 4 and 5). The advantage of thresholding the pilot estimator on a transformed scale is evident for all sample sizes, since the relative errors displayed in column 4 and 5 are always smaller than those of the other three estimators. Moreover, except for the case $n = 60$, choosing the transformation with smallest estimated sparsity rate delivers even better results than the log transform.

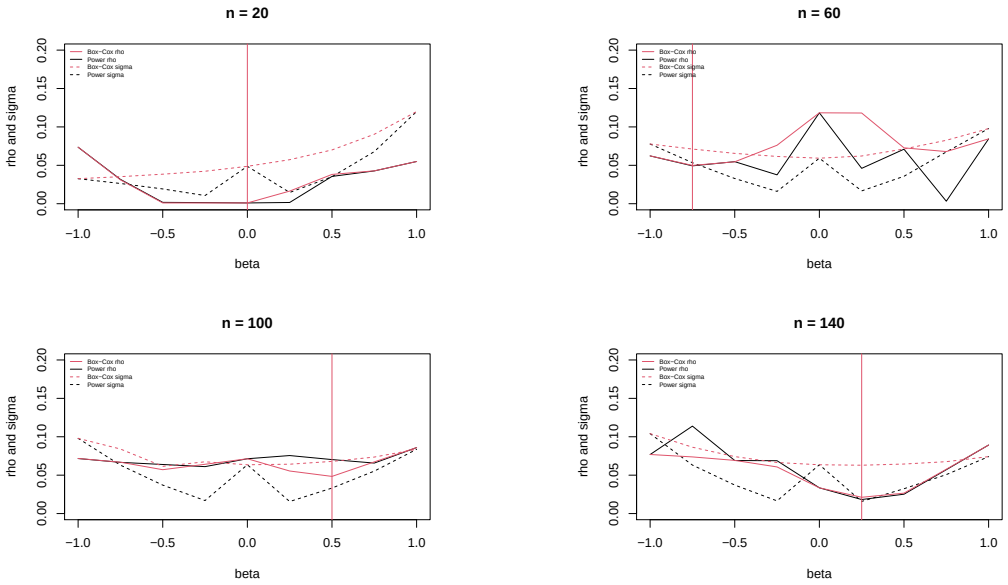


Figure 2.1: Plots of the estimated parameters ρ and σ associated to two different families of parametric transformations: the power transform $\lambda \to \lambda^\beta$ (black lines) and the Box-Cox transform $\lambda \to (\lambda^\beta - 1)/\beta$ (red lines). Here β ranges from -1, corresponding to the inverse transform, to 1, the identity transform, while $\beta = 0$ corresponds to the log transform. The vertical red line indicates the value of β with the lowest estimated ρ within the Box-Cox family.

(a) Power transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	7.36	3.26	1.58
-0.75	3.17	2.64	1.75
-0.50	0.18	1.92	1.85
-0.25	0.14	1.06	1.56
0.00	0.10	4.87	1.78
0.25	0.16	1.47	1.62
0.50	3.56	3.51	1.76
0.75	4.27	6.78	1.67
1.00	5.50	12.0	1.60

(b) Box-Cox transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	7.37	3.26	1.58
-0.75	3.11	3.52	1.78
-0.50	0.11	3.84	1.82
-0.25	0.10	4.23	1.81
0.00	0.10	4.87	1.78
0.25	1.64	5.73	1.73
0.50	3.81	7.00	1.77
0.75	4.28	9.04	1.67
1.00	5.50	12.0	1.60

Table 2.1: Maximum likelihood estimates when $n = 20$, $p = 200$, $\delta = 0.51$.

(a) Power transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	6.23	7.77	1.55
-0.75	4.93	5.34	1.61
-0.50	5.48	3.28	1.66
-0.25	3.76	1.60	1.58
0.00	11.84	5.93	1.79
0.25	4.60	1.67	1.62
0.50	7.10	3.57	1.66
0.75	0.33	6.74	0.10
1.00	8.44	9.78	1.50

(b) Box-Cox transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	6.23	7.77	1.55
-0.75	4.99	7.12	1.62
-0.50	5.46	6.55	1.65
-0.25	7.61	6.16	1.72
0.00	11.84	5.93	1.79
0.25	11.81	6.21	1.78
0.50	7.27	7.13	1.66
0.75	6.79	8.26	1.56
1.00	8.44	9.78	1.50

Table 2.2: Maximum likelihood estimates when $n = 60$, $p = 200$, $\delta = 0.3$.

(a) Power transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	7.16	9.80	1.53
-0.75	6.67	6.30	1.57
-0.50	6.39	3.70	1.59
-0.25	6.11	1.69	1.62
0.00	7.14	6.35	1.65
0.25	7.55	1.57	1.53
0.50	7.03	3.32	1.60
0.75	6.57	5.53	1.51
1.00	8.59	8.36	1.48

(b) Box-Cox transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	7.18	9.79	1.53
-0.75	6.69	8.40	1.57
-0.50	5.71	6.15	1.55
-0.25	6.42	6.74	1.62
0.00	7.14	6.35	1.65
0.25	5.54	6.43	1.58
0.50	4.84	6.79	1.50
0.75	6.73	7.36	1.52
1.00	8.59	8.36	1.48

Table 2.3: Maximum likelihood estimates when $n = 100$, $p = 200$, $\delta = 0.23$.

(a) Power transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	7.68	10.41	1.51
-0.75	11.39	6.31	1.78
-0.50	6.91	3.71	1.56
-0.25	6.88	1.65	1.60
0.00	3.35	6.35	1.40
0.25	1.83	1.58	1.17
0.50	2.52	3.23	1.24
0.75	5.71	5.07	1.43
1.00	8.92	7.41	1.43

(b) Box-Cox transform

β	$\hat{\rho} \cdot 10^2$	$\hat{\sigma} \cdot 10^2$	$\hat{\alpha}$
-1.00	7.69	10.41	1.52
-0.75	7.38	8.64	1.54
-0.50	6.92	7.42	1.56
-0.25	6.07	6.65	1.56
0.00	3.35	6.35	1.40
0.25	2.12	6.29	1.22
0.50	2.66	6.45	1.26
0.75	5.77	6.76	1.43
1.00	8.92	7.41	1.43

Table 2.4: Maximum likelihood estimates when $n = 140$, $p = 200$, $\delta = 0.19$.

n	Error	S_n	$\mathcal{T}(S_n)$	$\hat{\Sigma}_n^P$	$\hat{\Sigma}_{\log}^P$	$\hat{\Sigma}_{\beta^\star}^P$
20	$\|\Sigma^* - E\|_{\text{op}} / \|\Sigma^*\|_{\text{op}}$	340.36	198.71	160.36	78.52	78.52
		(29.42)	(32.79)	(13.63)	(4.14)	(4.14)
20	$\|\Sigma^* - E\|_{\text{F}} / \|\Sigma^*\|_{\text{F}}$	255.04	157.80	129.00	63.55	63.55
		(7.82)	(10.96)	(3.37)	(0.64)	(0.64)
60	$\|\Sigma^* - E\|_{\text{op}} / \|\Sigma^*\|_{\text{op}}$	157.14	93.55	106.04	70.34	86.42
		(11.12)	(11.59)	(7.15)	(5.24)	(1.33)
60	$\|\Sigma^* - E\|_{\text{F}} / \|\Sigma^*\|_{\text{F}}$	147.49	89.09	105.50	55.51	71.67
		(2.56)	((3.53)	(1.61)	(1.45)	(0.55)
100	$\|\Sigma^* - E\|_{\text{op}} / \|\Sigma^*\|_{\text{op}}$	111.59	67.98	81.94	55.90	42.66
		(8.97)	(9.17)	(6.28)	(4.74)	(4.97)
100	$\|\Sigma^* - E\|_{\text{F}} / \|\Sigma^*\|_{\text{F}}$	114.34	68.85	89.22	46.12	32.98
		(1.58)	(2.34)	(1.04)	(1.21)	(1.36)
140	$\|\Sigma^* - E\|_{\text{op}} / \|\Sigma^*\|_{\text{op}}$	90.12	55.45	68.98	49.17	42.93
		(6.01)	(6.45)	(4.45)	(4.33)	(4.50)
140	$\|\Sigma^* - E\|_{\text{F}} / \|\Sigma^*\|_{\text{F}}$	96.66	58.07	78.79	40.45	34.09
		(1.34)	(1.97)	(0.96)	(1.19)	(1.11)

Table 2.5: Comparison of estimators for $\Sigma^\star$. Relative error in spectral and Frobenius norm, averaged over 100 simulations. Standard errors shown in parenthesis.

Chapter 3

Higher-order sparse integral expansion

3.1 Introduction

As explained in the introduction, in McC&P, the definition of a sparse sequence of distributions $\{P_\nu\}_\nu$ identifies the pair (ρ, H) to be the rate function $\rho = \rho_\nu$ and the exceedance measure $H(dx)$ such that

$$\lim_{\nu \to 0} \rho_\nu^{-1} \int w(x)\, P_\nu(dx) = \int w(x)\, H(dx)\,,$$

for any function w in the class

$$\mathcal{W}^{\#} = \{w : \mathbb{R} \to \mathbb{R} \text{ s.t. } w(x) \text{ and } x^{-2} w(x) \text{ is bounded and continuous}\}\,.$$

Adopting an operator notation, another way of writing this definition of sparsity is

$$P_\nu(w) = \rho H(w) + o(\rho)\,,$$

where $F : \mathcal{W}^{\#} \to \mathbb{R}$

$$F(g) = \int g(x)F(dx)$$

is the operator form of the measure F acting on the function $g \in \mathcal{W}^{\#}$. With this notation, it is easy to see how the sparsity definition is constructing the $\mathcal{W}^{\#}$-integral asymptotic expansion of P_ν, driven by $\nu \to 0$, and the pair (ρ, H) characterizes the first term in this expansion. From this point of view, a reasonable question to ask is whether it is possible to find subsequent pairs (ρ_2, H_2), (ρ_3, H_3), and so on, for some sequence of decreasing rates $\{\rho_j\}_{j\geq 2}$, such that

$$P_\nu(w) = \rho H(w) + \rho_2 H_2(w) + \rho_3 H_3(w) + \cdots + \rho_k H_k(w) + o(\rho_k)\,.$$

In this context, we let H_k, $k \geq 2$, be any linear functional defined on the space of functions $\mathcal{W}^{\#}$. For this reason, we will refer to $H_k(w)$ as the k^{th}-order exceedance functional.

A related question arising from this higher-order integral expansion concerns equivalence relations. Indeed, as highlighted in McC&P, it is possible that two sparse families have the same first-order pair (ρ, H). For instance, the scaled Cauchy and the scaled horseshoe families are first-order equivalent in the sparse limit, since they share the inverse-power exceedance measure, and their sparsity parameters can be adjusted so to match the two rates. Therefore, it comes natural to ask whether the equivalence carries over to second and higher orders or whether, and at what point, this equivalence ceases to exist.

In the next sections, we examine different classes of sparse families separately: the measures having a density function which is analytic in the sparsity parameter; the scale sparse measures; the mixture measures. For each of these, we illustrate how to identify the terms beyond the first one, in their sparsity integral expansion. We conclude by giving two examples, one when the first-order equivalence relation carries over to higher orders, and one

when it does not.

3.2 Analytic density with Lévy integrable coefficients

When the sparse measure has a density $p_\nu(x)$ which is an analytic function of the sparsity parameter ν, one can write the density function using its Taylor series at $\nu = 0$,

$$p_\nu(x) = \sum_{k=0}^{\infty} \frac{\partial^k}{\partial \nu^k} p_\nu(x)\Big|_{\nu=0} \frac{\nu^k}{k!} .$$

Given this expansion, it is natural to identify subsequent sparsity rates as the increasing powers of ν, while the coefficients $u_k(x) = \frac{\partial}{\partial \nu} p_\nu(x)\Big|_{\nu=0}$ constitute the measures defining the corresponding exceedance functionals

$$H_k(w) = \int w(x) u_k(x)\, dx .$$

A case in point is the double gamma family

$$P_\nu(dx) = \frac{|x|^{\nu-1} e^{-|x|}}{2\Gamma(\nu)} dx .$$

In fact, the density function can be written as a convergent power series in the sparsity parameter ν,

$$P_\nu(dx) = \sum_{k=0}^{\infty} u_k(x) \frac{\nu^k}{k!} , \tag{3.1}$$

where

$$u_k(x) = |x|^{-1} e^{-|x|} \sum_{j=0}^{k-1} \frac{k!}{(k-j)!} c_j (\log x)^{k-j-1} .$$

The numbers c_j arise from the expansion $\frac{1}{\Gamma(z+1)} = \sum_{j=0}^{\infty} c_j z^j$. The functions $u_k(x)$ are all Lévy integrable, since $\int (x^2 \wedge 1) u_k(x) dx < \infty$ for all k. Therefore, for this sparse family, one can find a sequence of measures $\{H_k\}$, which are all exceedance measures according to

Definition 0.1, together with a sequence of rates $\{\rho_k\}$, with $\rho_k \propto \nu^k$, for which

$$P_\nu(w) = \rho H(w) + \rho_2 H_2(w) + \rho_3 H_3(w) + \cdots .$$

Interestingly, this expansion coincides with the symmetric version of the expansion proposed by Barndorff-Nielsen and Hubalek (2008) [7] for the infinitely divisible distribution on $\mathbb{R}_+$ associated to the gamma process, which has Lévy measure $\Lambda(dx) = e^{-x}x^{-1}dx$. The authors propose this expansion as a point-wise expansion of the density function. They seem to consider it valid in terms of integrals only against functions vanishing in a neighborhood of the origin, while we state it for the larger class $\mathcal{W}^{\#}$.

3.3 Scale sparse distributions

In this section we illustrate how to derive the sparsity integral expansion for those sparse measures whose sparsity is driven by their scale parameter going to zero. To give an insight of how we identify subsequent terms in the sparsity expansion for these families, we start presenting the derivation for two examples, the scaled Cauchy and scaled horseshoe. Following the same logic, we then extend it to the scaled Student's t with d degrees of freedom, $d \in (0, 2)$.

3.3.1 Scaled Cauchy and scaled horseshoe

Let $C_\nu(dx) = \frac{\nu}{\pi(x^2+\nu^2)}dx$ be the scaled Cauchy. As shown in the introduction, its first-order exceedance measure is $H(dx) = dx/(\sqrt{2\pi}|x|^2)$, while its rate function is $\rho = \nu\sqrt{2/\pi}$. We start by writing $C_\nu(dx) - \rho H(dx)$,

$$C_\nu(dx) - \rho H(dx) = \frac{\nu}{\pi(x^2+\nu^2)}dx - \nu\frac{1}{\pi|x|^2}dx = \frac{-\nu^2}{x^2}\left(\frac{1}{\nu}\frac{\nu^2}{\pi(x^2+\nu^2)}\right),$$

and notice that this can be written as

$$C_\nu(dx) - \rho H(dx) = \frac{-\nu^2}{x^2}\, P^\star_\nu(dx)\,, \tag{3.2}$$

where $P^\star_\nu = C_\nu$ is another scaled Cauchy. This clearly, by definition, is a sparse measure converging to $\delta_0(dx)$. Therefore, as $\nu \to 0$,

$$\frac{1}{-\nu^2}(C_\nu(dx) - \rho H(dx)) = \frac{1}{x^2} C_\nu(dx) \to \frac{1}{x^2}\delta_0(dx)\,.$$

Now, $|x|^{-2}\delta_0(dx)$ is not an exceedance measure but

$$H_2(w) = \int w(x)|x|^{-2}\, \delta_0(dx) \tag{3.3}$$

still defines a linear functional on $\mathcal{W}^\#$, which returns the value of $w(x)/x^2$ at zero. This value is finite as $w(x) = O(x^2)$ at the origin. If w has a second derivative at zero, then $H_2(w) = w''(0)/2$. In any case, one can identify $\rho_2 = -\nu^2$ and H_2 as in (3.3) to be the second-order pair in the integral expansion of C_ν,

$$C_\nu(w) = \rho H(w) + \rho_2 H_2(w) + o(\rho_2)\,.$$

At the same time, one can read $H_2(w)$ as the 'zero-term' in the expansion of the measure $P^\star_\nu = C_\nu$ for the integral of the function $s(x) = \frac{1}{x^2}w(x)$. Indeed, this function $s \notin \mathcal{W}^\#$, but, because $P^\star_\nu$ is itself sparse with first-order pair (ρ, H), one can still write

$$P^\star_\nu(s) = s(0) + \int (s(x) - s(0))\, P^\star_\nu(dx) = s(0) + \rho \int (s(x) - s(0))\, H(dx) \tag{3.4}$$

since $s(x) - s(0) = O(x^2)$ at the origin so that $\tilde{w}(x) = s(x) - s(0)$ belongs to $\mathcal{W}^\#$.

Putting (3.4) and (3.2) together,

$$\int w(x)\,(C_\nu(dx)-\rho H(dx)) = -\nu^2 \int s(x)\,P_\nu^\star(dx) = -\nu^2\left(s(0)+\rho\int (s(x)-s(0))H(dx)\right),$$

it comes natural to regard the integral appearing in the RHS of (3.4) as the third-order exceedance functional for the initial C_ν. In other words,

$$H_3(w) = \int\left(\frac{1}{x^2}w(x) - \int \frac{1}{x^2}w(x)\,\delta_0(dx)\right)H(dx) = \int (s(x)-s(0))\,H(dx)$$

is the linear functional on $\mathcal{W}^\#$ which, together with $\rho_3=\rho_2\rho$, constitute the third term in sparsity integral expansion of C_ν,

$$C_\nu(w) = \rho H(w) + \rho_2 H_2(w) + \rho_3 H_3(w) + o(\rho_3)\,.$$

Therefore, if w has a second derivative at zero, with a little sloppy notation, one can write the third-order sparsity expansion for the scaled Cauchy as

$$C_\nu(w(x)) = \sqrt{\frac{2}{\pi}}\,\nu H(w(x)) - \nu^2\,\frac{w''(0)}{2} - \sqrt{\frac{2}{\pi}}\,\nu^3 H\left(\frac{w(x)}{x^2} - \frac{w''(0)}{2}\right) + o(\nu^3)\,, \qquad (3.5)$$

from which one finds $\rho=\sqrt{\frac{2}{\pi}}\,\nu$, $\rho_2=-\nu^2$ and $\rho_3=-\sqrt{\frac{2}{\pi}}\,\nu^3$.

We now turn to look at the scaled horseshoe, $HS_\nu(dx) = \frac{1}{2\pi\,\nu}\log\left(1+\frac{\nu^2}{x^2}\right)dx$, whose first-order sparsity pair is $\rho=\frac{1}{\sqrt{2\pi}}\,\nu$ and $H(dx)=dx/(\sqrt{2\pi}|x|^2)$. As for the Cauchy, we start by writing

$$HS_\nu(dx)-\rho H(dx) = \frac{1}{2\pi\,\nu}\Big(\log\left(1+\frac{\nu^2}{x^2}\right)\Big)dx - \nu\,\frac{dx}{2\pi x^2} = \frac{-\nu^2}{x^2}\cdot\frac{1}{2\pi\nu}\Big(1-\frac{x^2}{\nu^2}\log\left(1+\frac{\nu^2}{x^2}\right)\Big)dx\,.$$

Now, the measure $\frac{1}{2\pi\nu}\left(1-\frac{x^2}{\nu^2}\log\left(1+\frac{\nu^2}{x^2}\right)\right)dx$ is not another scaled horseshoe. Nevertheless, it is still a scaled measure indexed by the sparsity parameter ν so, after normalization,

$$I=\frac{1}{2\pi}\int\left(1-x^2\log\left(1+\frac{1}{x^2}\right)\right)dx=\frac{1}{3}\,,$$

the probability measure

$$P_\nu^\star(dx)=3\,\frac{1}{2\pi\nu}\left(1-\frac{x^2}{\nu^2}\log\left(1+\frac{\nu^2}{x^2}\right)\right)dx\,,$$

is a sparse measure converging to the Dirac delta measure at zero as $\nu\to 0$. It thus follows that, as $\nu\to 0$,

$$\frac{3}{-\nu^2}\left(HS_\nu(dx)-\rho H(dx)\right)=\frac{1}{x^2}\,P_\nu^\star(dx)\to\frac{1}{x^2}\delta_0(dx)\,,$$

so, for the scaled horseshoe family, the second term in the sparsity integral expansion is given by

$$\rho_2 H_2(w)=-\frac{1}{3}\nu^2\int w(x)|x|^{-2}\,\delta_0(dx)\,.$$

At this stage, similarly to the Cauchy case, we can obtain the third term in the expansion of HS_ν by finding the first term $\rho^\star H^\star(w)$ in the expansion of the sparse measure $P_\nu^\star(dx)$. The first thing to notice is that $P_\nu^\star$ is another scale sparse family so, as mentioned in the introduction, its exceedance measure is an inverse-power measure and it can be determined by looking at the tail behavior of the unscaled density $p^\star(x)$ at infinity. Thus, it is sufficient to find the exponent γ of regular variation for which

$$\lim_{x\to\infty}\frac{p^\star(x)}{|x|^\gamma}=L(x)\,,$$

where $L(x)$ is a slowly varying function, i.e., $L(tx)/L(x)\to 1$ as $x\to\infty$ for every $t>0$ (see

Feller, 1966 [35]). Now,

$$\lim_{x\to\infty} \frac{\frac{3}{2\pi}\left(1 - x^2 \log\left(1 + \frac{1}{x^2}\right)\right)}{|x|^{-2}} = \frac{3}{4\pi},$$

so the tails of $p^\star$, at infinity, behave like $|x|^{-2}$, and therefore the exceedance measure $H^\star$ is again the inverse-square $H(dx) = dx/(\sqrt{2\pi}|x|^2)$, whereas its first-order rate is $\rho^\star = \frac{3}{4\pi}\sqrt{2\pi}\nu = \frac{3}{2\sqrt{2\pi}}\nu$. Thus, as for the Cauchy, writing

$$\int w(x)\,(HS_\nu(dx) - \rho H(dx)) = \frac{-\nu^2}{3}\int s(x) P^\star_\nu(dx) = \frac{-\nu^2}{3}\Big(s(0) + \rho^\star \int (s(x) - s(0)) H^\star(dx)\Big),$$

we arrive at identifying the third-order term in the expansion of HS_ν to be

$$\rho_2\rho^\star \int (s(x) - s(0))\, H^\star(dx) = \frac{-\nu^3}{2\sqrt{2\pi}} \int \left(\frac{1}{x^2} w(x) - \int \frac{1}{x^2} w(x)\, \delta_0(dx)\right) H(dx)\,.$$

Again, if w has a second derivative at zero, with the same notation of (3.5), we can write the third-order sparsity integral expansion for the scaled horseshoe as

$$HS_\nu(w(x)) = \frac{1}{\sqrt{2\pi}}\,\nu H(w(x)) - \frac{1}{3}\,\nu^2\,\frac{w''(0)}{2} - \frac{1}{2\sqrt{2\pi}}\,\nu^3\,H\left(\frac{w(x)}{x^2} - \frac{w''(0)}{2}\right) + o(\nu^3)\,, \quad (3.6)$$

from which one finds $\rho = \frac{1}{\sqrt{2\pi}}\,\nu$, $\rho_2 = -\frac{1}{3}\,\nu^2$ and $\rho_3 = -\frac{1}{2\sqrt{2\pi}}\,\nu^3$.

3.3.2 Scaled Student's t

We now derive the sparsity integral expansion for the scaled Student's t distribution with $\alpha \in (0, 2)$ degrees of freedom,

$$P_\nu(dx) = K_\alpha \frac{\nu^\alpha}{(x^2 + \alpha\nu^2)^{\frac{\alpha+1}{2}}}\,dx\,,$$

where $K_\alpha = \Gamma(\frac{\alpha+1}{2})\alpha^{\alpha/2}/(\sqrt{\pi}\Gamma(\alpha/2))$. This sparse family has first-order rate

$$\rho = K_\alpha \frac{\Gamma(1-\alpha/2)}{\alpha 2^{\alpha/2-1}} \nu^\alpha$$

and exceedance measure

$$H(dx) = \frac{\alpha 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)} |x|^{-\alpha-1} dx \,.$$

Taking the cue from the Cauchy and horseshoe examples, we start by writing the difference between $P_\nu(dx)$ and its first-order term $\rho H(dx)$ and try to obtain another sparse probability measure. So,

$$\begin{aligned} P_\nu(dx) - \rho H(dx) &= K_\alpha \frac{\nu^\alpha}{(x^2+\alpha\nu^2)^{\frac{\alpha+1}{2}}} - K_\alpha \frac{\nu^\alpha}{(x^2)^{\frac{\alpha+1}{2}}} \\ &= \frac{-\nu^2}{x^2} \frac{K_\alpha}{\nu} \left(\frac{x^2/\nu^2}{(x^2/\nu^2)^{\frac{\alpha+1}{2}}} - \frac{x^2/\nu^2}{(x^2/\nu^2+\alpha)^{\frac{\alpha+1}{2}}} \right) dx \\ &= \frac{-\nu^2}{x^2} I_\alpha \cdot \frac{1}{\nu} p^\star\left(\frac{x}{\nu}\right) dx \,, \end{aligned}$$

where I_α is the normalization constant

$$I_\alpha = \int K_\alpha \left(\frac{x^2}{(x^2)^{\frac{\alpha+1}{2}}} - \frac{x^2}{(x^2+\alpha)^{\frac{\alpha+1}{2}}} \right) dx = \frac{\alpha}{2-\alpha} \,,$$

and $p^\star(x)$ is the probability density function

$$p^\star(x) = \frac{1}{I_\alpha} K_\alpha \left(\frac{x^2}{(x^2)^{\frac{\alpha+1}{2}}} - \frac{x^2}{(x^2+\alpha)^{\frac{\alpha+1}{2}}} \right) .$$

So, similarly to the Cauchy and horseshoe cases, we have

$$P_\nu(dx) - \rho H(dx) = \frac{-\nu^2}{x^2} I_\alpha \cdot P^\star_\nu(dx) \,,$$

where $P^\star_\nu$ is another scale sparse measure: it converges to $\delta_0(dx)$ as $\nu \to 0$ and has a first-order sparsity pair $(\rho^\star, H^\star)$, with $H^\star$ being an inverse-power measure. Again, to find $H^\star$, we

just need to determine the tail behavior of $p^\star$. Since

$$\lim_{x\to\infty} \frac{p^\star(x)}{|x|^{-\alpha-1}} = \frac{K_\alpha}{I_\alpha}\frac{\alpha(\alpha+1)}{2}\,,$$

the exceedance measure of $P_\nu^\star$ is the α-inverse-power measure $H(dx) = \frac{\alpha 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}|x|^{-\alpha-1}dx$, and the rate is $\rho^\star = \frac{1}{I_\alpha}\,\frac{\alpha(\alpha+1)}{2}\,K_\alpha\,\frac{\Gamma(1-\alpha/2)}{\alpha 2^{\alpha/2-1}}\,\nu^\alpha = \frac{1}{I_\alpha}\,\frac{\alpha(\alpha+1)}{2}\,\rho$. Then

$$\int w(x)(P_\nu(dx)-\rho H(dx)) = -\nu^2 I_\alpha \int s(x)P_\nu^\star(dx) = -\nu^2 I_\alpha\Big(s(0)+\rho^\star\int (s(x)-s(0))H(dx)\Big)$$

leads to the third-order sparsity integral expansion of the scaled t_α of Student

$$P_\nu(w(x)) = \rho H(w(x)) + \rho_2\,\frac{w''(0)}{2} + \rho_3 H\left(\frac{w(x)}{x^2} - \frac{w''(0)}{2}\right) + o(\nu^{2+\alpha})\,, \tag{3.7}$$

where $\rho = K_\alpha\,\frac{\Gamma(1-\alpha/2)}{\alpha 2^{\alpha/2-1}}\,\nu^\alpha$, $\rho_2 = -\frac{\alpha}{2-\alpha}\,\nu^2$ and $\rho_3 = -\frac{\alpha(\alpha+1)}{2}\,K_\alpha\,\frac{\Gamma(1-\alpha/2)}{\alpha 2^{\alpha/2-1}}\,\nu^{2+\alpha}$.

Comparing (3.7) to (3.5) and (3.6), we can recognize the same pattern in the alternation of the measures. Indeed, after eliminating the first-order term, one is left with two functionals which are closely related to another sparse measure, having the same tail behavior as the starting one. All three cases share the same second-order exceedance functional, while the third one retains the inverse power characterizing the initial sparse measure. And, as needed, the expansion for the scaled t_α, with $\alpha = 1$, coincides with the expansion for the scaled Cauchy.

3.3.3 Normalization

To give an illustration of the sparsity integral expansions derived in the previous sections, consider the function $w(x) = (1 - e^{-x^2/2})$. By definition, any unitary exceedance measure

$H(dx)$ is normalized in such a way that

$$H(w) = \int (1 - e^{-x^2/2})\, H(dx) = 1\,.$$

Now, $s(x) = \frac{w(x)}{x^2} = \frac{(1-e^{-x^2/2})}{x^2}$ so that $s(0) = \frac{w''(0)}{2} = \frac{1}{2}$. Suppose that $H(dx)$ is the α-inverse-power measure, then

$$\begin{aligned}
&\int (s(x) - s(0))|x|^{-\alpha-1}\, dx = \\
&\int \left(\frac{1}{x^2}(1 - e^{-x^2/2}) - \frac{1}{2}\right) |x|^{-\alpha-1}\, dx = \\
&\int \left(\sum_{r=2}^{\infty} \frac{(-1)^r x^{2r-2}}{2^r r!}\right) |x|^{-\alpha-1}\, dx = \\
&\int \left(e^{-x^2/2} - 1 - x^2/2\right) (x^2)^{-1-\frac{\alpha+1}{2}}\, dx = \\
&\int \left(e^{-z/2} - 1 - z/2\right) z^{-1-\frac{\alpha+1}{2}-\frac{1}{2}}\, dx = \\
&- \frac{\Gamma(1-\alpha/2)}{\alpha(\alpha+2)2^{\alpha/2-1}}\,.
\end{aligned}$$

So

$$H\left(\frac{w(x)}{x^2} - \frac{w''(0)}{2}\right) = \int (s(x) - s(0)) \frac{\alpha 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)} |x|^{-\alpha-1}\, dx = \frac{-1}{\alpha+2}\,.$$

Then, for a scale sparse measure P_ν with rates ρ, ρ_2, ρ_3 and exceedance functionals H, H_2, H_3 as those in (3.6) and (3.7), the third-order sparsity expansion of $\int (1 - e^{-x^2/2})\, P_\nu(dx)$ gives

$$\rho + \rho_2 \frac{w''(0)}{2} + \rho_3 H\left(\frac{w(x)}{x^2} - \frac{w''(0)}{2}\right) + o(\rho_3) = \rho + \frac{1}{2}\,\rho_2 - \frac{1}{\alpha+2}\,\rho_3 + o(\rho_3)\,. \tag{3.8}$$

In light of (3.8), we could decide to renormalize the second and third-order exceedance functionals H_2 and H_3 in such a way that

$$H_2(1 - e^{-x^2/2}) = 1 \quad \text{and} \quad H_3(1 - e^{-x^2/2}) = 1\,.$$

More generally, any exceedance functional of order k such that

$$H_k(1 - e^{-x^2/2}) = 1$$

will be called a unitary exceedance functional. So, for the sparse scale family, the unitary second-order exceedance functional is

$$H_2(w) = 2 \int \frac{w(x)}{x^2} \, \delta_0(dx) \, ,$$

while the unitary third-order exceedance functional is

$$H_3(w) = -(\alpha + 2) \int \left(\frac{w(x)}{x^2} - \int \frac{w(x)}{x^2} \, \delta_0(dx) \right) H(dx) \, .$$

Clearly, this requires adjusting the second and third-order rates to be $\tilde{\rho}_2 = \frac{1}{2} \, \rho_2$ and $\tilde{\rho}_3 = -\frac{1}{\alpha+2} \, \rho_3$. In this way, the sparse approximation of order $o(\tilde{\rho}_3)$ for $\int (1 - e^{-x^2/2}) \, P_\nu(dx)$ is

$$\rho + \tilde{\rho}_2 + \tilde{\rho}_3 \, . \tag{3.9}$$

Suppose that $\rho = 0.1$, then (3.9) is equal to 0.0927 for the scaled Cauchy with scale-sparsity parameter $\rho\sqrt{\pi/2} = 0.125$, while (3.9) is equal to 0.0906 for the scaled horseshoe with scale-sparsity parameter $\rho\sqrt{2\pi} = 0.251$. On the other hand the first-order approximation 0.1 gets corrected downwards to 0.0649 when the scale sparse measure is Student's t with $\alpha = 1.5$ degrees of freedom.

3.3.4 Impact on signal-plus-noise marginal density

We conclude this part on scale sparse distributions investigating the impact of including higher-order terms in the sparsity expansion of a scale sparse measure on the inference for the signal-plus-noise model.

Let $Y = \mu + \eta$, where $\mu \sim P_\nu$ is a scale sparse signal, independent of $\eta \sim N(0,1)$. Then the marginal density of Y can be written as

$$m_\nu(y) = \phi(y) \left(\int_{\mathbb{R}} (\cosh(yx) - 1) e^{-x^2/2} \, P_\nu(dx) + 1 - \int_{\mathbb{R}} (1 - e^{-x^2/2}) \, P_\nu(dx) \right) .$$

Both functions appearing inside the two integrals are in the class $\mathcal{W}^\#$. Thus, we can apply the sparsity integral expansions derived in the previous sections to both of these functions, to obtain the third-order sparse approximation for m_ν.

For doing so, we suppose that P_ν has unitary exceedance functionals given by $H(w) = \int w(x) H(dx)$, $H_2(w) = w''(0)$ and $H_3(w) = -(\alpha+2) \int (w(x)/x^2 - w''(0)/2) H(dx)$ respectively, where $H(dx) = \frac{\alpha 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)} |x|^{-\alpha-1} dx$ is the inverse power measure with index α. This is the case when P_ν is any scaled t_α distribution, $\alpha \in (0,2)$, as well as when P_ν is a scaled horseshoe. Then, given the associated rates $\rho, \tilde{\rho}_2, \tilde{\rho}_3$, we can write

$$m_\nu(y) = \phi(y) \left(\rho \zeta(y) + \tilde{\rho}_2 y^2 + \tilde{\rho}_3 \zeta^{(3)}(y) + 1 - \rho - \tilde{\rho}_2 - \tilde{\rho}_3 \right) + o(\tilde{\rho}_3) \, . \tag{3.10}$$

Here $\zeta(y) = \int_{\mathbb{R}\setminus\{\mathbf{0}\}} (\cosh(yx) - 1) e^{-x^2/2} \, H(dx)$ is the zeta function introduced by McC&P, while

$$\zeta^{(3)}(y) = -(\alpha+2) \int_{\mathbb{R}\setminus\{\mathbf{0}\}} \left(\frac{(\cosh(yx) - 1) e^{-x^2/2}}{x^2} - \frac{y^2}{2} \right) H(dx)$$

is another integral transform of H_d, which can be similarly written as a power series

$$(\alpha+2)\frac{y^2}{2} - \sum_{r=2}^{\infty} \frac{y^{2r}}{(2r)!} \frac{\alpha \, (\alpha+2) \Gamma(r - 1 - \alpha/2) 2^{r-2}}{\Gamma(1-\alpha/2)} \, .$$

See the appendix for derivation.

Notice that because of the normalization chosen, all three functions appearing in (3.10),

$\zeta(y)$, y^2 and $\zeta^{(3)}(y)$, when multiplied by the Gaussian density, integrate to one over the real line. However, while both $\psi(y) = \phi(y)\zeta(y)$ and $\psi_2(y) = \phi(y)y^2$ are also always non negative, which means they are density functions, $\psi_3(y) = \phi(y)\zeta^{(3)}(y)$ is not always non negative. This is shown in Figure 3.1, where we plot all three psi functions for $\alpha = 0.5$, 1 and 1.5. Recall that $\alpha = 1$ is the index for the scaled Cauchy and scaled horseshoe.

Figure 3.2 shows the impact of the subsequent inclusion of higher-order terms on the tail-inflation factor in the sparse approximation to m_ν:

$$\phi(y)\left(\rho\,\zeta(y) + \tilde{\rho}_2\,y^2 + \tilde{\rho}_3\,\zeta^{(3)}(y)\right)\,.$$

In this case, we fix $\rho = 0.1$ and $\alpha = 1$, and compute $\tilde{\rho}_2$ and $\tilde{\rho}_3$ for the Cauchy case. The inclusion of the second-order term has the effect of lowering the first-order function at the two symmetric peaks around the origin. On the other hand, the impact of including also the third-order term goes in the opposite direction but it is much less appreciable.

Figure 3.3 instead shows the consequences of subsequently including higher-order components in the sparse approximation to the marginal density m_ν as in (3.10). Again, we consider the case when the signal distribution is the scaled Cauchy, with first-order rate $\rho = 0.1$. We compare the sparse approximations of different orders (colored solid lines) to the exact convolution density (black dashed line). We can see that the third-order approximation, depicted by the yellow line, is closer to the exact density than the first-order approximation, the blue line, both in the central part around the origin, as well as in the tails.

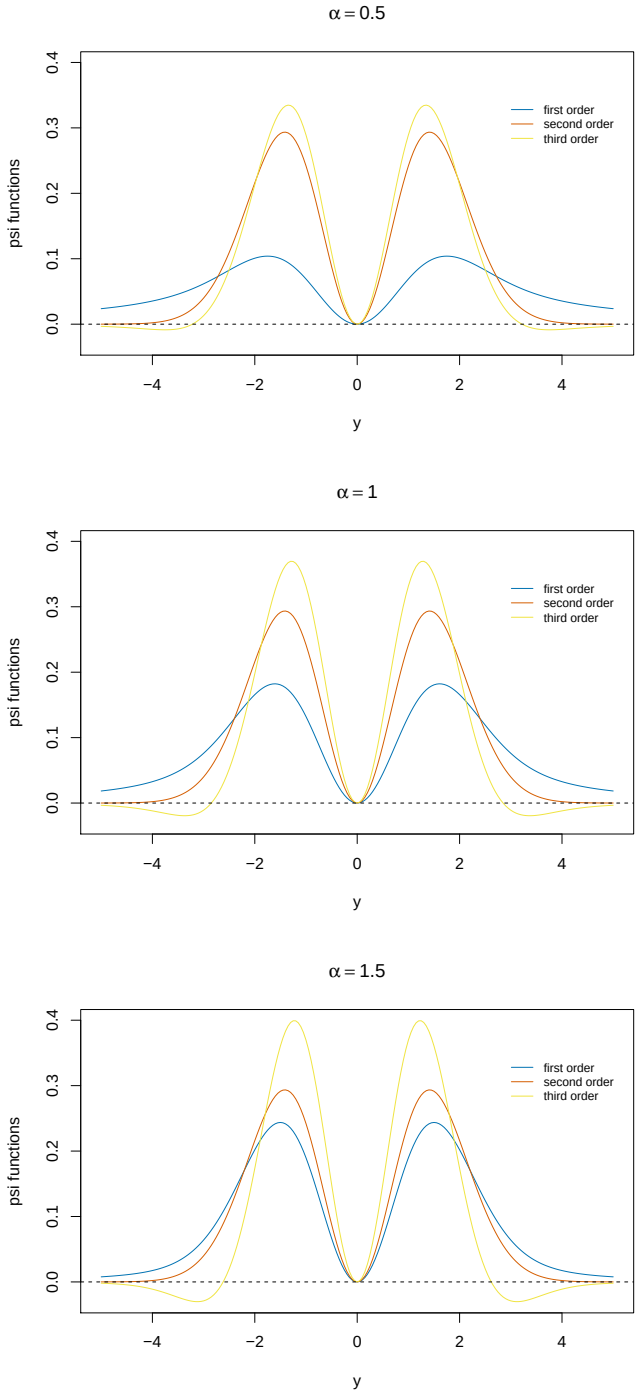


Figure 3.1: Psi functions appearing in the sparse approximation of $m_\nu(y)$, for $\alpha = 0.5, 1, 1.5$. Blue curves depict $\psi(y) = \phi(y)\zeta(y)$, brown curves depict $\psi_2(y) = \phi(y)y^2$, while yellow curves depict $\psi_3(y) = \phi(y)\zeta^{(3)}(y)$.

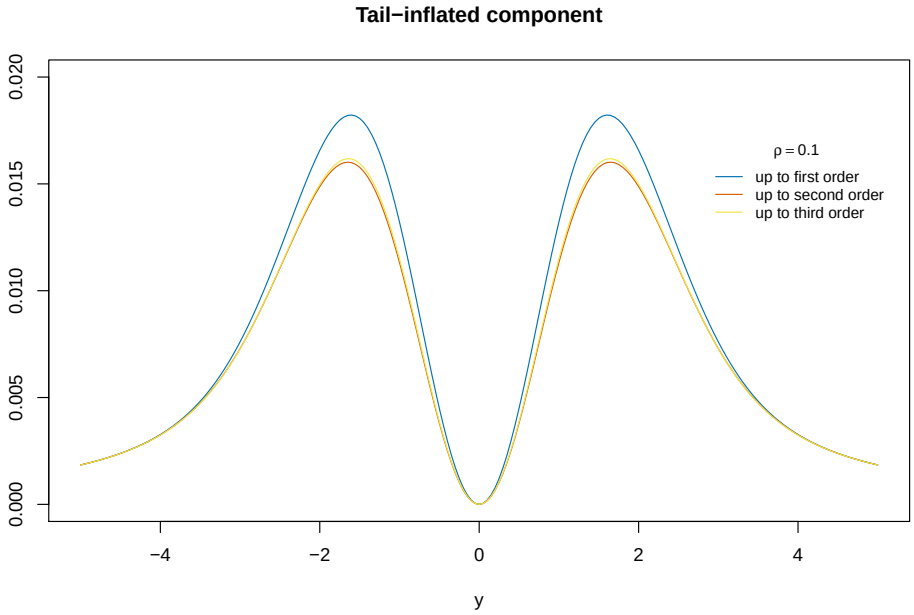


Figure 3.2: Impact on tail inflation factor of the sparse approximation for $m_\nu(y)$. Blue curve depicts only $\rho\phi(y)\zeta(y)$, brown curve includes second-order term, $\phi(y)(\rho\,\zeta(y) + \tilde{\rho}_2\,y^2)$, while yellow curve includes the third-order term as well, $\phi(y)(\rho\,\zeta(y) + \tilde{\rho}_2\,y^2 + \tilde{\rho}_3\,\zeta^{(3)}(y))$. We fix $\rho = 0.1$, $\alpha = 1$ and compute $\rho_2 = -0.005$ and $\rho_3 = 0.27 \cdot 10^{-3}$ as for the Cauchy case.

3.4 Mixtures

In this section, we examine another class of sparse families, namely those which can be written as a mixture of two distributions. This kind of mixture measures frequently appear frequently in the Bayesian literature on the signal-plus-noise model and all of its extensions. See for instance, Mitchell and Beauchamp (1988) [52], Johnstone and Silverman (2004) [45], Ročková and George (2018) [60], among many many others. Indeed, the mixture idea captures the prior information that, with some probability p, the signal might be zero, or very close to zero, and, with probability $1 - p$, the signal has a non-trivial distribution over $\mathbb{R}$. See Efron (2007) [27] and references therein for the two-group model and its use in the microarray data literature.

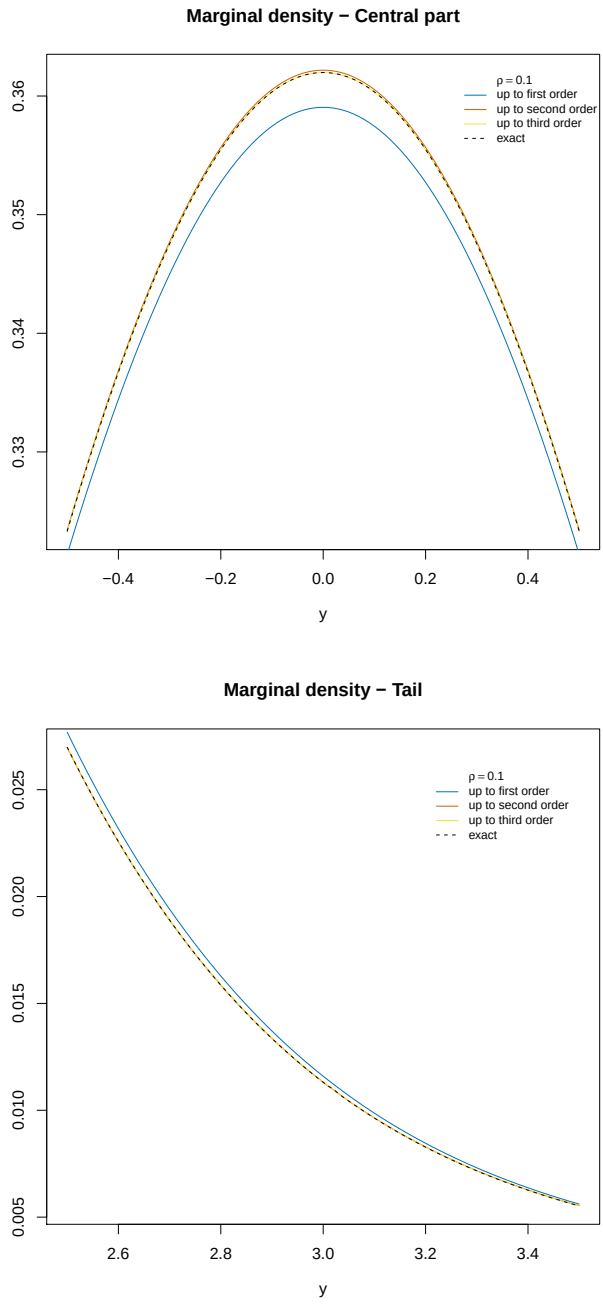


Figure 3.3: Impact of higher-order terms in the sparse approximation to the marginal density of Y, when it is the convolution of a scaled Cauchy and a standard Gaussian: $\phi(y)\left(\rho\zeta(y) + \tilde{\rho}_2 y^2 + \tilde{\rho}_3\zeta^{(3)}(y) + 1 - \rho - \tilde{\rho}_2 - \tilde{\rho}_3\right) + o(\tilde{\rho}_3)$, when first-order sparsity rate is $\rho = 0.1$. Blue curves include only first-order term, brown curves include up to second-order term while yellow curves include up to third-order term. The dashed black curves instead show the exact density for the convolution. Top panel: central part of the density. Bottom panel: tail of the density.

3.4.1 Atom-and-slab family

Consider the following atom-and-slab model

$$P_\nu(dx) = (1-\nu)\delta_0(dx) + \nu F(dx)\,, \tag{3.11}$$

where F is any symmetric distribution on $\mathbb{R}$. In this case, it is immediate to identify the first-order sparsity rate to be $\rho \propto \nu$ and finite exceedance measure proportional to $F(dx)$. Once we eliminate this first-order term νF from P_ν, we are left with just the Dirac delta measure at zero, which, being the limit itself, does not have a non-trivial sparse expansion. Thus, for this family, one can simply write

$$P_\nu(w) = \nu F(w) + o(\nu^\infty)\,.$$

Now instead, consider a slightly different family

$$P_\nu(dx) = (1-\nu)\delta_0(dx) + \nu F_\nu(dx)\,, \tag{3.12}$$

where the slab distribution F_ν is parametrized by the sparsity parameter. If F_ν is sparse with first-order pair (ρ, H), then the first-order term in the sparsity expansion of P_ν is given by the pair $(\nu\rho, H)$. Similarly, if F_ν has a non-trivial second-order term in its sparse expansion, say ρ_2 and H_2, then P_ν will also have a second-order term. More generally, if F_ν has a k^{th}-order sparse expansion, then one can write

$$P_\nu(w) = \nu(\rho H(w) + \rho_2 H_2(w) + \ldots \rho_k H_k(w)) + o(\nu\rho_k)\,.$$

On the other hand, if F_ν is not sparse, then it might be the case that, even if P_ν converges to the Dirac delta measure as $\nu \to 0$, P_ν does not have a first-order sparsity term. For instance, if F_ν is a Cauchy distribution with scale parameter $1/\nu$, i.e.,

$P_\nu(dx) = (1-\nu)\delta_0(dx) + \nu\frac{\nu}{\pi(\nu^2x^2+1)}dx$, then

$$\nu^{-2}\int w(x)\,P_\nu(dx) = \nu^{-2}\int w(x)\frac{dx}{\pi(\nu^2x^2+1)} = \int w(x)\,\frac{dx}{\pi}\,.$$

Yet, the functional $w \to \int w(x)\frac{dx}{\pi}$ is finite only if w is integrable with respect to Lebesgue measure on $\mathbb{R}$, which is not necessarily true for all functions in $\mathcal{W}^{\#}$.

In conclusion, for the atom-and-slab family in (3.12), the sparsity integral expansion of P_ν strictly depends on the behavior of the slab distribution F_ν, as ν goes to zero.

3.4.2 Spike-and-slab family

The spike-and-slab family is usually referred to as a mixture distribution of the kind

$$P_\nu(dx) = (1-\nu)G_\nu(dx) + \nu F(dx)\,, \tag{3.13}$$

where F is any symmetric distribution on $\mathbb{R}$, while G_ν is a sparse measure converging to the Dirac delta at zero as $\nu \to 0$. Depending on how fast this convergence occurs, we use different techniques to derive a sparsity integral expansion for P_ν.

Spike distribution with non-exponential tails

When the spike distribution G_ν is itself a sparse measure for which a non trivial sparsity integral expansion can be derived, then, besides the first-order term νF, the expansion of P_ν follows the expansion of G_ν. For example, if G_ν is the scaled Cauchy C_{ν^2} then, following (3.5), the first four terms in the expansion of P_ν are

$$P_\nu(w(x)) = \nu F(w(x)) + \sqrt{\frac{2}{\pi}}\,\nu^2 H(w(x)) - \nu^4\,\frac{w^{(2)}(0)}{2} - \sqrt{\frac{2}{\pi}}\,\nu^6 H\left(\frac{w(x)}{x^2} - \frac{w^{(2)}(0)}{2}\right) + o(\nu^6)\,.$$

Clearly, if in the expansion of G_ν, there is a number of rates ρ_k, say from $k = 1$ to $j - 1$, such that $\nu = o(\rho_k)$ and $\rho_j \propto \nu$, then the j^{th} term in the expansion of P_ν will be

$$\rho_j H_j(w) + \nu F(w)\,.$$

For example, if G_ν is the scaled Cauchy $C_{\sqrt{\nu}}$, then $\rho_1 = \sqrt{\frac{2}{\pi}}\sqrt{\nu}$ and $\rho_2 = -\nu$, so

$$P_\nu(w(x)) = \sqrt{\frac{2\nu}{\pi}}\,H(w(x)) + \nu\left(F(w(x)) - \frac{w^{(2)}(0)}{2!}\right) - \sqrt{\frac{2\nu^3}{\pi}}H\left(\frac{w(x)}{x^2} - \frac{w^{(2)}(0)}{2}\right) + o(\sqrt{\nu^3})\,.$$

Spike distribution with exponential tails

If G_ν has exponential tails, such as $N(0, \nu^2)$ (George and McCulloch, 1993 [40]), or Laplace with scale parameter ν^2 (George and Ročková, 2018 [60]), then there is no definite rate for which the sparsity limit integral definition of McC&P holds with a non trivial measure H. However, since we are now considering a broader class of operators for the higher-order exceedances, even for this class of measures, we can find a sparsity integral expansion, following our previous logic.

Because G_ν has exponential tails, given $X \sim G_\nu$, X has all finite moments. Thus, for those symmetrized functions $w \in \mathcal{W}^{\#}$ having a convergent Taylor expansion at zero, with non-zero coefficients only for even powers, it is possible to write

$$\int w(x)\,G_\nu(dx) = \int \sum_{k=0}^{\infty} \frac{w^{(2k)}(0)}{2k!}\,x^{2k}\,G_\nu(dx) = \sum_{k=1}^{\infty} \frac{w^{(2k)}(0)}{(2k)!}\,\mathbb{E}_\nu(X^{2k})\,,$$

where the first term $w(0) = 0$ because $w(x) = O(x^2)$ at the origin. Now, for each $k \geq 1$, one can write

$$\frac{w^{(2k)}(0)}{(2k)!} = \int w_{k-1}(x) \cdot \frac{1}{x^2}\,\delta_0(dx)\,,$$

where the functions $w_k(x)$ for $k \geq 1$, are defined recursively by

$$w_k(x) = w_{k-1}(x) \cdot \frac{1}{x^2} - \int w_{k-1}(x) \cdot \frac{1}{x^2}\, \delta_0(dx)\,, \tag{3.14}$$

and $w_0(x) = w(x)$. So, one has

$$w_1(x) = \sum_{k=1}^{\infty} \frac{w^{(2k)}(0)}{(2k)!}\, x^{2k-2} - \frac{w^{(2)}(0)}{2!} = \sum_{k=2}^{\infty} \frac{w^{(2k)}(0)}{(2k)!}\, x^{2k-2}\,,$$

$$w_2(x) = \sum_{k=2}^{\infty} \frac{w^{(2k)}(0)}{(2k)!}\, x^{2k-4} - \frac{w^{(4)}(0)}{4!} = \sum_{k=3}^{\infty} \frac{w^{(2k)}(0)}{(2k)!}\, x^{2k-4}\,,$$

et cetera. Therefore, for every $k \geq 1$, we can define the functional H_k on $\mathcal{W}^{\#}$ to be

$$H_k(w) = \int w_{k-1}(x) \cdot \frac{1}{x^2}\, \delta_0(dx) = \frac{w^{(2k)}(0)}{(2k)!}\,.$$

So the function w_k as in (3.14) can now be written as

$$w_k(x) = w_{k-1}(x) \cdot \frac{1}{x^2} - H_k(w)\,.$$

If the even moments of $X \sim G_\nu$ are denoted by

$$\begin{aligned} \rho_1 &= \mathbb{E}_\nu(X^2)\,, \\ \rho_2 &= \mathbb{E}_\nu(X^4)\,, \\ &\vdots \\ \rho_k &= \mathbb{E}_\nu(X^{2k})\,, \end{aligned}$$

then, provided $\rho_k = o(\nu)$ for all $k \geq 1$, we can write the sparsity integral expansion of P_ν as

$$P_\nu(w) = \nu F(w) + \sum_{k=1}^{\infty} \frac{w^{(2k)}(0)}{(2k)!}\, \mathbb{E}_\nu(X^{2k}) = \nu F(w) + \rho_1 H_1(w) + \rho_2 H_2(w) + \cdots\,.$$

3.5 Higher-order equivalences

We now turn to investigate the question whether the equivalence relationship established by the first-order sparsity pair carries over to higher-orders, and in what ways this can happen. We first recall the concept of sparsity equivalence as defined in McC&P. Given two sparse families of distributions, they are said to be first-order equivalent in the sparse limit if they share the same exceedance measure. In fact, this sharing allows one to reparametrize one of the two families in such a way that the two rates also match. Therefore, the first term $\rho H(w)$ in the sparsity integral expansion is exactly the same for both families. Below, we give two examples, one for which the relation carries over, and one for which this does not occur.

Scaled Cauchy versus scaled horseshoe

As highlighted in McC&P, the scaled Cauchy and scaled horseshoe are equivalent in their sparsity expansion since they both have the square inverse-power exceedance measure. Comparing (3.5) and (3.6), we can immediately notice that the two functionals H_2 and H_3 on $w \in \mathcal{W}^{\#}$ are the same for both families, so that their first-order equivalence carries over to higher orders, in terms of exceedance functionals. On the other hand, the higher-order rates of one family are scalar multiples of those of the other family. In fact, suppose we match the first-order rates

$$\frac{1}{\sqrt{2\pi}}\,\nu_{HS} = \rho^{HS} = \rho^{C} = \sqrt{\frac{2}{\pi}}\,\nu_{C}\,,$$

by setting the scale-sparsity parameter for the horseshoe family to be two times that of the Cauchy family: $\nu_{HS} = 2\nu_C$. Then the second-order and third-order rates are found to be in the following relations

$$\rho_2^{HS} = \frac{4}{3}\rho_2^{C}\,, \qquad \rho_3^{HS} = 2\rho_3^{C}\,.$$

Atom-and-Slab versus Spike-and-Slab

The atom-and-slab family and the spike-and-slab family as in (3.11) and (3.13), respectively, sharing the same slab distribution, are first-order equivalent. Yet, they are not second-order equivalent in their sparse limit. In fact, for the former family, after the first term νF, all remaining terms are just trivially equal to zero. On the contrary, for the latter family, if the spike distribution G_ν is, for instance, $N(0, \nu^2)$, then after the first term νF, one can identify the second-order term in the expansion of P_ν to be $\nu^2 w^{(2)}(0)/2$. This means that the equivalence relation between the two families holds up to first-order but it breaks down when we look at the second-order term.

3.6 Appendix

1. Here we derive the zeta function associated to the third-order exceedance functional,

$$H_3(w) = -(\alpha+2)\int\left(\frac{w(x)}{x^2} - \int \frac{w(x)}{x^2}\,\delta_0(dx)\right) H(dx)\,.$$

Let $w(x) = (\cosh(xy)-1)e^{-x^2/2}$. Then

$$s(x) = \frac{w(x)}{x^2} = \frac{(\cosh(xy)-1)e^{-x^2/2}}{x^2}\,,$$

so that $s(0) = \frac{w''(0)}{2} = \frac{y^2}{2}$. Letting $\kappa_\alpha = -\frac{(\alpha+2)\alpha 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}$,

$$\begin{aligned}
\zeta^{(3)}(y) &= -(\alpha+2)\int\left(\frac{(\cosh(xy)-1)e^{-x^2/2}}{x^2} - \frac{y^2}{2}\right) H(dx) \\
&= \int\left(\sum_{r=1}^{\infty}\frac{x^{2r-2}y^{2r-2}}{(2r)!/2}\,\frac{y^2}{2}e^{-x^2/2} - \frac{y^2}{2}\right)\frac{\kappa_\alpha}{|x|^{\alpha+1}}\,dx \\
&= \frac{y^2}{2}\int\left(\sum_{j=0}^{\infty}\frac{x^{2j}y^{2j}}{(2r)!/2}\,e^{-x^2/2} - (1-e^{-x^2/2}+e^{-x^2/2})\right)\frac{\kappa_\alpha}{|x|^{\alpha+1}}\,dx \\
&= \frac{y^2}{2}\int\left(\sum_{j=0}^{\infty}\frac{x^{2j}y^{2j}}{(2r)!/2} - 1\right)\,e^{-x^2/2}\frac{\kappa_\alpha}{|x|^{\alpha+1}}\,dx - \frac{y^2}{2}\int(1-e^{-x^2/2})\frac{\kappa_\alpha}{|x|^{\alpha+1}}\,dx \\
&= \frac{y^2}{2}\int\left(\sum_{j=1}^{\infty}\frac{x^{2j}y^{2j}}{(2r)!/2}\right)\,e^{-x^2/2}\frac{\kappa_\alpha}{|x|^{\alpha+1}}\,dx - \frac{y^2}{2}(-\alpha-2) \\
&= \frac{y^2}{2}\,\sum_{j=1}^{\infty}\frac{y^{2j}}{(2r)!/2}\int x^{2j}\,e^{-x^2/2}\frac{\kappa_\alpha}{|x|^{\alpha+1}}\,dx + \frac{y^2}{2}(\alpha+2) \\
&= \frac{y^2}{2}\,\sum_{j=1}^{\infty}\frac{y^{2j}}{(2r)!/2}\frac{\Gamma(j-\alpha/2)}{(1/2)^{j-\alpha/2}}\kappa_\alpha + \frac{y^2}{2}(\alpha+2) \\
&= \sum_{j=1}^{\infty}\frac{y^{2j+2}}{(2r)!}\frac{\Gamma(j-\alpha/2)}{(1/2)^{j-\alpha/2}}\frac{\alpha 2^{\alpha/2-1}(-\alpha-2)}{\Gamma(1-\alpha/2)} + \frac{y^2}{2}(\alpha+2) \\
&= \frac{y^2}{2}(\alpha+2) - \sum_{r=2}^{\infty}\frac{y^{2r}}{(2r)!}\frac{\Gamma(r-1-\alpha/2)2^{r-2}\,\alpha\,(\alpha+2)}{\Gamma(1-\alpha/2)}\,.
\end{aligned}$$

So the third-order sparsity expansion of $\int(\cosh(xy)-1)e^{-x^2/2}\,P_\nu(dx)$ is given by

$$\int(\cosh(xy)-1)e^{-x^2/2}\,P_\nu(dx) =$$

$$\rho\int(\cosh(xy)-1)e^{-x^2/2}\,H(dx)+\rho_2\frac{w''(0)}{2}+\rho_3\int\left(\frac{(\cosh(xy)-1)e^{-x^2/2}}{x^2}-\frac{y^2}{2}\right)H(dx) =$$

$$\rho\zeta(y)+\tilde{\rho}_2\,y^2+\tilde{\rho}_3\;\zeta^{(3)}(y)\,.$$

Part II

Component-wise sparsity

Chapter 4

Component-wise sparsity and negligibility

4.1 Introduction

With this chapter, we start our investigation of d-dimensional sparse distributions, which will be the main theme of the rest of the thesis. Therefore, we begin by giving the definition of multivariate sparsity, which is the natural extension of the original definition of sparsity of McC&P, to sequences of probability distributions defined on $\mathbb{R}^d$, $d > 1$. This definition is very general and we will refer to it also in Part III. In this part of the thesis, we focus our attention on those multivariate sparse distributions which are product of univariate sparse measures. We call this kind of multivariate sparsity, *component-wise sparsity*. Thus, in a sense, we keep the scalar notion of symmetry on the real line and regard the multivariate signal as a collection of d independent scalar signals.

In a context of independence like the present one, fairly often the interest lies in establishing which signals can be considered as 'active' and which can be considered as 'negligible'. However, as explained in the introduction, the univariate sparsity theory developed in

McC&P relies on the sparsity pair (ρ, H), which, by itself, does not allow one to identify the null atom $\{X = 0\}$. For this reason, we frame the question from a slightly different perspective, introducing a formal notion of negligibility. This idea of signal negligibility is to be understood as a limit-approaching description of the sparse sequence of measures over a sequence of positive-length intervals converging to the limit point $\{0\}$. Given this limiting notion of signal negligibility, we derive an alternative integral approximation for the univariate sparse measure P_ν. This alternative approximation is valid up to an error larger than the usual $o(\rho)$, but has a Dirac delta measure component in it, so that, in terms of integrals of bounded and continuous functions, negligibility is equivalent to being zero.

The integral approximation we develop within the negligibility-sparsity theory can be naturally seen as an alternative to many other Bayesian approaches developed for the so called two-groups model (Efron, 2007 [27]). Yet, despite the general structure resemblance, there are two main differences. The first most obvious discrepancy concerns the target event, which determines the mixture distribution for the signal. In our framework this event is the signal negligibility, whereas, in many formulations of the two-groups model, is the event of the signal being absolutely zero. The second difference follows from the first one to the extent that we obtain an atomic mixture (i.e., a Dirac delta measure component) for the signal distribution as an asymptotic approximation, driven by the sparsity limit, and true solely in terms of integrals of bounded and continuous functions.

This alternative sparse integral approximation turns out to be a useful tool in many different applied statistical contexts. In the next chapters, we will use it for: (i) constructing a multiple testing procedure to declare negligible and non negligible signals; (ii) obtaining a soft-thresholding estimator for the coefficients in wavelet regression; (iii) estimating the graphical structure in a Gaussian graphical model on the basis of the conditional probability of edge non-negligibility.

4.2 Multivariate sparsity

To extend the definition of sparsity of McC&P to sequences of distributions defined on $\mathbb{R}^d$, $d > 1$, we first introduce the definition of exceedance measure and Lévy-integrable function for the d-dimensional case.

Definition 4.2.1. A non-negative measure H_d defined on $(\mathbb{R}^d \setminus \{\mathbf{0}\}; \|\cdot\|)$ is termed an exceedance measure if $\int_{\mathbb{R}^d\setminus\{\mathbf{0}\}}(\|x\|^2 \wedge 1)H_d(dx) < \infty$. A measure satisfying $\int_{\mathbb{R}^d\setminus\{\mathbf{0}\}}(1 - e^{-\|x\|^2/2})H_d(dx) = 1$ is called a unit exceedance measure.

Here $\|\cdot\|$ is any norm induced by an inner product $\langle\cdot,\cdot\rangle$ defined on $\mathbb{R}^d$. As in the univariate case (see §2.1 in McC&P), there is a one-to-one correspondence between the exceedance measure H_d and the Lévy measure of an infinitely divisible distribution on $(\mathbb{R}^d; \|\cdot\|)$.

Definition 4.2.2. The space $\mathcal{W}_d^{\#}$ of Lévy-integrable functions on $\mathbb{R}^d$ consists of bounded and continuous functions $w : \mathbb{R}^d \to \mathbb{R}$ such that $\|x\|^{-2} w(x)$ is also bounded and continuous. Lévy-integrability implies $\int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} w(x)\, H_d(dx) < \infty$ for every $w \in \mathcal{W}_d^{\#}$ and every exceedance measure H_d.

We now naturally extend Definition 0.0.3 to a sequence of d-dimensional probability distributions $\{P_{\nu,d}\}_\nu$ which, as $\nu \to 0$, converges to the Dirac delta measure at the origin $\{\mathbf{0}\}$.

Definition 4.2.3. A sequence of probability distributions $\{P_{\nu,d}\}_\nu$, defined on $(\mathbb{R}^d; \|\cdot\|)$, is said to have a sparse limit with rate ρ_ν if there exists a unit exceedance measure H_d such that

$$\lim_{\nu\to 0} \rho_\nu^{-1} \int_{\mathbb{R}^d} w(x)\, P_{\nu,d}(dx) = \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} w(x)\, H_d(dx),$$

for every function $w \in \mathcal{W}_d^{\#}$.

This definition of multivariate sparsity is very general and comprises of many different families of distributions $\{P_{\nu,d}\}_\nu$. In this part of the thesis, we mostly study d-dimensional sparse measures which are product of scalar sparse measures. In Part III instead, we derive some theory for those d-dimensional sparse measures which are rotationally invariant and their sparsity is induced by the sparsity of their radial part. Clearly, these are just two instances of multivariate sparsity, and future research could be directed to study other different kinds of multivariate sparsity.

Before proceeding introducing component-wise sparsity, we define the d-dimensional analogs to the zeta function and zeta measure, which will appear in both this part of the thesis and the next.

For any given d-dimensional exceedance measure H_d on $(\mathbb{R}^d \setminus \{\mathbf{0}\}; \|\cdot\|)$, its zeta transform is

$$\zeta_d(y) = \int_{\mathbb{R}^d \setminus \{\mathbf{0}\}} (\cosh(\langle y, x\rangle) - 1) e^{-\|x\|^2/2} \, H_d(dx)\,, \tag{4.1}$$

while the associated zeta measure defined on $(\mathbb{R}^d \setminus \{\mathbf{0}\}; \|\cdot\|)$, is the integrand

$$\zeta_d(dx; y) = (\cosh(\langle y, x\rangle) - 1) e^{-\|x\|^2/2} H_d(dx)\,. \tag{4.2}$$

4.3 Component-wise sparsity

As anticipated in the introduction to this chapter, component-wise sparsity refers to those d-dimensional measures which are product of scalar sparse measures. Since this means that the d components of the random vector are independent, it is natural to take the inner product on $\mathbb{R}^d$ to be the standard Euclidean inner product $\langle y, x\rangle = y'x = \sum_{i=1}^d y_i x_i$. We now give a more formal definition of component-wise sparsity.

Definition 4.3.1. Let $P_{\nu,d}$ be a sequence of distributions defined on $(\mathbb{R}^d; \|\cdot\|)$, where $\|x\|^2 =$

$\sum_{i=1}^{d} x_i^2$. If it is possible to write

$$P_{\nu,d}(dx) = \prod_{i=1}^{d} P_\nu^i(dx_i)\,,$$

where, for each $i = 1, \ldots, d$, the one-dimensional sequence $\{P_\nu^i\}_\nu$ is symmetric around zero and sparse according to Definition 0.0.3, with rate $\rho_\nu^i = c_i \rho_\nu$, $c_i > 0$, and unit exceedance measure H^i, then we say that $P_{\nu,d}$ is component-wise sparse.

As a matter of fact, if $\{P_{\nu,d}\}_\nu$ is component-wise sparse then it is also sparse according to Definition 4.2.3. Indeed for any function $w \in \mathcal{W}_d^{\#}$, one has

$$\lim_{\nu \to 0} \rho_\nu^{-1} \int_{\mathbb{R}^d} w(x) \prod_{i=1}^{d} P_\nu^i(dx_i) = \int_{\mathbb{R}^d \setminus \{\mathbf{0}\}} w(x)\, H_d(dx),$$

where the rate $\rho_\nu = \rho_{\nu,d}$ is proportional to the sum of the rates of the scalar sparse measures, ρ_ν^i, while the d-dimensional unit exceedance measure defined on $\mathbb{R}_d \setminus \{\mathbf{0}\}$ is

$$H_d(dx) = \frac{1}{d} \sum_{i=1}^{d} H^i(dx_i) \prod_{j \neq i} \delta_0(dx_j)\,. \tag{4.3}$$

This exceedance measure is concentrated along the Cartesian axes, so it is singular with respect to Lebesgue measure defined on $\mathbb{R}^j$, for any $j \geq 2$. Indeed, integrating a function w defined on $\mathbb{R}^d$ against H_d is the same as projecting the function on each Cartesian axis,

$$w_i(x_i) = w((0, \ldots, 0, x_i, 0, \ldots, 0))\,,$$

integrating $w_i(x_i)$ against H^i, and then average the d integrals.

We check that H_d is a d-dimensional Lévy measure by computing the integral

$$\int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\|x\|^2 \wedge 1)\, H_d(dx) = \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\|x\|^2 \wedge 1) \left(\frac{1}{d} \sum_{i=1}^{d} H^i(dx_i) \prod_{j\neq i} \delta_0(dx_j) \right)$$
$$= \frac{1}{d} \sum_{i=1}^{d} \int_{\mathbb{R}\setminus\{0\}} (x_i^2 \wedge 1)\, H^i(dx_i)\,,$$

which is finite as long as, for each $i = 1, \ldots, d$, H^i is a Lévy measure.

The characteristic function of $P_{\nu,d}$ is, up to first-order sparsity, the same as the characteristic function of the infinitely divisible distribution on $\mathbb{R}^d$, having some scalar multiple of H_d as its Lévy measure Λ,

$$\begin{aligned}
\int e^{it'x}\, P_{\nu,d}(dx) &= \prod_{k=1}^{d} \int e^{it_k x_k}\, P_\nu^k(dx_k) \\
&= \prod_{k=1}^{d} \left(1 + \rho c_k \int (\cos(t_k x_k) - 1)\, H^k(dx_k) + o(\rho) \right) \\
&= \prod_{k=1}^{d} \left(e^{\rho c_k \int_{\mathbb{R}\setminus\{0\}} (\cos(t_k x_k)-1) H^k(dx_k)} + o(\rho) \right) \\
&= \exp\left\{ \rho \sum_{k=1}^{d} c_k \int_{\mathbb{R}\setminus\{0\}} (\cos(t_k x_k) - 1)\, H^k(dx_k) \right\} + o(\rho) \\
&= \exp\left\{ \rho \int_{\mathbb{R}_d\setminus\{\mathbf{0}\}} (\cos(t'x) - 1) \sum_{k=1}^{d} c_k\, H^k(dx_k) \prod_{j\neq k} \delta_0(dx_j) \right\} + o(\rho) \\
&= \exp\left\{ \rho \int_{\mathbb{R}_d\setminus\{\mathbf{0}\}} (\cos(t'x) - 1)\, \Lambda(dx) \right\} + o(\rho)\,.
\end{aligned}$$

An infinitely divisible distribution on $\mathbb{R}^d$ has Lévy measure proportional to H_d if and only if it has independent components (see Sato, 1999 [63], and Samorodnitsky and Taqqu, 1994 [62]). As a matter of fact, H_d can be written as the sum of d measures, each concentrated along one Cartesian axis,

$$\frac{1}{d} \sum_{i=1}^{d} \left(\frac{1}{2} \delta_{s_{i,+}}(d\,\tilde{x}) + \frac{1}{2} \delta_{s_{i,-}}(d\,\tilde{x}) \right) H^{R,i}(d\,\|x\|)\,.$$

Here $\tilde{x} = x/\|x\|$ is the direction of the vector on the unit sphere $\mathcal{S}^d = \{z : \|z\| = 1\}$, while, for each $i = 1, \ldots, d$,

$$s_{i,+} = (0, \cdots, 0, +1, 0, \cdots, 0) \in \{0, +1\}^d$$

and

$$s_{i,-} = (0, \cdots, 0, -1, 0, \cdots, 0) \in \{0, -1\}^d$$

denote the two intersections of the unit sphere with the i^{th} Cartesian axis. On the other hand, $H^{R,i}$ is the radial exceedance measure on $(0, \infty)$ corresponding to the scalar sparse measure P^i_ν, for each $i = 1, \ldots, d$. We consider $H^{R,i}$ to be two times the positive part of the symmetric exceedance measure H^i corresponding to P^i_ν.

Notice that if the d independent components of x are also identically distributed, then $P^i_\nu = P_\nu$ for all $i = 1, \ldots, d$. This means that we can write $H_d(dx)$ as

$$\frac{1}{d} \sum_{i=1}^{d} \left(\frac{1}{2} \delta_{s_{i,+}}(d\,\tilde{x}) + \frac{1}{2} \delta_{s_{i,-}}(d\,\tilde{x}) \right) \cdot H^R(d\,\|x\|)\,. \tag{4.4}$$

Thus H_d factorizes into a spectral measure on $\mathcal{S}^d$,

$$\Gamma(d\tilde{x}) = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{1}{2} \delta_{s_{i,+}}(d\,\tilde{x}) + \frac{1}{2} \delta_{s_{i,-}}(d\,\tilde{x}) \right)\,,$$

which is discrete and concentrated on the intersections of the axes with the unit sphere, and a radial measure on $(0, \infty)$,

$$H^R(d\,\|x\|)\,,$$

which is just two times the positive part of the exceedance measure H of the scalar sparse components. In Part III, we will see that the exceedance measure for vector-sparse distributions, i.e., sparse distributions that are rotationally invariant, can also be factorized in a spectral measure on $\mathcal{S}^d$ and a radial measure on $(0, \infty)$.

From now on, unless differently specified, we will consider the case where $P_\nu^i = P_\nu$ for all components $i = 1, \ldots, d$, so that the rate $\rho_{\nu,d}$ is $d\rho_\nu$, and the unit exceedance measure is

$$H_d(dx) = \frac{1}{d} \sum_{i=1}^{d} H(dx_i) \prod_{j \neq i} \delta_0(dx_j) \,.$$

4.3.1 Component-wise inverse-power exceedance

If the scalar exceedance measure H is the inverse-power measure

$$H(dx) = K_\alpha |x|^{-\alpha-1} dx \,,$$

with $K_\alpha = \frac{\alpha\, 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}$ and $\alpha \in (0,2)$, then

$$H_d(dx) = K_{d,\alpha} \sum_{i=1}^{d} |x_i|^{-\alpha-1} \prod_{j \neq i} \delta_0(dx_j) \tag{4.5}$$

is the component-wise inverse power exceedance measure. Here

$$K_{d,\alpha} = \frac{K_\alpha}{d} = \frac{1}{d} \frac{\alpha\, 2^{\alpha/2-1}}{\Gamma(1-\alpha/2)}$$

is the scalar such that H_d is a unit exceedance measure,

$$\int (1 - e^{-\|x\|^2/2}) H_d(dx) = \frac{1}{d} \sum_{i=1}^{d} \int (1 - e^{-x_i^2/2}) K_\alpha |x_i|^{-\alpha-1} = 1 \,.$$

The measure in (4.5) is, up to a multiplicative constant, the Lévy measure associated with the symmetric α-stable ($S\alpha S$) process having independent components (see Sato, 1999 [63], and Samorodnitsky and Taqqu, 1994 [62]), whose characteristic function is the exponential

of

$$\int_{\mathcal{S}^d}\int_0^\infty (e^{irt's}-1)\frac{dr}{r^{\alpha+1}}\Gamma(ds) = -\int_{\mathcal{S}^d}|t's|^\alpha \sum_{i=1}^d\left(\frac{1}{2}\delta_{s_{i,+}}(ds)+\frac{1}{2}\delta_{s_{i,-}}(ds)\right) = -\sum_{i=1}^d |t_i|^\alpha .$$

In particular, when $\alpha = 1$, H_d is proportional to the Lévy measure of the product of d Cauchy distributions

$$P_d(dx) = \prod_{i=1}^d \frac{dx_i}{\pi(1+x_i^2)} .$$

4.3.2 Component-wise zeta function

Following the definition given in Section 4.2, the zeta transform of a component-wise exceedance measure is

$$\begin{aligned}
\zeta_d(y) &= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(\langle y,x\rangle)-1)e^{-\|x\|^2/2}\, H_d(dx) \\
&= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(y'x)-1)e^{-x'x/2}\left(\frac{1}{d}\sum_{i=1}^d H(dx_i)\prod_{j\neq i}\delta_0(dx_j)\right) \\
&= \frac{1}{d}\sum_{i=1}^d \int_{\mathbb{R}\setminus\{\mathbf{0}\}} (\cosh(y_ix_i)-1)e^{-x_i^2/2}\, H(dx_i) \\
&= \frac{1}{d}\sum_{i=1}^d \zeta(y_i) ,
\end{aligned}$$

where $\zeta(y)$ is the usual univariate zeta function. Similarly, the corresponding d-dimensional zeta measure is

$$\zeta_d(dx;y) = \frac{1}{d}\sum_{i=1}^d \zeta(dx_i;y_i)\prod_{j\neq i}\delta_0(dx_j) .$$

Notice that the component-wise zeta measure has the same structure of the component-wise exceedance measure. So when integrating against it, we project the function onto each Cartesian axis, integrate the one-dimensional function $w_i(x_i)$ against $\zeta(dx_i;y_i)$, and then average the d integrals.

When the scalar exceedance H is the inverse-power measure, i.e., H_d is as in (4.5), then one can use the Taylor series for the univariate zeta transform of H, and write the zeta transform of H_d as

$$\zeta_d(y) = \frac{\alpha/2}{\Gamma(1-\alpha/2)} \frac{1}{d} \sum_{i=1}^{d} \left(\sum_{r=1}^{\infty} \frac{y_i^{2r}}{(2r)!} 2^r \Gamma(r-\alpha/2) \right) .$$

4.4 Component-wise sparse signal plus noise

In this section, we study the signal-plus-noise model in the case when the observations are independent and identically distributed, and can be written as

$$Y_i = \mu_i + \eta_i \,, \qquad i = 1, \ldots, n \,.$$

The signals μ_i are independent, with common distribution P_ν having sparsity pair (ρ, H), while η_i are independent standard Gaussian random variables. Using vector notation, letting $Y, \mu, \eta \in \mathbb{R}^n$, we can equivalently write

$$Y = \mu + \eta \,,$$

where the signal vector $\mu \sim P_{\nu,n}$, is component-wise sparse with scalar rate ρ and scalar exceedance measure H, and is independent of $\eta \sim N_n(0, I_n)$.

Then the first-order sparse approximation of the marginal density for Y at y is

$$\begin{aligned} m_{n,\nu}(y) &= \phi_n(y) \int e^{x'y} e^{-\|x\|^2/2} \, P_{\nu,n}(dx) \\ &= \phi_n(y) \left(n\rho \int (\cosh(x'y) - 1) e^{-\|x\|^2/2} \, H_n(dx) + 1 - n\rho \right) + o(n\rho) \\ &= \phi_n(y) \Big(n\rho \, \zeta_n(y) + 1 - n\rho \Big) + o(n\rho) \\ &= \phi_n(y) \Big(\rho \sum_{i=1}^{n} \zeta(y_i) + 1 - n\rho \Big) + o(n\rho) \, . \end{aligned}$$

Notice that this expression is equivalent, in first-order sparsity, to the product of the sparse approximations for the univariate $m_\nu(y_i)$,

$$m_\nu(y_i) = \phi(y_i)(\rho\zeta(y_i) + 1 - \rho) + o(\rho) \, .$$

In fact,

$$\begin{aligned} \prod_{i=1}^{n} m_\nu(y_i) = \prod_{i=1}^{n} \phi(y_i)(\rho\zeta(y_i) + 1 - \rho + o(\rho)) &= \phi_n(y) \prod_{i=1}^{n} (\rho\zeta(y_i) + 1 - \rho + o(\rho)) \\ &= \phi_n(y) \big(\rho\zeta(y_1) + 1 - \rho\big) \big(\rho\zeta(y_2) + 1 - \rho\big) \ldots \big(\rho\zeta(y_n) + 1 - \rho\big) + o(\rho) \\ &= \phi_n(y) \Big(\rho \sum_{i=1}^{n} \zeta(y_i) + 1 - n\rho \Big) + o(\rho) \, . \end{aligned}$$

4.4.1 Estimation of sparsity parameters

If one considers P_ν to be a scale sparse measure, then its exceedance measure is the inverse-power measure $H(dx) \propto |x|^{-\alpha-1}$ for $\alpha \in (0, 2)$. In practice, both ρ and α are unknown parameters and need to be estimated. In applied work, it is important that the estimates for ρ and α are not dependent of the scale on which the observations are made, as a change of scale simply corresponds to a change in the units of measurement.

Therefore, suppose that the actual observations are

$$\tilde{Y}_i = \tau Y_i\,, \qquad i = 1, \ldots, n\,,$$

where each $Y_i = \mu_i + \eta_i$ is scaled by a factor $\tau \neq 0$, so we can say that the group $\tau : y \mapsto \tau y$ acts on the observation space component-wise. Then each $\tilde{Y}_i$ has marginal density, in first-order sparse approximation, given by

$$1/|\tau|\, \phi(\tilde{y}/|\tau|)(1 - \rho + \rho\zeta(\tilde{y}/|\tau|)) + o(\rho)\,.$$

This marginal density is parametrized by the triplet $\theta = (\rho, \alpha, \tau^2)$ corresponding to the sparsity rate, the activity index of the exceedance measure and the scale of the observations. The sparse approximation to the log-likelihood for θ based on the scaled observations $\tilde{y}_1, \ldots, \tilde{y}_n$ is

$$l(\theta; \tilde{y}_1, \ldots, \tilde{y}_n) = -n\, \log(|\tau|) - \frac{1}{2}\sum_i \tilde{y}_i^2/\tau^2 + \sum_i \log(1 - \rho + \rho\zeta(\tilde{y}_i/|\tau|))\,, \qquad (4.6)$$

so that maximization of (4.6) with respect to θ delivers the maximum likelihood (ML) estimate. The ML estimator, as a function from the observation space $\mathcal{Y}$ to the parameter space Θ, is equivariant under the group action $\tau : (\rho, \alpha, \sigma^2) \mapsto (\rho, \alpha, \tau^2\sigma^2)$ for $\tau \neq 0$. In fact, given $y \in \mathbb{R}^n$, denoting by

$$\mathrm{ML}(y) = (\rho(y), \alpha(y), \sigma^2(y))$$

the maximum likelihood estimator for (ρ, σ^2, α), it is not hard to show that

$$\mathrm{ML}(\tau y) = (\rho(\tau y), \alpha(\tau y), \sigma^2(\tau y),) = (\rho(y), \alpha(y), \tau^2\sigma^2(y)) = |\tau|\, \mathrm{ML}(y)\,.$$

This means that, if we include the scale parameter in our estimation, then the maximum likelihood estimates for the sparsity parameters, the rate ρ and the inverse-power α, do not change if the observations are measured in different units. For this reason, in the following

chapters, we will always estimate the triplet (ρ, α, σ^2).

4.4.2 Signal conditional distribution

The first-order sparse approximation to the conditional distribution of the n-dimensional signal μ, given y, is

$$\begin{aligned} P_{\nu,n}(dx \mid y) &= \frac{n\rho\zeta_n(dx; y) + e^{-x'x/2} P_{\nu,n}(dx)}{n\rho\zeta_n(y) + 1 - n\rho} + o(n\rho) \\ &= \frac{\rho \sum_{i=1}^n \zeta(dx_i; y_i) \prod_{j \neq i} \delta_0(dx_j) + \prod_{i=1}^n e^{-x_i^2/2} P_\nu(dx_i)}{\rho \sum_{i=1}^n \zeta(y_i) + 1 - n\rho} + o(n\rho)\,. \end{aligned}$$

If we look at the n signals as a vector in $\mathbb{R}^n$, then we can compute the first-order sparse approximation to the conditional probability that the norm of the signal $\|\mu\|$ is greater than some threshold ϵ. In this case, supposing that the one-dimensional H is atom-free and does not have Gaussian nor sub-Gaussian tails, then for any positive ϵ, approximating $\chi_{\|x\|>\epsilon}$ with $1 - e^{-\|x\|^2/2\epsilon^2}$,

$$\begin{aligned} P_{n,\nu}(\|\mu\| > \epsilon \mid y) &\approx \int (1 - e^{-\|x\|^2/2\epsilon^2})\, P_{n,\nu}(dx \mid y) \\ &= \frac{\sum_{i=1}^n \int (1 - e^{-x_i^2/2\epsilon^2}) \left(\rho\zeta(dx_i; y_i) + e^{-x_i^2/2} \rho H(dx_i) \right)}{\rho \sum_{i=1}^n \zeta(y_i) + 1 - n\,\rho} + o(\rho)\,. \end{aligned}$$

The reason why H cannot have Gaussian or sub-Gaussian tails is explained in more detail in Section 8.5.2 of Chapter 8. In this context, not to get a trivial zero limit as $\rho \to 0$, we need a double limit regime under which, for all $i = 1, \ldots, n$, $\rho \to 0$ and $y_i \to \infty$ in such a way that $\rho\zeta(y_i) \to \lambda_i > 0$. In this case, the conditional probability of $\{\|\mu\| > \epsilon\}$, for a component-wise sparse signal, converges to

$$\frac{\rho \sum_{i=1}^n \zeta(y_i)}{1 + \rho \sum_{i=1}^n \zeta(y_i)}\,.$$

4.5 Negligibility

At the end of the previous section, a sparse double limit is given for the conditional probability of the event $\{\|\mu\| > \epsilon\}$. However, when the components of the signal are assumed to be independent, each with a sparse distribution, it is common in many statistical applications to be interested in establishing the 'activity/negligiblity' of each individual signal separately. In this section, we introduce a formal notion of signal negligibility, which in fact refers to any unidimensional sparse signal. So for the rest of this chapter, the number of observations (hence the number of signals) can really be just one, since the signal negligibility, as much as the sparsity of its distribution, is thought as a limiting notion, driven by the sparsity parameter rather than by the sample size getting large. To make this section self-contained, we report here part of the discussion presented in the introduction, when we first talked about negligibility.

Even if in the literature, there is not a universal consensus nor a formal mathematical definition of what constitutes signal activity, fairly often, in both theoretical and applied work, the dichotomy of signal non-activity/activity refers to the events that the signal be zero or not zero (see for instance, Efron, 2007, Johnstone and Silverman, 2004). However, in the sparsity theory developed in McC&P, the probability that the random signal with a sparse distribution, is exactly zero is not identifiable only from the sparsity pair (ρ, H). Indeed, when implied by the $\mathcal{W}^{\#}$-integral definition 0.0.3, the sparse approximation

$$P_{\nu}(\epsilon^{+}) = \rho H(\epsilon^{+}) + o(\rho)$$

only holds for strictly positive thresholds ϵ. Here $F(z^{+})$ is the probability that the random variable $X \sim F$ exceeds z, in absolute value. To see this, consider two sparse families such

as the spike-and-scaled slab

$$P_\nu^1(dx) = (1-\sqrt{\nu})\,\frac{e^{-|x|/\nu}}{2\nu}\,dx + \sqrt{\nu}\,\frac{\nu}{\pi(x^2+\nu^2)}\,dx\,,$$

and the atom-and-scaled slab

$$P_\nu^2(dx) = (1-\sqrt{\nu})\,\delta_0(dx) + \sqrt{\nu}\,\frac{\nu}{\pi(x^2+\nu^2)}\,dx\,.$$

These two sparse measures share the same first-order sparsity pair given by $\rho = \nu^{3/2}\sqrt{2/\pi}$ and $H(dx) = 1/\sqrt{2\pi}\,|x|^{-2}dx$. Therefore the first-order sparse approximations to the expectation of any function $w \in \mathcal{W}^{\#}$ with respect to P_ν^1 and P_ν^2, are exactly the same, insofar they are solely based on the pair (ρ, H). Yet, the two measures do give very different probability mass to the atom at zero, as $P_\nu^2(X=0) = 0$ while $P_\nu^2(X=0) = 1-\sqrt{\nu}$.

The reason why this happens is that the indicator function $\chi_{x=0}$ is a discontinuous function at the limit point and such discontinuity at the limit opens the door to different answers from the sparse measures even if their limiting behavior in approaching the Dirac delta limit is the same.

The way we propose to circumvent this zero-non-identifiability issue is to take a slightly different perspective in looking at the problem, and adopt a strategy that is similar in spirit to the "limit-approaching" standpoint, from which the sparsity theory by McC&P has been formulated in first place. The idea is to look at the atom $\{0\}$ as the limit point of a sequence of intervals $[-\epsilon_\nu, \epsilon_\nu]$, where $\epsilon_\nu \to 0$ as $\nu \to 0$, and to describe, as the limit takes place, the approaching behavior of the moving sequence of measures P_ν over the moving region $[-\epsilon_\nu, \epsilon_\nu]$. So, instead of asking for the probability at the limit point $P_\nu(X=0)$,

$$P_\nu(X=0) = \int \chi_{x=0}\,P_\nu(dx)\,,$$

which requires the expectation of a discontinuous function at zero, we ask for the probability that the signal is in a region converging to the limit point, which by contrast, can be approximated with arbitrary precision, by the expectation of a bounded and continuous function, such as

$$\int e^{-x^2/2\epsilon_\nu^2} P_\nu(dx)\,.$$

With some conditions on the speed of convergence to zero of the threshold sequence ϵ_ν, we can use the sparsity integral definition in (1) to obtain the sparse-negligibility approximation

$$\int e^{-x^2/2\epsilon_\nu^2} P_\nu(dx) = 1 - \rho \int (1 - e^{-x^2/2\epsilon_\nu^2})\, H(dx)\,,$$

where this approximation holds with an error $\rho \int (1 - e^{-x^2/2\epsilon_\nu^2})\, H(dx)$.

All this leads us to introduce a mathematical definition of signal negligibility. In what follows, we use the notation $F(w_z) = \int (1 - e^{-x^2/2z^2})\, F(dx)$, where F is some non-zero measure.

Definition 4.5.1. Let $\{P_\nu\}_\nu$ be a sparse sequence of symmetric distributions on $\mathbb{R}$ with sparsity pair (ρ, H), and let $\{\epsilon_\nu\}_\nu$ be a sequence of strictly positive thresholds, $\epsilon_\nu > 0$. We say that ϵ_ν is a negligibility sequence for P_ν if, as $\nu \to 0$,

1. $\epsilon_\nu \to 0\,,$
2. $P_\nu(w_{\epsilon_\nu}) \to 0\,,$
3. $P_\nu(w_{\epsilon_\nu}) = \rho H(w_{\epsilon_\nu}) + o(\rho H(w_{\epsilon_\nu}))\,.$

Given $X \sim P_\nu$, we say that X is negligible if $|X| \leq \epsilon_\nu$.

Such a sequence exists for every sparse family and, because of condition 2 in Definition 4.5.1, a signal $X \sim P_\nu$ is negligible with probability converging to one as $\nu \to 0$. The

reason why we need the third condition is that for some sparse families, $P_\nu(w_{\epsilon_\nu}) \to 0$ by itself does not impose any requirement on the negligibility sequence, so that it is not sufficient to guarantee the sparse-negligibility approximation $P_\nu(w_{\epsilon_\nu}) \sim \rho H(w_{\epsilon_\nu})$. An example of when this occurs is for the sparse family

$$P_\nu^2(dx) = (1-\sqrt{\nu})\,\delta_0(dx) + \sqrt{\nu}\,\frac{\nu}{\pi(x^2+\nu^2)}\,dx\,.$$

In fact,

$$P_\nu^2(w_{\epsilon_\nu}) = \sqrt{\nu}\int (1-e^{-x^2/2\epsilon_\nu^2})\frac{\nu}{\pi(x^2+\nu^2)}\,dx = \sqrt{\nu}\left(1-e^{\nu^2/2\epsilon_\nu^2}\ \mathrm{Erfc}(\nu/\sqrt{2}\epsilon_\nu)\right)\,,$$

where $\mathrm{Erfc}(z) = 1-\frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2/2}\,dt$. Now, this expression converges to zero as $\nu \to 0$, regardless of the behavior of ϵ_ν/ν. However, in order to have $P_\nu^2(w_{\epsilon_\nu}) = \rho H(w_{\epsilon_\nu}) + o(\rho H(w_{\epsilon_\nu}))$, given that

$$\rho H(w_{\epsilon_\nu}) = \nu^{3/2}\sqrt{2/\pi}\cdot\int (1-e^{-x^2/2\epsilon_\nu^2})\frac{1}{\sqrt{2\pi}x^2}\,dx = \sqrt{\frac{2}{\pi}}\,\nu^{3/2}\epsilon_\nu^{-1}\,,$$

one needs $\epsilon_\nu/\nu \to \infty$, in which case,

$$\sqrt{\nu}\left(1-e^{\nu^2/2\epsilon_\nu^2}\ \mathrm{Erfc}(\nu/\sqrt{2}\epsilon_\nu)\right) = \sqrt{\nu}\left(\sqrt{2/\pi}\cdot\nu/\epsilon_\nu + o(\nu/\epsilon_\nu)\right) = \sqrt{\frac{2}{\pi}}\,\nu^{3/2}\epsilon_\nu^{-1} + o(\nu^{3/2}/\epsilon_\nu)\,.$$

Now, the second and third conditions in Definition 4.5.1 require that

$$\rho H(w_{\epsilon_\nu}) \to 0\,,$$

even though $\epsilon_\nu \to 0$ might lead to $H(w_{\epsilon_\nu}) \to \infty$, if H is not finite. For instance, if H is the inverse-power measure, then $H(w_{\epsilon_\nu})$ is

$$\int (1-e^{-x^2/2\epsilon_\nu^2})K_\alpha\,|x|^{-\alpha-1}\,dx = \epsilon_\nu^{-\alpha}\,,$$

which indeed goes to infinity as $\epsilon_\nu \to 0$, since $\alpha \in (0,2)$. Thus, $\rho H(w_{\epsilon_\nu}) \to 0$ if and only if

$$\epsilon_\nu^\alpha = g(\rho)\,,$$

where $g(\rho) \to 0$ and $\rho/g(\rho) \to 0$, as $\rho \to 0$. Therefore, if there are no more stringent requirements on ϵ_ν from the other conditions, then one can choose the sequence ϵ_ν to behave, for instance, like $\log(1/\rho)^{-\eta/\alpha}$, for some positive η. In this way, $\rho\epsilon_\nu^{-\alpha} \sim \rho\log(1/\rho)^\eta$, and this converges to zero as $\rho \to 0$.

To see the advantage of considering signal negligibility rather than signal nullity, we look again at the two sparse families presented at the beginning of the section,

$$P_\nu^1(dx) = (1-\sqrt{\nu})\,\frac{e^{-|x|/\nu}}{2\nu}\,dx + \sqrt{\nu}\,\frac{\nu}{\pi(x^2+\nu^2)}\,dx\,,$$

and

$$P_\nu^2(dx) = (1-\sqrt{\nu})\,\delta_0(dx) + \sqrt{\nu}\,\frac{\nu}{\pi(x^2+\nu^2)}\,dx\,.$$

As already highlighted, despite having the same sparsity rate $\rho = \nu^{3/2}\sqrt{2/\pi}$ and same exceedance measure $H(dx) = |x|^{-2}/\sqrt{2\pi}dx$, the two families give different mass at the atom at zero: $P_\nu^1(X=0) = 0$ while $P_\nu^2(X=0) = 1-\sqrt{\nu}$. Instead, in terms of negligibility, one has that

$$\begin{aligned} P_\nu^1(w_{\epsilon_\nu}) &= (1-\sqrt{\nu})\int(1-e^{-x^2/2\epsilon_\nu^2})\frac{e^{-|x|/\nu}}{2\nu}\,dx + \sqrt{\nu}\int(1-e^{-x^2/2\epsilon_\nu^2})\frac{\nu}{\pi(x^2+\nu^2)}\,dx \\ &= (1-\sqrt{\nu})\left(1-\frac{\epsilon_\nu}{\nu}\sqrt{\pi/2}\,e^{\epsilon_\nu^2/2\nu^2}\,\mathrm{Erfc}(\epsilon_\nu/\sqrt{2}\nu)\right) + \sqrt{\nu}\,\left(1-e^{\nu^2/2\epsilon_\nu^2}\,\mathrm{Erfc}(\nu/\sqrt{2}\epsilon_\nu)\right)\,, \end{aligned}$$

and this last expression, provided that $\epsilon_\nu/\sqrt{\nu} \to \infty$, behaves as

$$(1-\sqrt{\nu})\left(\nu^2/\epsilon_\nu^2 + O(\nu^4/\epsilon_\nu^4)\right) + \sqrt{\nu}\,\left(\sqrt{2/\pi}\,\nu/\epsilon_\nu + O(\nu^2/\epsilon_\nu^2)\right) = \sqrt{2/\pi}\,\nu^{3/2}\,\epsilon_\nu^{-1} + O(\nu^2/\epsilon_\nu^2)$$

Now, if we compute the non-negligibility integral for the second family, given that $\epsilon_\nu/\nu \to \infty$, we have

$$\begin{aligned} P^2_\nu(w_{\epsilon_\nu}) &= \int (1 - e^{-x^2/2\epsilon^2_\nu}) \left((1-\sqrt{\nu})\, \delta_0(dx) + \sqrt{\nu}\, \frac{\nu}{\pi(x^2+\nu^2)}\, dx \right) \\ &= \sqrt{\nu} \int (1 - e^{-x^2/2\epsilon^2_\nu}) \frac{\nu}{\pi(x^2+\nu^2)}\, dx \\ &= \sqrt{\nu} \left(1 - e^{\nu^2/2\epsilon^2_\nu}\, \mathrm{Erfc}(\nu/\sqrt{2}\epsilon_\nu) \right) \\ &= \sqrt{2/\pi}\, \nu^{3/2}\, \epsilon_\nu^{-1} + O(\nu^{5/2}/\epsilon^2_\nu) \end{aligned}$$

Indeed,

$$\sqrt{2/\pi}\ \nu^{3/2}\, \epsilon_\nu^{-1} = \sqrt{2/\pi}\ \nu^{3/2} \cdot \int (1 - e^{-x^2/2\epsilon^2_\nu}) \frac{1}{\sqrt{2\pi} x^2}\, dx = \rho\, H(w_{\epsilon_\nu})\,,$$

so that, provided $\epsilon_\nu/\sqrt{\nu} \to \infty$,

$$P^1_\nu(w_{\epsilon_\nu}) = P^2_\nu(w_{\epsilon_\nu}) = 1 - \rho\, H(w_{\epsilon_\nu}) + o(\rho H(w_{\epsilon_\nu}))\,.$$

As already observed, in these cases of mixture measures, given $\epsilon_\nu \to 0$, one needs to ensure that both $P_\nu(w_{\epsilon_\nu}) \to 0$ and $P_\nu(w_{\epsilon_\nu}) \sim \rho H(w_{\epsilon_\nu})$, as $\nu \to 0$. By contrast, when the sparse measure is a scale sparse measure, conditions 2. and 3. of the negligibility definition 4.5.1 imply one another, and are met whenever $\epsilon_\nu/\nu \to \infty$. In fact, let $P_\nu(dx) = 1/\nu P(dx/\nu)$, where P is any symmetric distribution on the real line. Then, because in this case the distinction between hard and soft threshold functions is irrelevant, we can see that

$$P_\nu(\epsilon^+_\nu) = 2 \int_{\epsilon_\nu}^\infty P_\nu(dx) = 2 \int_{\epsilon_\nu/\nu}^\infty P(dx) \to 0\,,$$

if and only if $\epsilon_\nu/\nu \to \infty$. Moreover, because P_ν is scale sparse, the density of P is regularly varying at infinity so that its tail behavior is the same as that of the exceedance measure for

P_ν. So, provided $\epsilon_\nu/\nu \to \infty$, the tail integral of $P(dx)$ behaves as

$$P_\nu(\epsilon_\nu^+) = 2\int_{\epsilon_\nu/\nu}^{\infty} P(dx) = 2(\epsilon_\nu/\nu)^{-\alpha} + o(\nu^\alpha/\epsilon_\nu^\alpha)\,,$$

which in fact coincides with

$$\rho H_\nu(\epsilon_\nu^+) = 2\nu^\alpha \int_{\epsilon_\nu}^{\infty} |x|^{-\alpha-1} dx = 2\nu^\alpha \epsilon_\nu^{-\alpha}\,.$$

Therefore, $P_\nu(\epsilon_\nu^+) \sim \rho H_\nu(\epsilon_\nu^+)$ if and only if $P_\nu(\epsilon_\nu^+) \to 0$. This in turn means that, for scale sparse families, it is sufficient to choose the negligibility sequence ϵ_ν in such a way to guarantee $\rho H(\epsilon_\nu^+) \to 0$, i.e., $\epsilon_\nu/\nu \to \infty$.

Building upon this definition of negligibility, in the next section we show how to construct a new sparse integral approximation for the sparse measure P_ν. This alternative integral approximation will be less accurate than the $o(\rho)$ approximation, but will have a component given by the Dirac delta measure at zero.

4.5.1 A different integral expansion

Let P_ν be a sparse sequence of one-dimensional distributions with first-order sparsity pair (ρ, H) and let ϵ_ν be a negligibility sequence for P_ν. Because the signal $X \sim P_\nu$ is said to be negligible if $|X| \leq \epsilon_\nu$, once more, we approximate the hard-threshold function $\chi_{|x|\leq\epsilon_\nu}$ with the soft-threshold function $e^{-x^2/2\epsilon_\nu^2}$, and decompose P_ν into two parts

$$P_\nu(dx) = P_\nu(dx)e^{-x^2/2\epsilon_\nu^2} + P_\nu(dx)(1 - e^{-x^2/2\epsilon_\nu^2})\,. \tag{4.7}$$

The first component in the RHS of (4.7) is proportional to

$$\tilde{P}_\nu(dx) = \frac{e^{-x^2/2\epsilon_\nu^2} P_\nu(dx)}{\int e^{-x^2/2\epsilon_\nu^2}\, P_\nu(dx)}\,.$$

The measure $\tilde{P}_\nu$ is a sparse probability distribution converging weakly to the Dirac delta measure at zero exponentially fast. There is no definite sparsity rate for $\tilde{P}_\nu$ satisfying the integral definition of sparsity with a finite non-zero limit. Thus, for any bounded and continuous function, as $\nu \to 0$

$$\int_{\mathbb{R}} w(x)\, \tilde{P}_\nu(dx) \sim \int_{\mathbb{R}} w(x)\, \delta_0(dx)\,.$$

On the other hand, sparsity of P_ν implies

$$\int_{\mathbb{R}} w(x)(1 - e^{-x^2/2\epsilon_\nu^2})\, P_\nu(dx) \sim \rho \int_{\mathbb{R}\setminus\{0\}} w(x)(1 - e^{-x^2/2\epsilon_\nu^2})\, H(dx)\,,$$

so that the second component in (4.7) is, in the sense of integrals of bounded and continuous functions, equivalent to

$$\rho(1 - e^{-x^2/2\epsilon_\nu^2})H(dx)\,.$$

Thus, we write (4.7) as

$$\left(\int e^{-x^2/2\epsilon_\nu^2}\, P_\nu(dx)\right) \tilde{P}_\nu(dx) + \left(1 - \int e^{-x^2/2\epsilon_\nu^2}\, P_\nu(dx)\right) \frac{(1 - e^{-x^2/2\epsilon_\nu^2})P_\nu(dx)}{\int (1 - e^{-x^2/2\epsilon_\nu^2})\, P_\nu(dx)}\,. \tag{4.8}$$

Assuming P_ν to be a scale sparse measure, then H is the inverse-power exceedance measure so, as long as $\rho\epsilon_\nu^{-\alpha} \to 0$ as $\nu \to 0$, the negligibility integral is

$$\int_{\mathbb{R}} e^{-x^2/2\epsilon_\nu^2}\, P_\nu(dx) = 1 - \rho \int_{\mathbb{R}\setminus\{0\}} (1 - e^{-x^2/2\epsilon_\nu^2})\, H(dx) + o(\rho\epsilon_\nu^{-\alpha}) = 1 - \rho\epsilon_\nu^{-\alpha} + o(\rho\epsilon_\nu^{-\alpha})\,. \tag{4.9}$$

Therefore, taking the sparse approximation of each component appearing in (4.8), the integral of a bounded and continuous function w against the scale sparse measure P_ν is asymptotically equivalent to the integral of w against

$$(1 - \rho\epsilon_\nu^{-\alpha})\delta_0(dx) + \rho\epsilon_\nu^{-\alpha}\tilde{H}(dx)\,, \tag{4.10}$$

with an error of order $\rho\epsilon_\nu^{-\alpha}$. Here

$$\tilde{H}(dx) = \frac{(1 - e^{-x^2/2\epsilon_\nu^2})H(dx)}{\int_{\mathbb{R}\setminus\{0\}}(1 - e^{-x^2/2\epsilon_\nu^2})\,H(dx)}$$

is the weighted exceedance measure, defined on $\mathbb{R} \setminus \{0\}$ and finite for every $\epsilon_\nu > 0$. Notice that the error in the integral approximation $o(\rho\epsilon_\nu^{-\alpha})$ is larger than the usual approximation error $o(\rho)$.

Therefore, negligibility is a "limit-approaching" notion, in the sense that:

1. the signal is small, but it is not declared to be necessarily zero;
2. at the same time, whether the signal is negligible or is exactly zero, it does not matter as long as expectations of bounded and continuous functions are involved.

4.5.2 Signal plus noise revisited

Given the alternative integral approximation to the scale sparse distribution P_ν in (4.10), we can derive the corresponding approximation for the relevant functionals in the one-dimensional signal-plus-noise model,

$$Y = \mu + \eta\,.$$

Here $\mu \sim P_\nu$ is sparse with sparsity pair (ρ, H_α), and is independent of $\eta \sim N(0,1)$. We

would like to stress that now the error is of order $o(\rho\epsilon_\nu^{-\alpha})$.

The joint probability of the signal μ and the symmetrized observation $|Y|$ can be approximated by

$$\begin{aligned}\frac{\mathbb{P}(\mu \in dx; |Y| \in dy)}{\phi(y)} &= \cosh(yx)e^{-x^2/2}P_\nu(dx) \\ &= \rho\epsilon_\nu^{-\alpha}\cosh(yx)e^{-x^2/2}\tilde{H}(dx) + (1-\rho\epsilon_\nu^{-\alpha})\delta_0(dx) + o(\rho\epsilon_\nu^{-\alpha})\,.\end{aligned} \tag{4.11}$$

It follows that the sparse approximation to the marginal of Y of order $\rho\epsilon_\nu^{-\alpha}$ is

$$\begin{aligned}\mathbb{P}(Y \in dy) &= \int \mathbb{P}(\mu \in dx; Y \in dy) \\ &= \phi(y)\left(\rho\epsilon_\nu^{-\alpha}\int \cosh(yx)e^{-x^2/2}\,\tilde{H}(dx) + (1-\rho\epsilon_\nu^{-\alpha})\right) + o(\rho\epsilon_\nu^{-\alpha})\,.\end{aligned}$$

Analogously to the zeta function, we introduce the A function defined as

$$A(y) = \int \cosh(yx)e^{-x^2/2}\,\tilde{H}(dx)\,,$$

which is the normalization constant of the A measure

$$A(dx; y) = \cosh(yx)e^{-x^2/2}\tilde{H}(dx)\,.$$

Since H is the inverse-power measure, $A(y)$ can be computed as

$$A(y) = \epsilon_\nu^\alpha(\zeta(y) - \tau^\alpha\zeta(y/\tau) - 1 + \tau^\alpha)\,,$$

where $\tau^2 = 1 + 1/\epsilon_\nu^2$, and $\zeta(y)$ is the usual zeta function (see the appendix). With this, we

can rewrite the above approximation to the marginal of Y as a mixture of two components

$$(1 - \rho\epsilon_\nu^{-\alpha})\phi(y) + \rho\epsilon_\nu^{-\alpha}\phi(y)A(y) + o(\rho\epsilon_\nu^{-\alpha})\,. \tag{4.12}$$

As for $\psi(y) = \phi(y)\zeta(y)$, also $\phi(y)A(y)$ is a probability density function, as it is non negative for all y and it integrates to one (see the appendix).

Figure 6.5 shows two comparisons: the left panels depict $A(y)$ together with $\zeta(y)$, plotted on the log scale; the right panels instead show the product $\phi(y)A(y)$ versus $\phi(y)\zeta(y)$. We choose ϵ_ν to behave like $\log(1/\rho)^{-\eta/\alpha}$, $\eta = \alpha/2+0.01$, so that $\rho\epsilon_\nu^{-\alpha} \sim \rho\log(1/\rho)^{\alpha/2+0.01}$. Now $A(y)$ depends on ρ through ϵ_ν, so in each panel we fix the value of α and consider the set of values $\{0.01, 0.025, 0.05, 0.10, 0.20\}$ for ρ, corresponding to the different colors ranging from dark blue to light yellow, respectively. One can notice that, even if the two functions, $A(y)$ and $\zeta(y)$, have the same tail behavior, their difference around the origin is clearly reflected in the difference of shape of $\phi(y)\zeta(y)$ and $\phi(y)A(y)$, where the first one is zero at zero and bimodal, while the second one is unimodal with its mode at zero.

Notice that, except for the error term, the approximation in (4.12) is equivalent to

$$(1 - \rho)\phi(y) + \rho\phi(y)\zeta(y) + o(\rho)\,,$$

as long as the sequence of negligibility thresholds ϵ_ν converges to zero in such a way that $\epsilon_\nu^{2-\alpha}y^2 = o(\rho\epsilon_\nu^{-\alpha})$. In fact, as $\epsilon_\nu \to 0$, $\tau \sim \epsilon_\nu^{-1}$ so that $y/\tau \sim y\epsilon_\nu$. Moreover, the zeta function at the origin behaves like a quadratic and this implies

$$A(y) = \epsilon_\nu^\alpha(\zeta(y) - \tau^\alpha\zeta(y/\tau) - 1 + \tau^\alpha) \sim \epsilon_\nu^\alpha\zeta(y) - y^2\epsilon_\nu^2 - \epsilon_\nu^\alpha + 1\,,$$

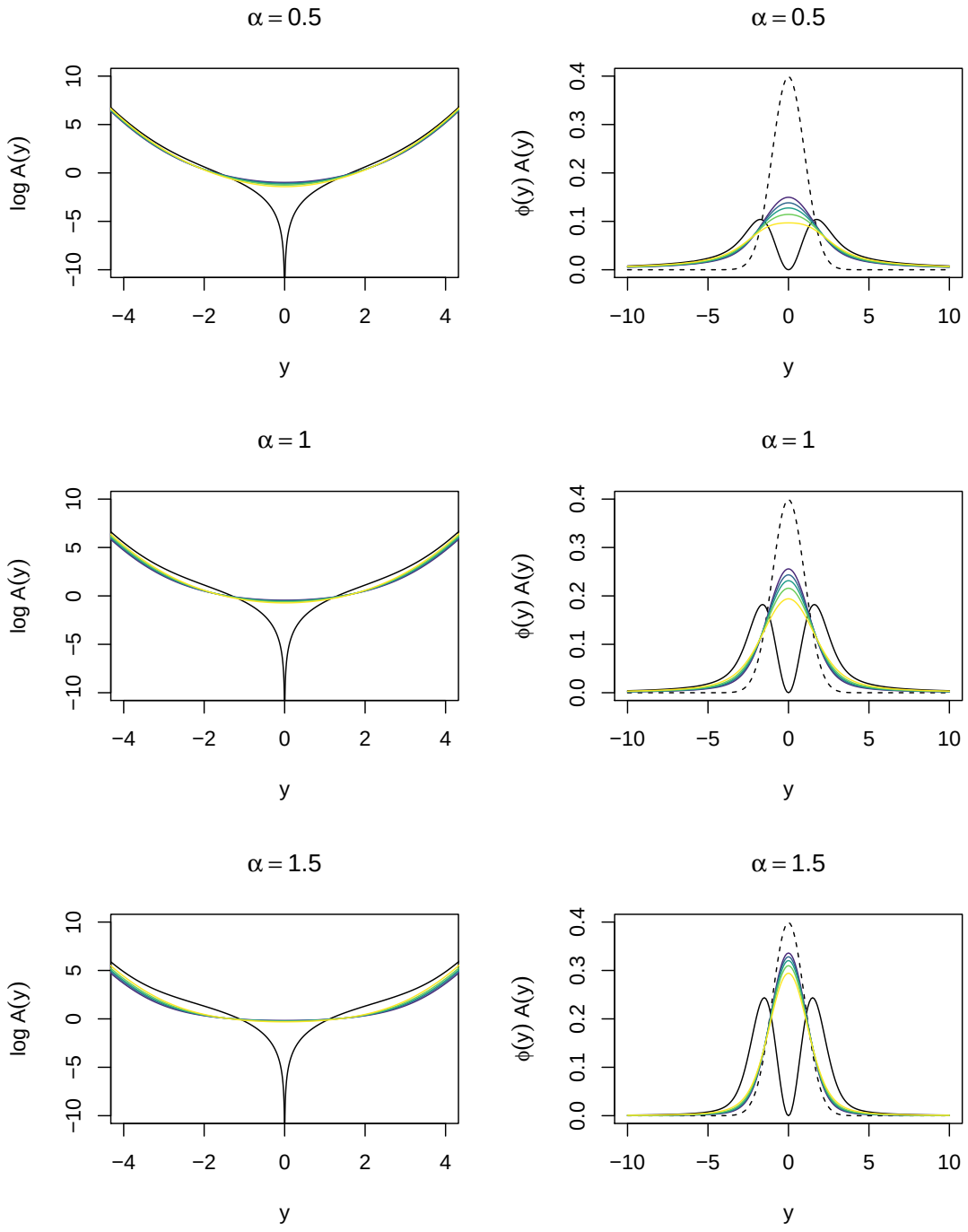


Figure 4.1: Left panels: comparison, on the log scale, between $\zeta(y)$ (black curve) and $A(y)$ (colored curves). Right panels: comparison between $\phi(y)\zeta(y)$ (black curve) and $\phi(y)A(y)$ (colored curves). Different colors for the $A(y)$ and $\phi(y)A(y)$ functions correspond to different values of ρ which imply different values of $\epsilon_\nu = (\log(1/\rho))^{-\eta/\alpha}$, with $\eta = \alpha/2 + 0.01$. From dark blue to light yellow, the ρ parameter is $0.01, 0.025, 0.05, 0.10, 0.20$. The dashed black line depicts the standard Gaussian density.

so that

$$\rho\epsilon_\nu^{-\alpha}A(y) + 1 - \rho\epsilon_\nu^{-\alpha} \sim \rho\zeta(y) - \rho\epsilon_\nu^{2-\alpha}y^2 - \rho + \rho\epsilon_\nu^{-\alpha} + 1 - \rho\epsilon_\nu^{-\alpha}$$
$$\sim \rho\zeta(y) + 1 - \rho + o(\rho\epsilon_\nu^{-\alpha})\,.$$

4.5.3 Signal conditional distribution revisited

From (4.11), it is easy to see that the sparse approximation to the conditional distribution of μ given $|Y|$ of order $\rho\epsilon_\nu^{-\alpha}$ is

$$\mathbb{P}(\mu \in dx \mid |Y| \in dy) = \frac{\rho\epsilon_\nu^{-\alpha}\cosh(yx)e^{-x^2/2}\tilde{H}(dx) + (1-\rho\epsilon_\nu^{-\alpha})\delta_0(dx)}{\rho\epsilon_\nu^{-\alpha}A(y) + 1 - \rho\epsilon_\nu^{-\alpha}} + o(\rho\epsilon_\nu^{-\alpha})\,. \tag{4.13}$$

Clearly, the approximation in (4.13) is meant in terms of integrals of bounded and continuous functions. It can also be written as a mixture of the Dirac-delta measure at zero and the normalized A measure, $\bar{A}(dx;y) = A(dx;y)/A(y)$,

$$w(y)\bar{A}(dx;y) + (1-w(y))\delta_0(dx) + o(\rho\epsilon_\nu^{-\alpha})\,,$$

where

$$w(y) = \frac{\rho\epsilon_\nu^{-\alpha}A(y)}{\rho\epsilon_\nu^{-\alpha}A(y) + 1 - \rho\epsilon_\nu^{-\alpha}}\,. \tag{4.14}$$

In Figure 4.2, the weight function $w(y)$ is plotted for different values of α and ρ, which in turn imply different values for $\epsilon_\nu = \log(1/\rho)^{-\eta/\alpha}$, $\eta = \alpha/2 + 0.01$. The three panels are for $\alpha = 0.5, 1$ and 1.5, while the different colors ranging from dark blue to light yellow, correspond to different values of ρ, ranging from 1% to 20%. It can be observed that, regardless of α, for larger values of ρ, $w(y)$ is consistently higher over the whole range of y. This is as expected since smaller *a priori* ρ, and therefore smaller *a priori* non-negligibility probability $\rho\epsilon_\nu^{-\alpha}$, requires more extreme observations to support the claim of signal non negligibility. On the other hand, the inverse power α mostly affects the value of the weight function $w(y)$ for small values of $|y|$: when $\alpha = 1.5$, $w(y)$ is larger relative to the cases when α is smaller.

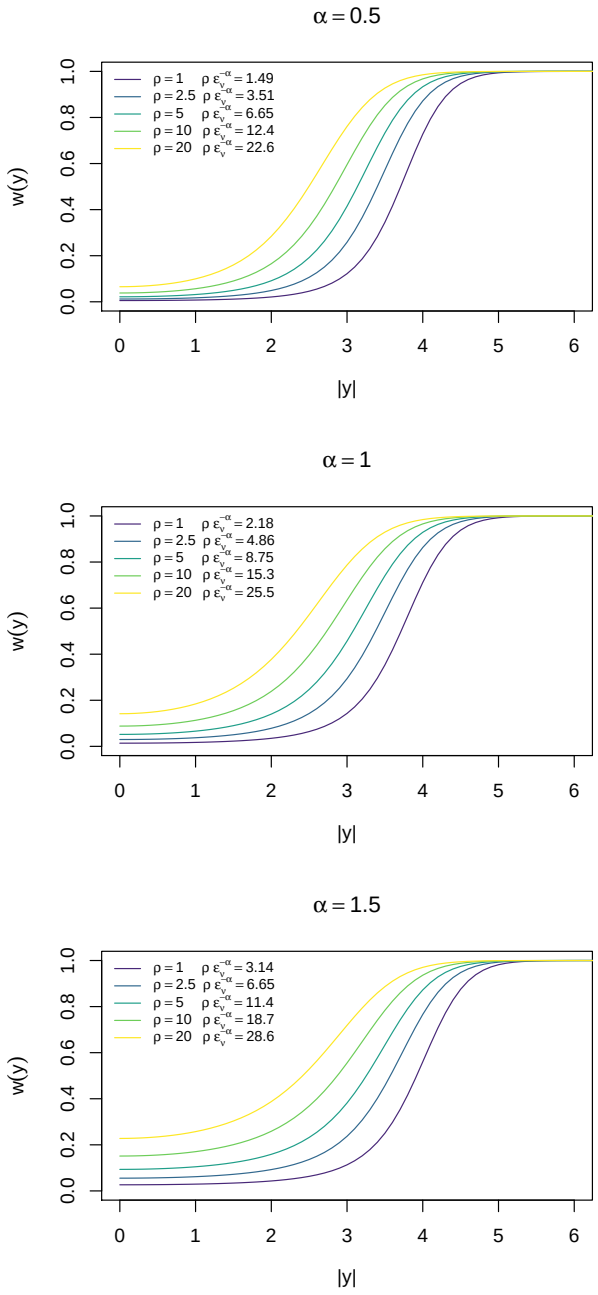


Figure 4.2: Plots of the weight function $w(y)$ as in (4.14). Different colors correspond to different values of ρ which imply different values of $\epsilon_\nu = (\log(1/\rho))^{-\eta/\alpha}$, with $\eta = \alpha/2 + 0.01$. From dark blue to light yellow, the ρ parameter is $0.01, 0.025, 0.05, 0.10, 0.20$. We also report the values of ρ and $\rho\epsilon_\nu^{-\alpha} = \rho \log(1/\rho)^{\alpha/2+0.01}$, in percentages.

Now, if we want to compute the conditional probability of non-negligibility, this can be approximated by

$$P_\nu(|\mu| > \epsilon_\nu \mid |Y| = y) = \frac{\rho\epsilon_\nu^{-\alpha} \int w_{\epsilon_\nu}(x)\, A(dx; y)}{\rho\epsilon_\nu^{-\alpha} A(y) + 1 - \rho\epsilon_\nu^{-\alpha}} + o(\rho\epsilon_\nu^{-\alpha})\,, \tag{4.15}$$

where $w_{\epsilon_\nu}(x) = 1 - e^{-x^2/2\epsilon_\nu^2}$. For small ϵ_ν, the integral $\int w_{\epsilon_\nu}(x)\, A(dx; y)$ behaves like $A(y) - y^2\epsilon_\nu^2$ (see the appendix), so that the behavior of the numerator in (4.15) is the same as

$$\rho\epsilon_\nu^{-\alpha} A(y) - \rho y^2 \epsilon_\nu^{2-\alpha}\,.$$

Now, the double limit regime under which $\rho\zeta(y) \to \lambda$, with $\lambda > 0$, requires $y^2 \sim \log(1/\rho)$. So suppose we choose the sequence of thresholds to behave like

$$\epsilon_\nu \sim \log(1/\rho)^{-\eta/\alpha}\,,$$

then

$$y^2\epsilon_\nu^2 \sim \log(1/\rho)^{1-2\eta/\alpha}\,.$$

Choosing η such that $1 - 2\eta/\alpha < 0$, i.e., $\eta > \alpha/2$, the function $\log(1/\rho)^{1-2\eta/\alpha} \to 0$ as $\rho \to 0$. In this case, we have

$$\rho y^2\, \epsilon_\nu^{2-\alpha} \sim \rho\epsilon_\nu^{-\alpha} g(\rho) = o(\rho\epsilon_\nu^{-\alpha})\,.$$

It follows that, under the double limit regime for which $\rho\zeta(y) \to \lambda$, provided the negligibility sequence converges to zero at an appropriate rate, then the conditional probability of signal non-negligibility can be approximated by

$$P_\nu(|\mu| > \epsilon_\nu \mid |Y| = y) = \frac{\rho\epsilon_\nu^{-\alpha} A(y)}{1 + \rho\epsilon_\nu^{-\alpha} A(y)} + o(\rho\epsilon_\nu^{-\alpha})\,. \tag{4.16}$$

From this, one can write the approximation for the conditional odds for $\{|\mu| > \epsilon_\nu\}$ to be

$$\frac{P_\nu(|\mu| > \epsilon_\nu \mid |Y| = y)}{P_\nu(|\mu| \leq \epsilon_\nu \mid |Y| = y)} = \rho \epsilon_\nu^{-\alpha} A(y)\,,$$

so that the Bayes Factor reduces to

$$\mathrm{BF}_{\epsilon^+}(y) = \frac{\mathrm{odds}(|\mu| > \epsilon_\nu \mid y)}{\mathrm{odds}(|\mu| > \epsilon_\nu)} = A(y)\,, \tag{4.17}$$

when the initial odds for $\{|\mu| > \epsilon_\nu\}$ are $\rho \epsilon_\nu^{-\alpha}$ to one.

4.6 Appendix

1. Here we derive the sparse approximation of order $\rho\epsilon_\nu^{-\alpha}$ to the marginal density of Y,

$$\int \cosh(yx)e^{-x^2/2}\,P_\nu(dx) =$$
$$\int \cosh(yx)e^{-x^2/2}e^{-x^2/2\epsilon_\nu^2}\,P_\nu(dx) + \int \cosh(yx)e^{-x^2/2}(1-e^{-x^2/2\epsilon_\nu^2})\,P_\nu(dx) =$$
$$(1-\rho\epsilon_\nu^{-\alpha})\int \cosh(yx)e^{-x^2/2}e^{-x^2/2\epsilon_\nu^2}\,\delta_0(dx) + \rho\epsilon_\nu^{-\alpha}\int \cosh(yx)e^{-x^2/2}\,\tilde{H}(dx) + o(\rho\epsilon_\nu^{-\alpha}) =$$
$$(1-\rho\epsilon_\nu^{-\alpha}) + \rho\int \cosh(yx)e^{-x^2/2}(1-e^{-x^2/2\epsilon_\nu^2})\,H(dx) + o(\rho\epsilon_\nu^{-\alpha}) =$$
$$(1-\rho\epsilon_\nu^{-\alpha}) + \rho\int \big((\cosh(yx)-1)e^{-x^2/2}+e^{-x^2/2}\big)(1-e^{-x^2/2\epsilon_\nu^2})\,H(dx) + o(\rho\epsilon_\nu^{-\alpha}) =$$
$$(1-\rho\epsilon_\nu^{-\alpha}) + \rho\left(\int (\cosh(yx)-1)e^{-x^2/2}(1-e^{-x^2/2\epsilon_\nu^2})\,H(dx) + \int e^{-x^2/2}(1-e^{-x^2/2\epsilon_\nu^2})\,H(dx)\right)$$
$$+ o(\rho\epsilon_\nu^{-\alpha}) =$$
$$(1-\rho\epsilon_\nu^{-\alpha}) + \rho\left(\zeta(y) - \tau^\alpha\zeta(y/\tau) + \int (e^{-x^2/2}-e^{-\tau^2x^2/2})\,H(dx)\right) + o(\rho\epsilon_\nu^{-\alpha}) =$$
$$(1-\rho\epsilon_\nu^{-\alpha}) + \rho\left(\zeta(y) - \tau^\alpha\zeta(y/\tau) - 1 + \tau^\alpha\right) + o(\rho\epsilon_\nu^{-\alpha})\,.$$

2. Here we show that $\int \phi(y)A(y)dy = 1$,

$$\begin{aligned}\int \phi(y)A(y)\,dy &= \epsilon_\nu^\alpha\left(\int \phi(y)\zeta(y)\,dy - \tau^\alpha\int \phi(y)\zeta(y/\tau)\,dy - (1-\tau^\alpha)\int \phi(y)\,dy\right)\\ &= \epsilon_\nu^\alpha\left(1-\tau^\alpha\int \phi(\tau z)\zeta(z)\tau\,dz\; - 1 + \tau^\alpha\right)\\ &= \epsilon_\nu^\alpha\tau^\alpha\left(1-\tau\int \phi(\tau z)\zeta(z)\,dz\right)\,.\end{aligned} \tag{4.18}$$

Now, the integral appearing in the last expression can be computed as

$$\begin{aligned}
\int \phi(\tau y)\zeta(y)dy &= \int \phi(\tau y) \int (\cosh(yx) - 1)e^{-x^2/2}\, H(dx)\, dy \\
&= \int\int \phi(\tau y)e^{yx}dy\, e^{-x^2/2}H(dx) - \int\int \phi(\tau y)dy\, e^{-x^2/2}\, H(dx) \\
&= \int\int \frac{1}{\sqrt{2\pi}}e^{-1/2(\tau^2y^2-yx+x^2/\tau^2)}dy\, e^{-x^2/2(1-1/\tau^2)}\, H(dx) - \int \frac{1}{\tau}\, e^{-x^2/2}\, H(dx) \\
&= \int \frac{1}{\tau}e^{-x^2/2(1-1/\tau^2)}\, H(dx) - \int \frac{1}{\tau}\, e^{-x^2/2}\, H(dx) \\
&= \frac{1}{\tau}\int (e^{-x^2/2(1-1/\tau^2)} - 1 + 1 - e^{-x^2/2})\, H(dx) \\
&= \frac{1}{\tau}\left(-(1-1/\tau^2)^{\alpha/2} + 1\right) \\
&= \frac{1}{\tau}\left(1 - \frac{(\tau^2-1)^{\alpha/2}}{\tau^\alpha}\right).
\end{aligned}$$

So, plugging this expression back in (4.18), we obtain

$$\begin{aligned}
\int \phi(y)A(y)\, dy &= \epsilon_\nu^\alpha \tau^\alpha \left(1 - \tau \int \phi(\tau z)\zeta(z)\, dz\right) \\
&= \epsilon_\nu^\alpha \tau^\alpha \left(1 - \left(1 - \frac{(\tau^2-1)^{\alpha/2}}{\tau^\alpha}\right)\right) \\
&= \epsilon_\nu^\alpha (\tau^2-1)^{\alpha/2} = 1
\end{aligned}$$

since $\tau^2 = 1 + 1/\epsilon_\nu^2$ so that $(\tau^2-1)^{\alpha/2} = (1/\epsilon_\nu^2)^{\alpha/2} = \epsilon_\nu^{-\alpha}$.

3. Here we report the few passages for establishing the behavior of

$$\int w_{\epsilon_\nu}(x)\, A(dx; y) = \int (1 - e^{-x^2/2\epsilon_\nu^2})\, A(dx; y)\,,$$

appearing in the conditional probability of non-negligibility. So,

$$\int (1 - e^{-x^2/2\epsilon_\nu^2})\, A(dx; y) = A(y) - \epsilon_\nu^\alpha \int e^{-x^2/2\epsilon_\nu^2}\, \cosh(yx)e^{-x^2/2}(1 - e^{-x^2/2\epsilon_\nu^2})\, H(dx)\,,$$

and because ϵ_ν is small, the integral in the RHS behaves like

$$\int_{|x|\le\epsilon_\nu} \frac{y^2x^2}{2}\,\frac{x^2}{2\epsilon_\nu^2}\,H(dx) \;\sim\; \frac{y^2}{2}\,\frac{1}{2\epsilon_\nu^2}\int_{|x|\le\epsilon_\nu} |x|^{4-\alpha-1}\,dx \;\sim\; y^2\,\epsilon_\nu^{2-\alpha}\,.$$

Therefore, putting things together,

$$\int (1-e^{-x^2/2\epsilon_\nu^2})\,A(dx;y) \sim A(y) - y^2\,\epsilon_\nu^2\,.$$

Chapter 5

Multiple testing for negligible signals

5.1 Introduction

In the context of many observations generated from a sparse signal plus noise model, identifying simultaneously which signals are negligible and which, on the contrary, can be considered as active, is often an object of primary interest. In this section, we make use of the negligibility notion we introduced in Chapter 4, to outline a multiple testing procedure to simultaneously test a collection of null hypotheses, where each of these null hypotheses states the negligibility of a single signal.

We present this testing procedure to show how the negligibility notion can be exploited in the sparsity theory developed by McC&P, but the procedure is not thought to be innovative in its formulation. As a matter of fact, it mimics the testing procedure first proposed by Benjamini & Hochberg (1995) [9] to control the false discovery rate (FDR). The BH procedure, in turn, despite its initial formulation in a frequentist framework, has been shown, by many authors, to have close connections with empirical Bayes methods, which control the Bayesian Fdr. See for instance Efron and Tibshirani (2001) [30].

The algorithm we develop within the negligibility-sparsity theory might be labelled as an empirical Bayes procedure. Yet, despite the general structure resemblance, there are two main differences between our method and other Bayesian approaches proposed for the two-groups model. The first most obvious discrepancy concerns the null hypothesis being tested, which, in our framework is on the signal negligibility rather than on the signal being absolutely zero. The second difference follows from the first one to the extent that the mixture, with a Dirac delta measure at zero, for the signal distribution, is an asymptotic approximation, driven by the sparsity limit, and only true in terms of expectations of bounded and continuous functions.

5.2 Multiple testing procedure

Let the observations be

$$Y_i = \mu_i + \eta_i \qquad i = 1, \ldots, n\,. \tag{5.1}$$

The signals μ_i are independent and identically distributed according to P_ν, which is a scale sparse distribution with rate ρ and inverse-power exceedance measure H, with exponent α. The errors η_i, on the other hand, are independent standard Gaussian, and each μ_i is independent of each η_i. Then, using the $o(\rho\epsilon_\nu^{-\alpha})$ integral approximation for the signal distribution P_ν,

$$(1 - \rho\epsilon_\nu^{-\alpha})\delta_0(dx) + \rho\epsilon_\nu^{-\alpha}\tilde{H}(dx)\,,$$

the sparse approximation of order $o(\rho\epsilon_\nu^{-\alpha})$, to the marginal density of each Y_i, can be written as

$$m_\nu(y) = (1 - \rho\epsilon_\nu^{-\alpha})\phi(y) + \rho\epsilon_\nu^{-\alpha}\phi(y)A(y) + o(\rho\epsilon_\nu^{-\alpha})\,,$$

where

$$A(y) = \int \cosh(yx)e^{-x^2/2}\tilde{H}(dx)\,.$$

Let ϵ_ν be a negligibility sequence for the sparse measure P_ν, and consider the collection of null hypotheses

$$H_i : |\mu_i| \leq \epsilon_\nu \qquad i = 1, \ldots, n\,,$$

which we would like to test based on the corresponding observed values $y_1, \ldots, y_n$. To this end, we propose the following testing procedure.

1. Choose a desired level $q \in (0, 1]$.

2. Denote by $y_{(i)}$ the i^{th} largest observation in absolute value, i.e., $y_{(i)} = y_{\sigma(i)}$ where $\sigma \in \mathcal{S}^n$ is the permutation $\sigma : [n] \to [n]$ which ranks the collection $|y_1|, |y_2|, \ldots, |y_n|$ in decreasing order, so that $y_{(1)} \geq y_{(2)} \geq \cdots \geq y_{(n)} \geq 0$. Denote by $\mu_{(i)}$ and $H_{(i)} : |\mu_{(i)}| \leq \epsilon$, the signal and the null hypothesis corresponding to $y_{(i)}$.

3. For each $i = 1, \ldots, n$, compute

$$f_{(i)} = (1 - \rho\epsilon_\nu^{-\alpha})\Phi(y_{(i)}+)\,,$$

where $\Phi(z+) = 2\,\mathbb{P}(Z > z)$ for $Z \sim N(0, 1)$. This is the sparse approximation of order $o(\rho\epsilon_\nu^{-\alpha})$ to the joint probability that the signal is negligible and the corresponding observation Y is larger than $y_{(i)}$, in absolute value. In fact,

$$\begin{aligned}
\mathbb{P}(|\mu| \leq \epsilon_\nu, |Y| \geq y) &= \mathbb{P}(|Y| \geq y \mid |\mu| \leq \epsilon_\nu)\mathbb{P}(|\mu| \leq \epsilon_\nu) \\
&= 2\int_y^\infty \mathbb{P}(\mu + \eta \in dz \mid |\mu| \leq \epsilon)\, dz \cdot \mathbb{P}(|\mu| \leq \epsilon_\nu) \\
&= 2\int_y^\infty \int_{\mathbb{R}} \mathbb{P}(\eta \in d(z-u))\mathbb{P}(\mu \in du \mid |\mu| \leq \epsilon_\nu)\, dz \cdot \mathbb{P}(|\mu| \leq \epsilon_\nu) \\
&= 2\int_y^\infty \int_{\mathbb{R}} \phi(z-u)\delta_0(du)\, dz \cdot (1 - \rho\epsilon_\nu^{-\alpha}) + o(\rho\epsilon_\nu^{-\alpha}) \\
&= \Phi(y+)(1 - \rho\epsilon_\nu^{-\alpha}) + o(\rho\epsilon_\nu^{-\alpha})\,.
\end{aligned}$$

Observe that, by monotonicity of $\Phi(z+)$, the increasing ranking

$$f_{(1)} \leq f_{(2)} \leq \cdots \leq f_{(n)}\,,$$

corresponds to the decreasing ranking of the observed absolute values

$$y_{(1)} \geq y_{(2)} \geq \cdots \geq y_{(n)}.$$

4. Find
$$y_n^q = \inf\{y > 0 : \frac{(1-\rho\epsilon_\nu^{-\alpha})\Phi(y+)}{\#\{i \in [n] : y_{(i)} \geq y\}/n} \leq q\}$$
and the corresponding
$$f_n^q = (1-\rho\epsilon_\nu^{-\alpha})\Phi(y_n^q+)\,.$$

5. Reject all null hypothesis $H_{(i)}$ corresponding to $f_{(i)} \leq f_n^q$.

6. If
$$\{y > 0 : \frac{(1-\rho\epsilon_\nu^{-\alpha})\Phi(y+)}{\#\{i \in [n] : y_{(i)} \geq y\}/n} \leq q\} = \emptyset\,,$$
then no hypothesis is rejected.

Now, for any testing procedure which identifies a rejection region Γ on some domain, one can define $R_n(\Gamma)$ and $V_n(\Gamma)$ to be the total number of rejections and the number of wrong rejections, respectively, out of n hypothesis. Then, the false discovery proportion (FDP) is defined as a function of the rejection region,

$$\mathrm{FDP}(\Gamma) = \frac{V_n(\Gamma)}{R_n(\Gamma) \vee 1}\,, \tag{5.2}$$

as the proportion of wrong rejections among all rejections.

From this point of view, it is easy to see that, analogously to the Benjamini-Hochberg procedure which is discussed in more detail in the next section, for given values of ρ, ϵ_ν and α, the procedure outlined above proceeds in the following way. For each data point y_k, it estimates the false discovery proportion corresponding to the rejection region defined by that point, by computing

$$\frac{(1-\rho\epsilon_\nu^{-\alpha})\Phi(y_k+)}{\#\{i\in[n]:y_{(i)}\geq y_k\}/n}\,;$$

based on these estimates, it chooses the rejection region $\Gamma=\{y:|y|>y_n^q\}$, in such a way that the FDP is controlled at some predetermined level q.

In practice, however, we do not know the sparsity parameters ρ, ϵ_ν and α, so we estimate them by maximum likelihood. As a matter of fact, because we need the negligibility sequence ϵ_ν to converge to zero at a slower rate than the sparsity rate ρ, we can just set ϵ_ν to be some function of ρ and α, such as $\epsilon_\nu=\log(1/\rho)^{-1/2\alpha}$, and only estimate ρ and α. Yet, as explained in Chapter 4 Section 4.4.1, in applied work, it is important that the estimates for the sparsity parameters are not dependent of the scale on which the observations are made. So we assume that the independent observations are

$$\sigma Y_i=\mu_i+\eta_i\qquad i=1,\ldots,n\,,$$

where μ_i and η_i are the same as in (5.1), while σ is an unknown scale parameter. From this formulation, one can write a sparse approximation to the log likelihood function for the triplet (ρ,α,σ^2), and maximize it to obtain the maximum likelihood estimates for these parameters.

In the next section, we investigate the multiple connections of the sparsity-negligibility procedure outlined above, with only a few of the many existing approaches which have been developed in this context.

5.3 Connections with some literature

5.3.1 Benjamini-Hochberg's FDR and Bayesian Fdr

The concept of false discovery rate (FDR) was introduced by Benjamini and Hochberg's seminal paper [9] (B&H from now on) as a new approach to simultaneous testing. The introduction of FDR control was proposed as an extension of the frequentist hypothesis testing framework to the setting where a large number of independent hypothesis are to be tested simultaneously.

For a given multiple testing procedure, the false discovery rate (FDR) is defined as the expected proportion of the wrongly rejections decided by the procedure. In other words, the FDR is the expected value of the false discovery proportion defined in (5.2),

$$\mathrm{FDR}(\Gamma) = \mathbb{E}\left(\frac{V_n(\Gamma)}{R_n(\Gamma) \vee 1}\right) . \tag{5.3}$$

In the Benjamini and Hochberg's frequentist approach, the signals μ_i are unknown but fixed parameters, so the only randomness in Y_i is given by the noise component η_i. Therefore the expectation in (5.3) is taken over the noise distribution. Taking the cue from Simes (1986) [65], to test a collection of sharp null hypotheses $H_i : \mu_i = 0$, for any given proportion of true nulls π_0, B&H proposed the following algorithm:

1. Choose a desired level $q \in (0, 1]$.

2. Compute the p-values $p_1, \ldots, p_n$ and sort them in increasing order

$$p_{(1)} \leq \cdots \leq p_{(n)} .$$

3. Find

$$i_q = \arg\max_{i=1,\ldots,n} \{p_{(i)} \leq \frac{i/n}{\pi_0} q\} \,.$$

4. Reject all null hypotheses $H_{(i)}$ having $p_{(i)} \leq p_{(i_q)}$.

5. If

$$\{i : p_{(i)} \leq \frac{i/n}{\pi_0} q\} = \emptyset \,,$$

then no hypothesis is rejected.

B&H prove that, following this procedure, one is guaranteed to control FDR at any required level q. In that paper, the proportion of true null hypotheses π_0 is fixed to one, corresponding to the most conservative approach. However, in later works such as Benjamini and Yekutieli (2001) [10], it is suggested to estimate π_0 from the data in order to reduce the conservative bias.

After B&H, a large number of papers, such as Efron et al. (2001A [30], 2001B [31]), Efron and Tibshirani (2002) [29], Genovese and Wasserman (2002) [39], Storey (2003) [68], among many others, showed a close relationship between empirical Bayes methods and FDR theory. The common starting point is to consider for each hypothesis, some univariate summary statistic Z_i. In empirical Bayes approaches, the assumed model is the so called two-groups model, for which $Z_1, \ldots, Z_n$ are independent and identically distributed according to the mixture density f

$$f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i) \,.$$

This mixture model can be interpreted in terms of a hierarchical model where, for each $i = 1, \ldots, n$, the latent variable

$$r_i \sim Ber(\pi_0)$$

determines which class, null or not null, the signal μ_i belongs to, so that

$$\mu_i \mid r_i = 0 \sim \delta_0\,, \quad \mu_i \mid r_i = 1 \sim G\,, \tag{5.4}$$

where G is some non zero distribution. Then $Z_i = \mu_i + \eta_i$ is such that

$$Z_i \mid r_i = 0 \sim F_0\,, \quad Z_i \mid r_i = 1 \sim F_1\,,$$

where F_0 is the distribution function of the η_i's while F_1 is the convolution of G with F_0. Denoting by F the cumulative distribution function (c.d.f.) of f, Efron and Tibshirani (2002) [29] (E&T henceforth) defined the Bayesian Fdr to be

$$\mathrm{Fdr}(z) = \frac{\pi_0 F_0(z)}{F(z)}\,, \tag{5.5}$$

for rejection regions $\Gamma = \{z_i : z_i \leq z\}$. The non-parametric estimate of $\mathrm{Fdr}(z)$ is

$$\overline{\mathrm{Fdr}}(z) = \frac{\pi_0 F_0(z)}{\#\{z_i : z_i \leq z\}/n}\,,$$

where $\#\{z_i : z_i \leq z\}/n$ is the empirical version of $F(z)$. The *Equivalence Theorem* in E&T states that, for known π_0 and F_0, the rule in step 4. of the B&H procedure is equivalent to rejecting all those hypotheses with $z_i \leq z_q$ where

$$z_q = \max_z \{\overline{\mathrm{Fdr}}(z) \leq q\}\,.$$

In fact, if we sort $z_{(1)} \leq \cdots \leq z_{(n)}$, then $\#\{z_i : z_i \leq z_{(i)}\}/n = i/n$ and $F_0(z_{(i)}) = p_{(i)}$. So rejecting all hypotheses with $z_i \leq z_q$ is equivalent to rejecting all hypothesis with

$$\frac{\pi_0\, p_{(i)}}{i/n} \leq q\,.$$

In the testing procedure described in Section 5.2, the signals are considered to be random with a sparse distribution P_ν. Keeping in mind that the hypotheses being tested there are on the negligibility of the μ_i's rather than on the sharp events $\{\mu_i = 0\}_i$, *mutatis mutandis*, it may come natural to identify $(1 - \rho\epsilon_\nu^{-\alpha})$ with π_0 and $\Phi(y_{(i)}+)$ with $F_0(z_{(i)})$. However, besides the shift from testing the hypothesis $\{\mu_i = 0\}$ to testing $\{|\mu_i| \leq \epsilon_\nu\}$, there is another main conceptual difference, which is that in Section 5.2, we use sparse approximations rather than exact distributions: $(1 - \rho\epsilon_\nu^{-\alpha})$ is the sparse asymptotic approximation to $\mathbb{P}(|\mu_i| \leq \epsilon_\nu)$ and Φ is the sparse approximation to the conditional distribution of $Y_i \mid |\mu_i| \leq \epsilon_\nu$.

Since we are going to estimate the sparsity parameters (ρ, α) using maximum likelihood, the procedure in Section 5.2 can be seen as an empirical Bayes approach to multiple testing. Yet, the *Equivalence Theorem* makes the connection between the Bayesian and the frequentist approach. So, supposing that the true null proportion is $\pi_0 = 1 - \rho\epsilon_\nu^{-\alpha}$ and the distribution of Y, under the null hypothesis, is Φ, then, if the rejection region $\Gamma = \{y_{(i)} \geq y\}$ is chosen as large as possible subject to the constraint that

$$\frac{(1 - \rho\epsilon_\nu^{-\alpha})\Phi(y+)}{\#\{i : y_{(i)} \leq y\}/n} \leq q\,,$$

then the expected proportion of wrong rejections, from a frequentist point of view, is also less than q.

5.3.2 Connections with Storey's q-value and Stephens (2017)

Since, when $R_n(\Gamma) = 0$, the FDR is also equal to zero, then not rejecting any hypothesis would always guarantee FDR control at any level. For this reason, Storey (2003) [68] argues

for controlling another quantity,

$$\text{pFDR}(\Gamma) = \mathbb{E}\left(\frac{V_n(\Gamma)}{R_n(\Gamma)} \,\middle|\, R_n(\Gamma) > 0\right), \tag{5.6}$$

which is called positive false discovery rate, since it is conditioning on having a positive number of rejections. Here Γ is the rejection region on the sample space of the Z statistic. To further make the frequentist-Bayesian connection clear, Storey (2003) [68] shows that, under the two-groups model in (5.4) and the Z_i's independence assumption,

$$\text{pFDR}(\Gamma) = \mathbb{P}(\mu_i = 0 \mid Z_i \in \Gamma)\,.$$

We report here the short proof. Write $\text{pFDR}(\Gamma)$ as

$$\mathbb{E}\left(\frac{V_n(\Gamma)}{R_n(\Gamma)}\Big|R_n(\Gamma) > 0\right) = \sum_{k=1}^{n} \mathbb{E}\left(\frac{V_n(\Gamma)}{k}\Big|R_n(\Gamma) = k\right) \mathbb{P}(R_n(\Gamma) = k \mid R_n(\Gamma) > 0)\,.$$

Now,

$$\begin{aligned}
\mathbb{E}\left(V_n(\Gamma)\big|R_n(\Gamma) = k\right) &= \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{z_{(i)}\in\Gamma,\mu_i=0}\big|\, Z_{(1)},\ldots,Z_{(k)} \in \Gamma\,; Z_{(k+1)},\ldots,Z_{(n)} \notin \Gamma\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{k} \mathbf{1}_{\mu_i=0}\big|\, Z_{(1)},\ldots,Z_{(k)} \in \Gamma\,; Z_{(k+1)},\ldots,Z_{(n)} \notin \Gamma\right) \\
&= \sum_{i=1}^{k} \mathbb{E}\left(\mathbf{1}_{\mu_i=0}\big|\, Z_{(i)} \in \Gamma\right) \\
&= k\ \mathbb{P}\left(\mu_i = 0\big|\, Z_i \in \Gamma\right).
\end{aligned}$$

So then

$$\begin{aligned}
\text{pFDR}(\Gamma) &= \sum_{k=1}^{n} \mathbb{E}\left(\frac{V_n(\Gamma)}{R_n(\Gamma)}\Big|R_n(\Gamma) = k\right) \mathbb{P}(R_n(\Gamma) = k \mid R_n(\Gamma) > 0) \\
&= \sum_{k=1}^{n} \mathbb{P}\left(\mu_i = 0\big|\, Z_i \in \Gamma\right) \mathbb{P}(R_n(\Gamma) = k \mid R_n(\Gamma) > 0) \\
&= \mathbb{P}\left(\mu_i = 0\big|\, Z_i \in \Gamma\right),
\end{aligned}$$

as desired.

Notice that, under (5.4), since $\mathbb{P}(\mu_i = 0,\, Z_i \in \Gamma) = \pi_0 F_0(\Gamma)$, this proves that the pFDR(Γ) defined in (5.6), with $\Gamma = \{z_i : z_i \leq z\}$, is actually equivalent to the Bayesian FDR defined in (5.5).

Together with pFDR, Storey (2003) [68] also introduced the q-value, defined as

$$q\text{-value}(z) = \inf_{\{\Gamma_\lambda :\, z \in \Gamma_\lambda\}} \text{pFDR}(\Gamma_\lambda)\,.$$

Yet, in light of the equivalence proved above, an alternative definition is

$$q\text{-value}(z) = \inf_{\{\Gamma_\lambda :\, z \in \Gamma_\lambda\}} \mathbb{P}(\mu_i = 0 \mid Z_i \in \Gamma_\lambda)\,,$$

whose empirical counterpart, for rejection regions of the type $\Gamma = \{z_i : z_i \leq z\}$, is indeed the non parametric estimate of the Bayesian Fdr

$$\hat{q}\text{-value}(z) = \frac{\pi_0 F_0(z)}{\#\{z_i : z_i \leq z\}/n} = \overline{\text{Fdr}}(z)\,.$$

Once again, with the necessary caution highlighted at the end of the last section, if we want to match this setting with the sparsity-negligibility framework of Section 5.2, then $Z_i = |Y_i|$ and

$$\hat{q}\text{-value}(y) = \frac{(1 - \rho\epsilon_\nu^{-\alpha})\Phi(y+)}{\#\{i : y_{(i)} \leq y\}/n}\,.$$

It is now evident that the multiple testing procedure of Section 5.2 finds the data-dependent threshold

$$y_n^q = \inf\{y > 0 : \hat{q}\text{-value}(y) \leq q\}\,,$$

and so any rejected null hypothesis $\{|\mu_i| \leq \epsilon_\nu\}$ has corresponding $\hat{q}$-value$(y_{(i)}) \leq q$. Yet,

while the theoretical version of the tail conditional probability $\mathbb{P}(|\mu_i| \leq \epsilon_\nu \mid |Y_i| \geq y)$ is necessarily a decreasing function of y, the empirical version may not be so. Thus, it is possible that some of the not rejected null hypotheses have a corresponding $\hat{q}$-value$(y_{(i)})$ which is still less or equal than the level q.

Still within the empirical Bayes paradigm, but coming from a somewhat different perspective, perhaps more focussed on estimation rather than testing, Stephens (2017) [66] puts forward three main ideas in approaching large-scale studies. The first one is the so called *unimodal assumption* (UA) for the distribution of all the signals. So instead of pre-specifying a given distribution for the signals, Stephens (2017) [66] only assumes the unimodality of such unknown distribution g, symmetric around the origin, and estimates it in a non-parametric fashion. This is done by considering a large fixed grid of scale parameters $\{\sigma_0, \sigma_1, \ldots, \sigma_K\}$, and construct a scale mixture of zero-mean Gaussian distributions, whose variances correspond to the grid points:

$$g(dx; \pi) = \sum_{k=0}^{K} \pi_k N(dx; 0, \sigma_k^2)\,,$$

where $N(dx; 0, \sigma_k^2)$ denotes the Gaussian distribution with mean zero and variance σ_k^2, and $\sigma_0 = 0$ allows one to include the Dirac delta measure at zero, $\delta_0(dx)$. The mixture weights $\pi_0, \pi_1, \ldots, \pi_K$ are estimated by maximum likelihood. Even if the actual implementation for estimating the signal distribution is far from the sparsity-negligibility framework, both approaches share the same instance of comprising more than just one specific model, since also the sparse distribution is only identified through its sparsity pair (ρ, H). However, we have seen that two different scale Gaussian mixtures can have the same sparsity pair; thus, this choice of g does not characterize the tail behavior of the signal distribution as it does the exceedance measure.

The second idea is to retain two statistics, both the signal sizes and their standard errors,

rather than converting them into z-values or p-values. This approach has the potential of being more informative, as it allows one to account for differences in measurements precision. In Stephens (2017) [66], the standard errors are used as empirical Bayes estimates for the unknown error variances $\eta_i \sim N(0, \tau_i^2)$. In the sparsity-negligibility framework presented here, we assume $\eta_i \sim N(0, 1)$ for all $i = 1, \ldots, n$, and estimate by maximum likelihood a common scale parameter for all observations. Yet, if strong reasons are given to believe in heteroskedasticity, then one could follow the t-statistic generalization presented in Chapter 1.

Lastly, Stephens (2017) [66] introduces the *local false sign rate* (lfsr), which is analogous to Efron's local FDR we discuss in the next section, but the emphasis is put on the sign of the signal, rather than on its being non-zero. Even in this change of perspective, we can trace a resemblance with the negligibility-sparsity approach, whose standpoint is indeed that of considering the negligibility hypothesis rather than the sharp zero hypothesis.

5.3.3 Efron's local fdr and empirical null

In Efron et al. (2001A [30], 2001B [31]), the local FDR at point z, is defined as

$$\mathrm{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)}\,,$$

which is the conditional probability of the signal being null given that the observed statistic is exactly equal to z. More generally, for any subset $\mathcal{Z}$ of the sample space of the Z statistic, one can define

$$\mathrm{Fdr}(\mathcal{Z}) = \frac{\pi_0 \mathbb{P}_{f_0}(\mathcal{Z})}{\mathbb{P}_f(\mathcal{Z})}\,.$$

According to E&T, the advantage of $\mathrm{fdr}(z)$ is its 'specificity' since it provides a measure of belief on the i^{th} hypothesis that only depends on the exact value of Z_i. However, there is one difficulty arising from using $\mathrm{fdr}(z)$ rather than $\mathrm{Fdr}(z)$, which is the estimation of $f(z)$. A non parametric option implemented in E&T, described in more detail by Efron (2007) [27],

is to estimate f using a Poisson spline regression. So, after binning the z-values in, say, K bins, one obtains a maximum likelihood fit to the histogram of these bin counts, by fitting a J-parameter exponential family,

$$f_\beta(z) = \exp\left\{\sum_{j=0}^{J} \beta_j z^j\right\}.$$

Quite naturally, one can see that the expression

$$\frac{(1-\rho\epsilon_\nu^{-\alpha})\phi(y)}{m_\nu(y)} = \frac{1-\rho\epsilon_\nu^{-\alpha}}{1-\rho\epsilon_\nu^{-\alpha}+\rho\epsilon_\nu^{-\alpha}A(y)}$$

is the sparsity-negligibility analog to the local fdr, being the sparse approximation of order $\rho\epsilon_\nu^{-\alpha}$ to $\mathbb{P}(|\mu_i| \leq \epsilon_\nu)\mathbb{P}(y \mid |\mu_i| \leq \epsilon_\nu)/\mathbb{P}(y)$.

Besides estimating the marginal mixture f, Efron (2004) [26] proposes to also estimate f_0, the density of Z under the null hypothesis $H_i : \mu_i = 0$. His idea is to fix the kernel of f_0 to be Gaussian, and then, as explained below, use various techniques, such as central matching, to estimate the mean and variance parameters, (δ_0, σ_0^2). In this way, one obtains what Efron calls the *empirical null density*. The motivations for doing so are multiple. One of these is that correlation among the z-values and unobserved covariates arising in observational studies, might make $N(0,1)$ a non appropriate choice for describing the empirical distribution of the observed $(z_1, \ldots, z_n)$, even when marginally $Z_i \sim N(0,1)$ for all i. See Efron (2012) [28], p. 105, for a complete and more detailed list of reasons. For a different treatment of the problem of having correlated z-values due to correlation among the noise components, see Stephens and Sun (2018) [70].

5.3.4 Estimation of the null atom

Regarding the proportion of true null hypothesis $H_i : \mu_i = 0$, Efron et al. (2001A) [30], and Storey (2002) [67] suggest obtaining an estimate of π_0 via the following reasoning. Let $\mathcal{A}_0$ be a region around the origin. The *zero assumption* states that all the non-null cases give z-values outside $\mathcal{A}_0$; in other words, the distribution of Z under the alternative gives zero mass to the region $\mathcal{A}_0$. Then the expected number of z-values in $\mathcal{A}_0$, N_0, is n times $\pi_0 F_0(\mathcal{A}_0)$, where $F_0(\mathcal{A}_0)$ is the probability of $\mathcal{A}_0$ under the null distribution F_0. This suggests estimating π_0 with

$$\hat{\pi}_0 = \frac{n_0}{n \cdot F_0(\mathcal{A}_0)} ,$$

where n_0 is the actual number of z-values observed in $\mathcal{A}_0$.

Efron (2007 [27], 2012 [28]) on the other hand, proposes to estimate π_0 together with the two parameters of the empirical null $N(\delta_0, \sigma_0^2)$, either by central matching or by maximum likelihood. The former method assumes the *zero assumption*, so that around the origin $\log(f(z)) = \log(\pi_0 f_0(z))$, and it also assumes that $\log(f(z))$ is quadratic near $z = 0$. So, first it estimates the parameters of the parabola

$$\log\left(f(z)\right) = \beta_0 + \beta_1 z + \beta_2 z^2 ,$$

from the histogram counts of the z_i's around zero. Then it matches $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ with the corresponding terms in the expression for $\log\left(\pi_0 f_0(z)\right)$

$$\log \pi_0 - \tfrac{1}{2}\log(2\pi\sigma_0^2) - \tfrac{1}{2}\delta_0^2/\sigma_0^2 + \tfrac{1}{\sigma^2}\delta_0 z - \tfrac{1}{2\sigma^2} z^2 .$$

to obtain the estimate for $(\pi_0, \delta_0, \sigma_0^2)$. The maximum likelihood (ML) estimate method, on the other hand, also starts with the *zero assumption*, $f_1(z) = 0$ for $z \in \mathcal{A}_0$, and estimates

$(\pi_0, \delta_0, \sigma_0^2)$ from the likelihood of the z-values observed in $\mathcal{A}_0$

$$\binom{n}{n_0}\theta^{n_0}(1-\theta)^{n-n_0}\prod_{z_i\in\mathcal{A}_0}\frac{\phi_{\delta_0,\sigma_0^2}(z_i)}{\Phi_{\delta_0,\sigma_0^2}(\mathcal{A}_0)}\,.$$

In this expression, $\theta = \mathbb{P}(Z \in \mathcal{A}_0) = \pi_0\Phi_{\delta_0,\sigma_0^2}(\mathcal{A}_0)$, whereas

$$\phi_{\delta_0,\sigma_0^2}(z_i)/\Phi_{\delta_0,\sigma_0^2}(\mathcal{A}_0)$$

is the probability density of Z at z_i given that $Z \in \mathcal{A}_0$, assuming $Z \sim N(\delta_0, \sigma_0^2)$.

Notice that the *zero assumption*, meaning $f_1(z) = 0$ for $z \in \mathcal{A}_0$, is not satisfied by the hierarchical model in (5.4) if F_0 is assumed to be a Gaussian distribution. In fact, if F_1 is the convolution of F_0 with a non-zero distribution G, then clearly $F_1(\mathcal{A}_0) = G * \Phi_{\delta_0,\sigma_0^2}(\mathcal{A}_0) > 0$ and not zero. The violation of the *zero assumption* in a setting like this, introduces some bias in both ML and central matching estimates. Still Efron (2012) [28] claims that the bias in those obtained using central matching is not very large as long as $1 - \pi_0$ is small.

5.3.5 Comparison for Leukemia data

In this section, we would like to give an idea of how the methods described in previous sections compare to each other. To this end, we analyze a dataset coming from a leukemia study (Golub et al., 1999 [41]), freely available on Efron's website. The Leukemia dataset is an example of microarrays data, which is one kind of data that first posed the statistical problem of testing thousands of hypothesis simultaneously (see for instance Efron et al., 2001B [31], E&T). As described in Efron (2012) [28], for this dataset, high density oligonucleotide microarrays provided expression levels on 7128 genes for 72 patients, 45 with acute lymphoblastic leukemia (ALL) and 27 with acute myeloid leukemia (AML). The scientific

interest lies on establishing whether there are some genes which show a different expression level between the AML and ALL groups of patients. To this end, the raw expression levels on each microarray, X_{ij} for gene $i = 1, \ldots, 7128$ on patient array $j = 1, \ldots, 72$, were first transformed to a normal score value x_{ij} via an empirical quantile-matching transform. These kinds of standardization steps are very common in microarrays data for removing response disparities among the microarrays as well as some wild outlying values. Following this, two-sample t-statistics t_i were computed, gene by gene, for comparing the sample mean from the AML group with the sample mean from the ALL group,

$$t_i = \frac{\bar{x}_i^{\mathrm{ALL}} - \bar{x}_i^{\mathrm{AML}}}{s_i} \,.$$

Here s_i is the estimate of the standard error of the numerator of t_i and the pooled sample variance is used to estimate the variance, which is assumed to be the same for both groups. These t-statistics were then transformed into z-values so that the actual data consists of

$$z_i = \Phi^{-1}(F_{t_{70}}(t_i)) \quad i = 1, \ldots, 7128 \,.$$

Within the sparsity-negligibility setting, we assume that, for each $i = 1, \ldots, 7128$,

$$Z_i = \sigma(\mu_i + \eta_i)$$

where $\mu_i \overset{i.i.d.}{\sim} P_\nu$ sparse with first-order pair (ρ, H_α) and $\eta_i \overset{i.i.d.}{\sim} N(0,1)$, with μ_i and η_i independent of each other. The sparse approximation of order $o(\rho\epsilon_\nu^{-\alpha})$ to the marginal of Z_i is

$$(1 - \rho\epsilon_\nu^{-\alpha})\, \tfrac{1}{\sigma}\phi(z/\sigma) + \rho\epsilon_\nu^{-\alpha}\, \tfrac{1}{\sigma}\psi_A(z/\sigma) \,, \tag{5.7}$$

where $\psi_A(z) = \phi(z)A(z)$, while the sparse approximation of order $o(\rho)$ is

$$(1 - \rho)\, \tfrac{1}{\sigma}\phi(z/\sigma) + \rho\, \tfrac{1}{\sigma}\psi(z/\sigma) \,, \tag{5.8}$$

where $\psi(z) = \phi(z)\zeta(z)$.

On the other hand, the empirical Bayes approach assumes that, for each $i = 1, \ldots, 7128$, the marginal density of Z_i is

$$\pi_0 f_0 + (1 - \pi_0) f_1(z)\,,$$

where, following Efron (2004 [26], 2007 [27]), f_0 is the density function of the empirical null distribution $N(\delta_0, \sigma_0^2)$, while f_1 is assumed to satisfy the *zero assumption*, i.e., $F_1(\mathcal{A}_0) = 0$ for some region $\mathcal{A}_0$ around the origin.

It is interesting to notice that the tail-inflation component in the $o(\rho)$-sparse approximation to the marginal density of Z_i in (5.8), does satisfy

$$\psi(0) = \phi(0)\zeta(0) = 0\,, \tag{5.9}$$

whereas, as mentioned before, if F_1 is the convolution of F_0 with some distribution G, then

$$f_1(0) = \int f_0(-x)\, G(dx) = \int \phi_{\delta_0, \sigma_0^2}(x)\, G(dx) > 0$$

whenever $G(dx)$ is not identically zero. From (5.9), one can also see that, if the exceedance measure H was known, so would be ζ, and one could estimate the sparsity rate ρ exploiting the fact that $1 - \rho = m_\nu(0)/\phi(0)$. All the same, one should still bear in mind that, even if the signal distribution was in fact an atom-and-slab mixture with mixing parameter π_0, as in the two-groups model in (5.4), ρ would not be the proportion of non-null hypothesis, since $\rho = \pi_0 \int (1 - e^{-x^2/2})\, G(dx) < \pi_0$.

As a way of comparing the two approaches, we can derive what negligibility threshold is required in the sparsity-negligibility procedure in order to identify the same rejection region

as that obtained using Efron's empirical null estimation. In other words, we derive what negligibility hypothesis $|\mu| \leq \epsilon_\nu$, the 'sharp' null hypothesis $\mu = 0$ corresponds to. So, given the estimates for the two-groups model $(\hat{\pi}_0, \hat{\delta}_0, \hat{\sigma}_0^2)$, we first estimate the sparsity parameters (ρ, α, σ^2), and then find the threshold ϵ_ν such that

$$\inf\left\{z > 0 : \frac{(1 - \hat{\rho}\epsilon_\nu^{-\hat{\alpha}})\Phi_{\hat{\sigma}}(z+)}{\#\{i \in [n] : |z_i| \geq z\}/n} \leq q\right\} = z^E\,,$$

where the value

$$z^E = \inf\left\{y > 0 : \frac{\hat{\pi}_0 \hat{F}_0(z+)}{\#\{i \in [n] : |z_i| \geq z\}/n} \leq q\ \right\}$$

is the data dependent threshold for controlling the Bayesian Fdr at level q using the estimated empirical null $\hat{F}_0 = N(\hat{\delta}_0, \hat{\sigma}_0^2)$. So, given z^E, we solve for ϵ_ν

$$\frac{(1 - \hat{\rho}\epsilon_\nu^{-\hat{\alpha}})\Phi_{\hat{\sigma}}(z^E+)}{\#\{i \in [n] : |z_i| \geq z^E\}/n} = q\,,$$

and obtain the 'matching' threshold

$$\epsilon_\nu^E = \left(\hat{\rho}^{-1}\left(1 - q \cdot \#\{i \in [n] : |z_i| \geq z^E\}/n \cdot \frac{1}{\Phi_{\hat{\sigma}}(z^E+)}\right)\right)^{-1/\hat{\alpha}}\,. \tag{5.10}$$

To estimate the sparsity parameters, we maximize the log likelihood from the whole sample

$$l(\rho, \alpha, \sigma^2; z_1, \cdots, z_n) = \sum_{i=1}^{7128} \log\left(\tfrac{1}{\sigma}\phi(z_i/\sigma)(1 - \rho + \rho\zeta_\alpha(z_i/\sigma)\right) + o(\rho)\,.$$

This is the log likelihood derived using the sparse approximation of order $o(\rho)$ and not order $o(\rho\epsilon_\nu^{-\alpha})$. In this way, the log likelihood is free of the negligibility threshold ϵ_ν, and so are the estimates of (ρ, α, σ^2), so that we can then set ϵ_ν as in (5.10) and match the sparsity-negligibility rejection region with that found by Efron's empirical null method.

Yet, because there is an appreciable asymmetry in the Leukemia z-values, we estimate

the sparsity parameters (ρ, α) separately for the 3719 positive observations, and the 3403 negative ones, holding fixed the common estimate for the scale parameter σ^2. From the estimation on the whole sample, we get $\hat{\sigma}^2$ about 1.86^2, and then obtain $(\hat{\rho}, \hat{\alpha}) = (0.02, 1.39)$ for the positive observations and $(\hat{\rho}, \hat{\alpha}) = (0.11, 1.68)$ for the negative ones.

The sparse marginal densities m_ν of order $o(\rho)$ corresponding to these estimates are shown in Figure 5.1, depicted by the light blue curves. We also plot the estimated sparse approximations of order $o(\rho\epsilon_\nu^{-\alpha})$ (dashed yellow curves). For these last ones, we set the negligibility thresholds to ϵ_ν^E, using the two triplets $(\hat{\rho}, \hat{\sigma}^2, \hat{\alpha})$, estimated separately for positive and negative z-values. The difference between the two sparse approximations is appreciable only for the negative part. In Figure 5.2 instead, we only plot the mixture component $(1-\hat{\rho})\frac{1}{\hat{\sigma}}\phi(z/\hat{\sigma})$ of the approximation of order $o(\rho)$, once again split for positive and negative z-values (dark green lines), and compare them with Efron's estimate $\hat{\pi}_0\hat{F}_0$, which is $0.93 \cdot N(0.09, 1.68^2)$ (red curve).

The data threshold z^E to control the estimated Bayesian Fdr

$$\overline{\mathrm{Fdr}}(z) = \frac{\hat{\pi}_0\hat{F}_0(z)}{\#\{z_i : z_i \leq z\}/n}\,,$$

at level $q = 10\%$ is found to be around 5.31. In order to match the rejection regions, using (5.10), the negligibility threshold for the signals generating the positive observations needs be 0.074, while that for those signals generating the negative z-values needs be 0.408. The corresponding sparse approximations to the probability of non negligibility are: $\mathbb{P}(|\mu| > 0.074) = 0.02 \cdot 0.074^{-1.39} = 0.68$ for the signals generating the positive observations; $\mathbb{P}(|\mu| > 0.408) = 0.11 \cdot 0.408^{-1.68} = 0.5$ for the signals generating the negative ones.

In some sense, this means that testing the sharp null hypothesis $\mu_i = 0$ after estimating

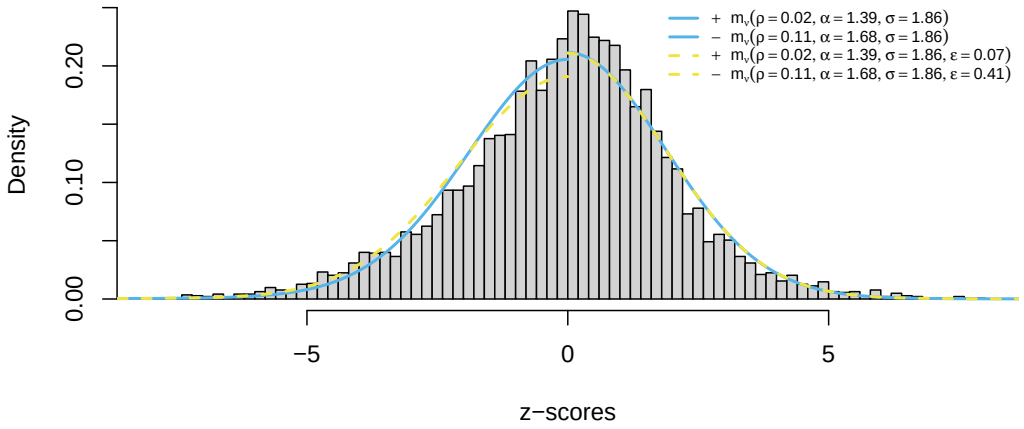


Figure 5.1: Estimated densities for Leukemia z-scores. Light blue solid curves correspond to the sparse approximations of order ρ to the marginal densities for positive and negative observations, respectively. Yellow dashed curves instead show the sparse approximation of order $\rho\epsilon_\nu^{-\alpha}$ to these two marginals.

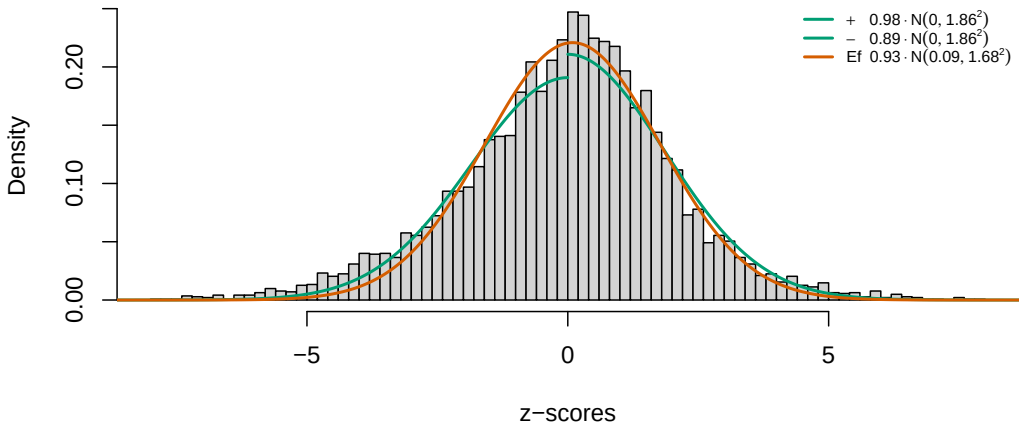


Figure 5.2: Estimated null densities for Leukemia z-scores. Dark green curves show the mixture components $(1-\rho)\frac{1}{\sigma}\phi(z/\sigma)$ appearing in the order ρ sparse approximations, for both positive and negative observations. The red curve, on the other hand, depicts the Efron's estimate of the empirical (sharp) null distribution, common to positive and negative values.

$\mathbb{P}(\mu_i \neq 0) = 0.07$, and after estimating the density of the Z_i's under the sharp null to be $N(0.09, 1.68^2)$, is equivalent to test the sparse negligibility hypothesis

- $|\mu_i| \leq 0.074$ for positive observations having sparse marginal density of order ρ

$$\tfrac{1}{1.86}\phi(z/1.86)\big(1 - 0.02 + 0.02\,\zeta_{1.39}(z/1.86)\big)\,, \quad z \geq 0$$

- $|\mu_i| \leq 0.408$ for negative observations having sparse marginal density of order ρ

$$\tfrac{1}{1.86}\phi(z/1.86)\big(1 - 0.11 + 0.11\,\zeta_{1.68}(z/1.86)\big)\,, \quad z < 0\,.$$

In Table 5.1 we report the empirical version of the conditional tail probability of the signal negligibility, given by

$$\hat{\mathbb{P}}(|\mu| \leq \epsilon_\nu \mid |Z| \geq |z|) = \frac{(1 - \hat{\rho}\epsilon_\nu^{-\hat{\alpha}})\Phi_{\hat{\sigma}}(|z|+)}{\#\{i \in [n] : |z_i| \geq |z|\}/n}\,,$$

corresponding to the top ten rejected z-values, negative and positive respectively. Notice that this also coincides with our empirical estimate of Storey's q-value.

Figure 5.3 shows $\hat{\mathbb{P}}(|\mu| \leq \epsilon_\nu \mid |Z| \geq |z|)$ for the Leukemia z-scores, split according to their sign. The dashed vertical lines indicate the smallest (in absolute value) z-value rejected by the sparsity-negligibility procedure, with the negligibility thresholds as specified in the previous paragraph. As expected, the empirical conditional tail probability of signal negligibility, which is also the sparse version of $\overline{\mathrm{Fdr}}(z)$, is below the level $q = 10\%$ for all z-values whose signal negligibility hypothesis is rejected.

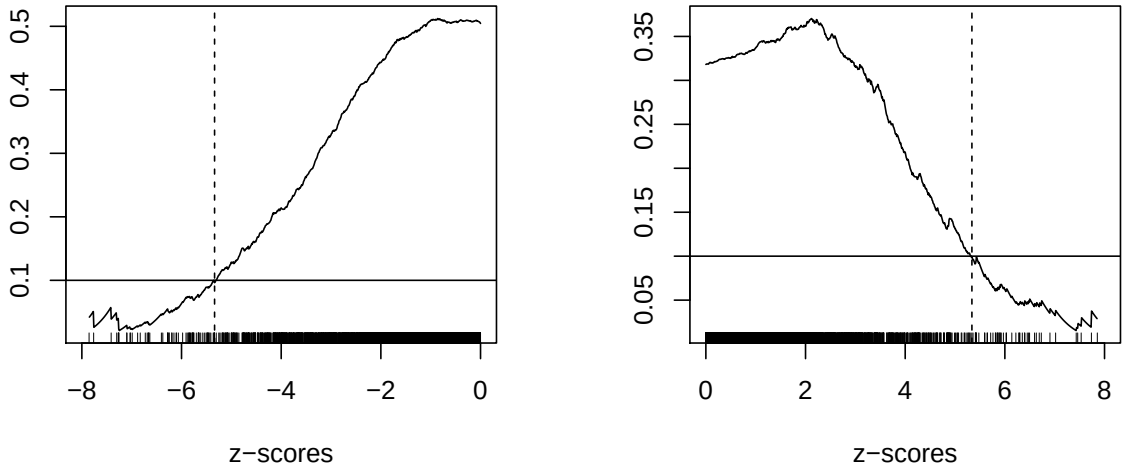


Figure 5.3: Estimated conditional tail probability of signal negligibility for the Leukemia dataset z-scores. The left panel shows $\hat{\mathbb{P}}(|\mu| \leq 0.408 \mid |Z| \geq |z|)$ as a function of $z < 0$ while the rug reports the observed negative z-scores. The right panel shows $\hat{\mathbb{P}}(|\mu| \leq 0.074 \mid |Z| \geq |z|)$ as a function of $z > 0$ together with the observed positive z-scores. The dashed vertical lines indicate the smallest (in absolute value) z-score rejected by the sparsity-negligibility procedure having as negligibility thresholds those specified above.

z-score	$\hat{\mathbb{P}}(\lvert\mu\rvert \leq 0.408 \mid \lvert Z\rvert \geq \lvert z\rvert)$	z-score	$\hat{\mathbb{P}}(\lvert\mu\rvert \leq 0.074 \mid \lvert Z\rvert \geq \lvert z\rvert)$
-7.855	0.042	7.855	0.029
-7.763	0.026	7.737	0.019
-7.412	0.039	7.531	0.020
-7.300	0.037	7.462	0.018
-7.268	0.032	7.432	0.015
-7.267	0.027	7.019	0.032
-7.264	0.023	6.905	0.035
-7.255	0.021	6.731	0.044
-7.097	0.026	6.690	0.042
-7.097	0.023	6.631	0.043

Table 5.1: Top ten rejected z-values, positive and negative, with the corresponding tail conditional probability that the underlying signal is negligible.

Chapter 6

Sparsity for wavelet regression

6.1 Introduction

In the Gaussian non-parametric regression problem, the observations of an unknown function f, are taken at regularly spaced points, and are subject to noise. In formulae, one can write

$$Y(t_i) = f(t_i) + \xi(t_i) \quad i = 1, \dots, T\,, \tag{6.1}$$

where $\xi(t_i)$ are independent random variables with distribution $N(0, \sigma_\xi^2)$, the variance of the errors being constant, while f is the unknown mean function, which is the object of interest. The index $i \in [T]$ is derived from the ordered set $\{t_i\}$ of T regularly spaced points in a one-dimensional space.

The wavelet approach to the estimation of f entails expressing $f \in L^2(\mathbb{R})$ in terms of an orthonormal basis of $L^2(\mathbb{R})$, $\{\psi_{jk}\}_{j,k\in\mathbb{Z}}$, called wavelet basis,

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \theta_{jk}\psi_{jk}(x)\,,$$

and estimating the wavelet coefficients

$$\theta_{jk} = \langle f, \psi_{jk} \rangle = \int f(x)\psi_{jk}(x)\, dx \quad j, k \in \mathbb{Z}\,, \tag{6.2}$$

to obtain an estimate of f. The underlying rationale for carrying out the estimation in the wavelet domain rather than in the original domain is that if the unknown function has some kind of spatial structure, then it is possible to well approximate f by a function whose wavelet representation is sparse, in the sense that the majority of its wavelet coefficients are zero or very near zero, and relatively few of them are in fact non-negligible. The second reason why the estimation in the wavelet domain is convenient, is that, when represented in wavelet form, a Gaussian function is still a Gaussian function so that, in the wavelet domain, the observed function can indeed be seen as the superposition of a sparse signal function and a Gaussian function. This in turn, allows one to carry out signal detection in the wavelet domain, and be able to reconstruct separately, the two components in the original domain, f and ξ.

Now, since wavelet regression in some sense reduces to signal processing, in its form of shrinking or thresholding the observed coefficients to estimate the unknown function coefficients, it goes without saying that a plethora of methods are available to perform this task. In this chapter, after giving some background on wavelet regression and a short review of some Bayesian approaches to this problem, we illustrate how it can be formulated within the sparsity and negligibility framework. In fact, we take the problem of estimating the wavelet coefficients θ_{jk} as an opportunity to compare the two sparse integral approximations, which we first discussed in the introduction. The original sparse approximation of order $o(\rho)$ can in fact be used to obtain a shrinking estimator for θ_{jk}, while the negligibility sparse approximation of order $o(\rho\epsilon_\nu^{-\alpha})$ can be employed to obtain a soft-thresholding estimator for θ_{jk}. To have an external benchmark, we also compare these two sparsity proposals to a

cornerstone in the literature of wavelet estimation, which is the empirical Bayes approach proposed by Johnstone and Silverman (2005) [47].

6.2 Wavelet regression

The discrete wavelet basis functions $\{\psi_{jk}\}_{j,k\in\mathbb{Z}}$ are generated by dilation and translation of the mother wavelet ψ as follows

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j(x - k/2^j)) \quad j,k \in \mathbb{Z}\,.$$

Here the index j defines the resolution level, in that it determines the width 2^{-j} of the equal-length intervals in which the domain of the mother wavelet is split up. Higher values of j correspond to finer resolutions while lower values correspond to coarser ones. The index k instead determines the location, since for each level of resolution j, it identifies which of the 2^j intervals to consider. Because of this construction by dilation and translation, wavelets are said to be localized in both time and frequency domains. The mother wavelet ψ has two peculiar properties: it oscillates above and below zero in such a way that has at least the first moment equal to zero, $\int \psi(x)dx = 0$, and has fast decaying tails, if not compact support. The first property aims at capturing a variation feature in the function, while the second property constraints this variation to be local. With this in mind, for any given j, the wavelet coefficients in (6.2) capture the amount of local variation happening in f at resolution level j.

So now suppose that f is spatially structured, in the sense that $f(t)$ and $f(t')$ are expected to be similar if $|t-t'|$ is small and abrupt changes are allowed as quite rare exceptions. Then, by virtue of the properties of the wavelet basis described above, f in the wavelet domain, will have coefficients at finer resolutions which will be mostly close to zero, with only few large exceptions. Intuitively, this is why the wavelet coefficients can be effectively characterized by sparsity patterns when the underlying function exhibits some spatial structure.

Borrowing from the engineering language the notion of energy of a vector as its L^2 norm, by Parseval's identity, one can show that the energy of the original signal is conserved in the wavelet domain but gets compressed into few high-energy coefficients. (*)

Yet, as mentioned above, the amount of sparsity induced in the wavelet representation of a function crucially depends on the function itself. For instance, if $\xi(t)$ is a Gaussian process (GP) with zero mean and covariance function $C(t, t') = \delta_{t,t'}$, then its wavelet transform is also a GP with zero mean and same covariance function. Therefore, no sparsity is induced in a Gaussian random function when passing from the original domain to the wavelet domain.

In the context of data observed at a finite number of points $\{t_1, \ldots, t_T\}$, usually $T = 2^J$, the discrete wavelet transform (DWT) maps a vector Y of length T to a vector $\tilde{Y}$ of equal length, containing its wavelet coefficients, all but the first, corresponding to a translation-dilation of the mother wavelet function ψ:

$$\tilde{Y}_{jk} = \langle Y, \psi_{jk} \rangle \quad j = 0, \ldots, J-1,\, k = 0, \ldots, 2^j - 1\,.$$

Mallat (1989) showed that the DWT can be represented by an orthogonal matrix W which stores in each row a scaled wavelet vector, i.e., a vector with entries given by the values of a wavelet function ψ_{jk} at the T equally-spaced points $t_i = i/T$. In other words, identifying each row index $l \in \{1, \ldots, T\}$ with a pair (j, k), the W matrix is such that $\sqrt{T}W_{l,i} = \sqrt{T}W_{jk,i} = \psi_{jk}(i/T) = 2^{j/2}\psi(2^j i/T - k)$.

If the observation process $Y(t)$ is the superposition of a signal process $f(t)$ and a white noise process $\xi(t)$, as in (6.1), then the discrete wavelet transform of the observed vector

$Y = (Y(t_1), \ldots, Y(t_T))$,

$$WY = Wf + W\xi\,,$$

is the sum of a compressed signal $\tilde{\theta} = W(f(t_1), \ldots, f(t_T))$, and an unchanged Gaussian noise $\tilde{\eta} = W(\xi(t_1), \ldots, \xi(t_T))$. Since $\tilde{\theta}$ concentrates the energy of f in just few components, while the energy of ξ remains equally distributed among the components of $\tilde{\eta}$, the observed sum in the wavelet domain,

$$\tilde{Y} = \tilde{\theta} + \tilde{\eta}\,,$$

exhibits an appreciably improved signal-to-noise ratio between the two components. This in turn means that we can estimate the wavelet coefficients $\tilde{\theta}$, by applying some form of thresholding or shrinkage to $\tilde{Y}$ in order to remove the noisy part and be left with the sparse wavelet representation of f. After carrying out this denoising procedure on $\tilde{Y}$, the last step is to return to the original domain by taking the inverse discrete wavelet transform W^{-1}, so to obtain an estimate of f.

In practice, it is not necessary to perform matrix multiplications as both the DWT and its inverse can be computed using the pyramid filtering algorithm proposed by Mallat (1989) [50]. This allows one to reduce the computational time from $O(T^2)$ to $O(T)$, beating the computational time of another very common transform in signal processing, the fast Fourier transform, which instead takes $O(T \log T)$.

(*) Note: the reason why the wavelet expansion gives rise to fewer high-energy coefficients than other orthogonal function basis has to do with the fact the wavelet coefficient sequences are highly symmetric about the coefficient axis and any rotation applied to these sequences would result in a loss of such symmetry. For an exhaustive and rigorous discussion of this fascinating topic, we refer to Donoho (1993) [22].

6.3 Bayesian approaches to wavelet regression

Within the Bayesian paradigm, the expected sparsity in the representation of the unknown function in the wavelet domain is naturally accomodated by a precise choice of the prior distribution for the underlying wavelet coefficients. Since the beginning of its development, the Bayesian approach has seen many different proposals for this prior distribution. For example, Clyde, Parmigiani and Vidakovic (1998) [18], and Abramovich, Sapatinas and Silverman (1998) [1], among others, considered an atom-and-slab mixture prior with a Gaussian distribution as the slab component, while Chipman, Kolaczyk and McCulloch (1997) [17] proposed a mixture of two Gaussian distributions, one concentrated around zero and the other dispersed. More recently, Xing, Carbonetto and Stephens (2021) [79] extended the two-component scale mixture of normals to a K-component scale Gaussian mixture, with the possibility of including the Dirac delta measure as a degenerate Gaussian distribution. A cornerstone in the Bayesian wavelet estimation literature is the paper by Johnstone and Silverman (2005) [47] (J&S henceforth), who proposed a class of prior distributions given by the mixture

$$(1-w)\delta_0(du) + w\Gamma(du)\,,$$

where the non-zero measure Γ is assumed to have a unimodal symmetric density γ. Clearly, the normal density is a viable option for γ but, following Wainwright, Simoncelli and Willsky (2001) [73], J&S emphasize the advantage, in the wavelet context, of choosing γ to have heavier tails, such as those of the Laplace density. As an another possibility for γ, the authors proposed a scale mixture of normals, $N(0, \kappa^{-1} - 1)$ with a Beta prior on $\kappa \sim Be(1/2, 1)$. The resulting density

$$\gamma(u) = (2\pi)^{-1/2}\left(1 - |u|\big(1 - \Phi(|u|)\big)/\phi(u)\right)\,, \tag{6.3}$$

has tails decaying as u^{-2}, for which reason they name it 'quasi-Cauchy density'. This scale Gaussian mixture was in fact already considered by Strawderman (1971) [69], as well as Berger (1980) [11].

For any given choice of the prior for the signal wavelet coefficients $\tilde{\theta}$, the corresponding conditional distribution, given the observed coefficients $\tilde{Y}$, can be utilized to obtain an estimator for $\tilde{\theta}$. The conditional mean, considered by Clyde, Parmigiani and Vidakovic (1998) [18] and Chipman, Kolaczy and McCulloch (1997) [17] among others, shrinks the observed coefficients towards zero and has been shown to give good results. As an alternative, Abramovich, Sapatinas and Silverman (1998) [1] proposed to use the conditional median. The advantage of this last choice is that, if the prior contains an atom at zero, then the conditional median takes the form of a soft-thresholding operator. In this case, a subset of observed coefficients is thresholded to zero while the remaining coefficients are shrunk to zero by an amount depending on their size.

In the next section, we describe how we adapt the approach of J&S to the sparsity-negligibility framework, taking the wavelet coefficient estimation as a chance to make a comparison between two sparse integral approximations: the original sparse $o(\rho)$ approximation, and the sparse-negligibility $o(\rho\epsilon_\nu^{-\alpha})$ approximation.

6.4 Model assumptions and estimation

6.4.1 Sparsity assumptions

Since wavelet coefficients are naturally grouped by their level of resolution, it is appropriate to model each level separately, and, in light of the discussion of Section 6.2, we only model the coefficients of the higher levels of resolution, say from level m to level J. For each of these higher levels, we assume that the coefficients are independently distributed with a

level-specific sparse distribution P_ν^j. This means that for $j = m, \ldots, J$, the observed j-level coefficients

$$\tilde{Y}_{jk} = \tilde{\theta}_{jk} + \tilde{\eta}_{jk} = \sigma_j(\theta_{jk} + \eta_{jk}) \qquad k = 1, \ldots, 2^j \,,$$

are the scaled sum of a sparse signal $\theta_{jk} \sim P_\nu^j$ and an independent noise $\eta_{jk} \sim N(0,1)$. Here P_ν^j is a sparse distribution with first-order sparsity pair (ρ_j, H_j); H_j is an inverse-power exceedance measure; σ_j is the j-level scale parameter. Allowing a different scale parameter for each level of resolution is appropriate, especially when the noise function exhibits some form of autocorrelation in the original domain. In fact, Johnstone and Silverman (1997) [44] explains that, even when there is an appreciable autocorrelation in the Gaussian noise process $\xi(t)$, its wavelet transform yield coefficients with much less dependence. We refer the reader to Section 6 of the aforementioned paper for a formal treatment. For simplicity of exposition, from now on, we suppress the j-level script, and refer to $\tilde{Y}_k$, θ_k and η_k as the random variables at level j, implicitly meaning that the same estimation procedure can be carried out for each level of resolution separately.

As mentioned in the introduction to this chapter, depending on what kind of regularization rule we want to adopt, we consider different sparse integral approximations to P_ν. In fact, if the smoothing of the observed function is to achieve by shrinking the observed wavelet coefficients, then we can use the conditional mean of θ_k. For this shrinkage to happen, no atom at zero is needed in the integral sparse approximation to $P_\nu(dx)$, so that the approximation of order $o(\rho)$,

$$\rho H(dx) \,, \tag{6.4}$$

can be employed. Under this approximation, the marginal density of $\tilde{Y}_k$ at $\tilde{y}$ is

$$m_\nu(\tilde{y}) = \phi_\sigma(\tilde{y}) \, (1 - \rho + \rho\zeta(\tilde{y}/\sigma)) + o(\rho) \,, \tag{6.5}$$

where $\phi_\sigma(z) = 1/\sigma\,\phi(z/\sigma)$, and the moment generating function of the conditional distribution of θ_k, given $\tilde{Y}_k = \tilde{y}$, is

$$\int e^{tx} P_\nu(dx \mid \tilde{y}) = \int \frac{e^{tx} e^{x\tilde{y}/\sigma - x^2/2}}{m_\nu(\tilde{y})/\phi_\sigma(\tilde{y})}\, P_\nu(dx) = \frac{\rho\zeta(\tilde{y}/\sigma + t)}{1 - \rho + \rho\zeta(\tilde{y}/\sigma)} + o(\rho)\,.$$

By Eddington-Dyson's formula, the conditional mean of θ_k is

$$\mu(\tilde{y}/\sigma) = \frac{\rho\zeta'(\tilde{y}/\sigma)}{1 - \rho + \rho\zeta(\tilde{y}/\sigma)}\,,$$

so the shrinkage estimator of the k-translated coefficient at level j is

$$\sigma\mu(\tilde{Y}_k/\sigma)\,. \tag{6.6}$$

If instead one wants to denoise the observed function by means of some sort of thresholding in the wavelet domain, then the conditional median of θ_k can be used as a soft-thresholding rule, provided that the sparse integral approximation to P_ν has an atom at zero. In this case, we consider the integral approximation of order $o(\rho\epsilon_\nu^{-\alpha})$ introduced in Chapter 4,

$$(1 - \rho\epsilon_\nu^{-\alpha})\delta_0(du) + \rho\epsilon_\nu^{-\alpha}\tilde{H}(du)\,. \tag{6.7}$$

Here ϵ_ν is a negligibility sequence for P_ν according to Definition 4.5.1. Henceforth, we consider ϵ_ν to be the function $(\log(1/\rho))^{-1/2\alpha}$, so that the unconditional probability of non-negligibility $\rho\epsilon_\nu^{-\alpha} = \rho\sqrt{\log(1/\rho)} \to 0$ as $\nu \to 0$. Under this approximation, the marginal density of $\tilde{Y}_k$ at $\tilde{y}$ is

$$m_\nu(\tilde{y}) = \phi_\sigma(\tilde{y})(1 - \rho\epsilon_\nu^{-\alpha} + \rho\epsilon_\nu^{-\alpha}A(\tilde{y}/\sigma)) + o(\rho\epsilon_\nu^{-\alpha})\,, \tag{6.8}$$

where we recall that

$$A(y) = \int \cosh(yx) e^{-x^2/2} \, \tilde{H}(dx) = \epsilon_\nu^\alpha (\zeta(y) - \tau^\alpha \zeta(y/\tau) - 1 + \tau^\alpha) \,,$$

for $\tau^2 = 1 + 1/\epsilon_\nu^2$. Then the sparse approximation of order $o(\rho \epsilon_\nu^{-\alpha})$ to the conditional distribution of θ_k, given $|\tilde{Y}_k| = \tilde{y}$, is

$$(1 - w(\tilde{y}/\sigma))\delta_0(dx) + w(\tilde{y}/\sigma)\bar{A}(dx; \tilde{y}/\sigma) \,, \tag{6.9}$$

which is a mixture of a Dirac delta measure at zero, and the non-zero measure

$$\bar{A}(dx; y) = \frac{\cosh(yx) e^{-x^2/2} \tilde{H}(dx)}{\int \cosh(yx) e^{-x^2/2} \, \tilde{H}(dx)} \,,$$

with relative weight

$$w(y) = \frac{\rho \epsilon_\nu^{-\alpha} A(y)}{\rho \epsilon_\nu^{-\alpha} A(y) + 1 - \rho \epsilon_\nu^{-\alpha}} \,.$$

Denote by $q^{0.5}(\tilde{y}/\sigma)$ the conditional median of θ_k given $\tilde{Y}_k = \tilde{y}$, i.e., $q^{0.5}(\tilde{y}/\sigma)$ is the value q for which

$$\int_{-\infty}^{q} P_\nu(dx \mid \tilde{y}) = 0.5 \,.$$

Strictly speaking, the indicator function $\chi_{(-\infty,q]}(x)$ is not a continuous function; nevertheless it can be approximated to any level of precision by some bounded and continuous function. So, because the conditional distribution is

$$P_\nu(dx \mid \tilde{Y}) = \frac{e^{x\tilde{y}/\sigma}}{\cosh(x\tilde{y}/\sigma)} P_\nu(dx \mid |\tilde{Y}|) \,,$$

applying the integral approximation (6.9) to $P_\nu(dx \mid |\tilde{Y}|)$, we obtain the sparse approximation

$$\int_{-\infty}^{q} P_\nu(dx \mid \tilde{y}) = (1 - w(\tilde{y}/\sigma)) \int_{-\infty}^{q} \delta_0(dx) + w(\tilde{y}/\sigma) \int_{-\infty}^{q} \frac{e^{x\tilde{y}/\sigma}}{\cosh(x\tilde{y}/\sigma)} \bar{A}(dx; \tilde{y}/\sigma) + o(\rho \epsilon_\nu^{-\alpha})$$

So, letting $a(\mu; y) = \int_{-\infty}^{\mu} \frac{e^{xy}}{\cosh(xy)} \bar{A}(du; y)$, consider $\tilde{y} > 0$. The conditional median $q^{0.5}(\tilde{y})$ is zero if

$$1 - w(\tilde{y}/\sigma) + w(\tilde{y}/\sigma)\, a(0; \tilde{y}/\sigma) \geq 0.5\,;$$

otherwise $q^{0.5}(\tilde{y})$ is equal to the value $q^{\star}$ for which

$$1 - w(\tilde{y}/\sigma) + w(\tilde{y}/\sigma)\, a(q^{\star}; \tilde{y}/\sigma) = 0.5\,.$$

Putting these last two observations together, we have that the conditional median, given $\tilde{y} > 0$, can be written as

$$q^{0.5}(\tilde{y}/\sigma) = \begin{cases} 0 & \text{if } 1 - w(\tilde{y}/\sigma) + w(\tilde{y}/\sigma)\, a(0; \tilde{y}/\sigma) \geq 0.5\,, \\ q^{\star} & \text{else.} \end{cases}$$

By antisymmetry of the conditional median, for $\tilde{y} < 0$, $q^{0.5}(\tilde{y}/\sigma) = -q^{0.5}(-\tilde{y}/\sigma)$. So, the conditional median naturally gives rise to the soft-thresholding estimator for the k-translated coefficient at level j,

$$\sigma q^{0.5}(\tilde{Y}_k/\sigma)\,. \tag{6.10}$$

In Figure 6.1, we plot the conditional mean in (6.6), and conditional median in (6.10), as functions of the observed $\tilde{y}$, for values of $\rho = 0.05, 0.1, 0.2$ and $\alpha = 0.5, 1, 1.5$, while σ is fixed to one. We can see how the conditional mean (red curve) shrinks all values but does not threshold any of them to zero; by contrast, the conditional median (light blue curve) defines a symmetric region around the origin,

$$\{\tilde{y} : w(|\tilde{y}|)\ (1 - a(0; |\tilde{y}|)) \leq 0.5\}\,, \tag{6.11}$$

which gets thresholded to zero. The curve $w(|\tilde{y}|)\ (1 - a(0; |\tilde{y}|))$ is plotted for positive $\tilde{y}$ in Figure 6.2.

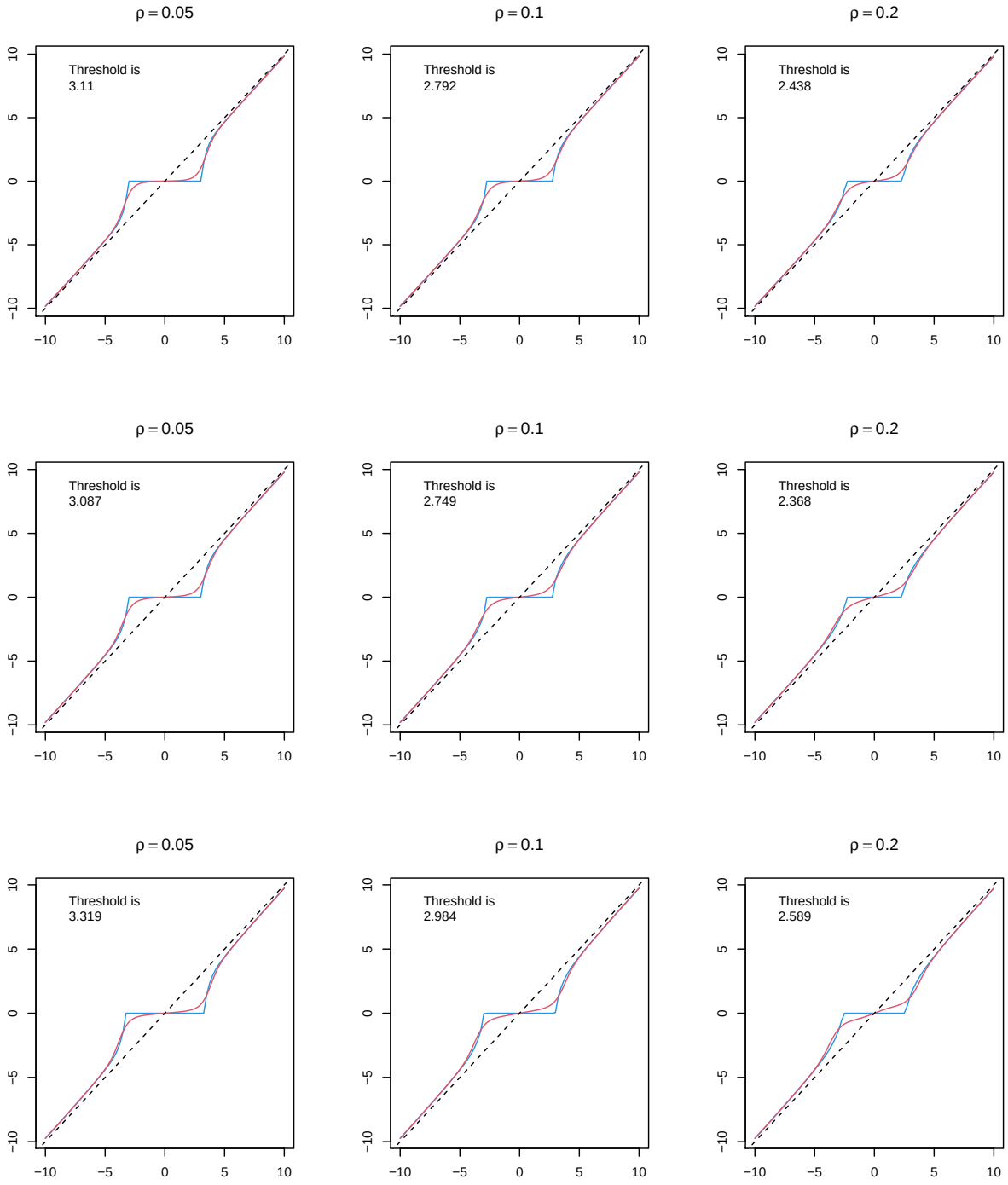


Figure 6.1: Conditional median $q^{0.5}(\tilde{y})$ (blue curve) and conditional mean $\mu(\tilde{y})$ (red curve) for different values of the sparsity parameters ρ and α. The top three plots have $\alpha = 0.5$, the middle ones have $\alpha = 1$ while the bottom ones correspond to $\alpha = 1.5$. In each plot, the threshold refers to the $\tilde{y}$-value that determines the region in (6.11) which is thresholded to zero by the conditional median.

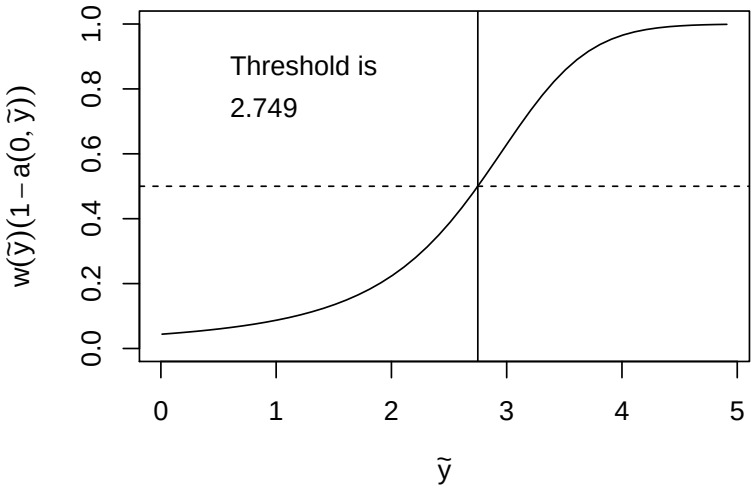


Figure 6.2: Plot of the curve $w(\tilde{y})\,(1 - a(0; \tilde{y}))$. The black vertical line depicts the $\tilde{y}$-threshold that defines the region of $\tilde{y}$-values for which the corresponding signal estimate gets thresholded to zero, bu the conditional median. The parameters are set to $\rho = 0.1$, $\alpha = 1$, and $\sigma = 1$.

6.4.2 Translation-invariant wavelet

Despite their well developed theory, which includes results of near optimality in comparison to other methods, wavelet de-noising methods were criticized for producing some kind of artifacts in proximity of the discontinuities in the underlying function. As explained in Coifman and Donoho (1995) [19], wavelet reconstructions can excessively oscillate up and down, whenever a discontinuity in the signal function occurs not exactly at one point of the dyadic segmentation of the domain. In other words, the alignment of the wavelet basis elements with the discontinuities of the signal function is crucial for the success of the signal reconstruction. From this perspective, it comes natural to consider shifting the signal in such a way to produce a better alignment and eliminate the unwanted artifacts. This means that one can apply a circulant shift operator on the signal, denoise the shifted version and then shift this back. However, when the function exhibits multiple discontinuities, it is not obvious

which shift would be best to apply, as a shift aligning one discontinuity, could be detrimental for moving another one. For this reason, Coifman and Donoho (1995) [19] proposed to apply a range of possible shifts and then average over the results obtained from each of these shifts. When the average is taken over all T possible circulant shifts, this procedure, called 'cycle-spinning', is translation invariant, and produces a T-long vector of coefficients for each level of resolution, since at level j, the number of coefficients 2^j gets multiplied by the number of possible shifts 2^{J-j}. This translation invariant wavelet transform (TIDWT) can be computed rapidly with only $O(TJ) = O(T \log_2 T)$ computations. For a different exposition of the TIDWT, see Section 2.7.2 of Nason (2008) [55] and references therein.

6.4.3 Parameter estimation

As in J&S and many others after them taking an empirical Bayes approach, for each level of resolution, we estimate the sparsity parameters by maximum likelihood. However in our setting, as explained in Section 6.4.1, depending on the regularization rule we want to use to smooth the observed function, we consider two different integral approximations to P_ν, which in turn imply two different approximations to the marginal density of $\tilde{Y}_k$. With the former approximation in (6.5), given the observed coefficients $\{\tilde{y}_k\}_k$ from the standard DWT, the log likelihood for the j-level sparsity parameters is

$$\log L(\rho, \alpha, \sigma^2; \tilde{y}_0, \ldots, \tilde{y}_{2^j-1}) = \sum_{k=0}^{2^j-1} \log\{\phi_\sigma(\tilde{y}_k)\,(1 - \rho + \rho\zeta(\tilde{y}_k/\sigma))\}\,. \tag{6.12}$$

Instead, if we use the approximation in (6.8), the log likelihood is

$$\log L(\rho, \alpha, \sigma^2; \tilde{y}_0, \ldots, \tilde{y}_{2^j-1}) = \sum_{k=0}^{2^j-1} \log\{\phi_\sigma(\tilde{y}_k)\,\big(1 - \rho\epsilon_\nu^{-\alpha} + \rho\epsilon_\nu^{-\alpha} A(\tilde{y}_k/\sigma)\big)\}\,. \tag{6.13}$$

Yet, if we work with the TIDWT, then the T coefficients obtained for each level j are not independent, and in principle one should consider separately the different packets corre-

sponding to the different choices of the origin. However, the tendency in the literature has been instead to treat the entries of the whole T-long coefficient vector as independent and maximize a log likelihood averaged over the choice of origin. So letting $\{\tilde{y}_{p,0}, \ldots, \tilde{y}_{p,2^j-1}\}$ denote the observed coefficients in the p packet of level j, for each j, we maximize

$$\bar{l}(\rho, \alpha, \sigma^2) = \frac{1}{2^{J-j}} \sum_{p=1}^{2^{J-j}} \log L(\rho, \alpha, \sigma^2; \tilde{y}_{p,0}, \ldots, \tilde{y}_{p,2^j-1}) \tag{6.14}$$

in such a way to borrow strength in the estimation of (ρ, α, σ) between the different locations of the origin. For a more detailed discussion about this 'as-if-independence' strategy, we refer the reader to J&S. The estimated parameters can then be used to perform one of the shrinkage or thresholding procedures discussed above, on the observed coefficients of the TIDWT. After this, the smoothed coefficients are transformed back on the original scale using an average basis approach, as mentioned at the end of the previous section.

6.5 Simulation study

In this section, we simulate some data to illustrate the functioning of the sparsity-negligibility methods described in the previous section. For the underlying signal function, we use the four test functions, Doppler, Bumps, Blocks and Heavisine, first considered by Donoho and Johnstone (1994) [23].

We set $T = 512$ and generate the signal vector μ over the interval $[0, 1]$ so that the points $t_1, \ldots, t_T$ are equally spaced by $1/2^9$. Then we simulate Y by adding to μ the vector η whose entries are independent and identically distributed as $N(0, \sigma_\eta^2)$, where σ_η^2 is set so that the signal-to-noise ratio is one. We transform the noisy signals using the Daubechies least-asymmetric wavelet basis with ten vanishing moments, which is the default option implemented by the wd command of the **wavethresh** R package (Nason, 2016 [56]). These

wavelets belong to a larger family of wavelets introduced by Daubechies (1988) [20], constructed with the aim of having compactly supported functions which were smoother than the original Haar wavelets.

Since $J = \log_2 T = 9$, after applying the translation invariant wavelet transform to Y, we estimate the sparsity parameters following Section 6.4.3, for the resolution levels ranging from four to eight. In Figure 6.3, we compare the estimates for (ρ, α, σ) obtained by maximizing $\bar{l}(\rho, \alpha, \sigma^2)$ using the two options for the log likelihood as in (6.12) and (6.13). As expected, the estimate for the sparsity rate ρ is decreasing as the resolution level increases, with the only exception of the case when a very large scale parameter is estimated for the coarsest level four. If we were to fix the scale parameter across levels, for instance estimating it by the mean absolute deviation (MAD) of the finest level coefficients, an option considered by J&S, the decreasing pattern of the sparsity rate would be much more pronounced.

Regarding the differences between using (6.12) versus (6.13) as log likelihood, one can observe that there are no major differences for the estimates of the α parameter and almost identical estimates for the scale σ. On the contrary, the second log likelihood, derived from using the atomic mixture integral approximation (6.7), leads to an estimated weight $\rho\epsilon_\nu^{-\alpha}$ for the component $A(y)$ in

$$m_\nu(\tilde{y}) = \phi_\sigma(\tilde{y})(1 - \rho\epsilon_\nu^{-\alpha} + \rho\epsilon_\nu^{-\alpha} A(\tilde{y}/\sigma)) + o(\rho\epsilon_\nu^{-\alpha}),$$

which is systematically higher than the estimated weight ρ for the component $\zeta(y)$ in

$$m_\nu(\tilde{y}) = \phi_\sigma(\tilde{y})\,(1 - \rho + \rho\zeta(\tilde{y}/\sigma)) + o(\rho)\,.$$

This is more clearly shown in Figure 6.4 where we plot these two estimated weights, or rates. To get a sense of why the weight in the first mixture is estimated to be higher than the

weight in the second one, it is helpful to look at the product $A(y)\phi(y)$ versus the product $\zeta(y)\phi(y)$, which are plotted in Figure 6.5, for different values of α and ρ. Indeed, even if the two functions have the same tail behavior, around the origin the behavior is very different, the former being unimodal while the latter is bimodal.

In order to give an overall idea of what the two estimation strategies, shrinkage and soft-thresholding, produce in terms of signal reconstruction, in Figure 6.6, over the original signal vector μ, we plot the estimated vector $\hat{\mu}$. This latter is obtained by averaging the transformed-back cycle-spinning coefficients, after de-noising. The dark blue line depicts the signal reconstructed using the conditional median of the TIDWT coefficients, while the light blue line shows the denoised signal obtained using the conditional mean. We also report two measures of error, the mean absolute error (MAE) and the mean square error (MSE). In general, the shrinkage method leads to lower errors, but this is at the cost of a loss in smoothness of the curve, as in fact, no coefficient gets shrunk to exactly zero when using the conditional mean.

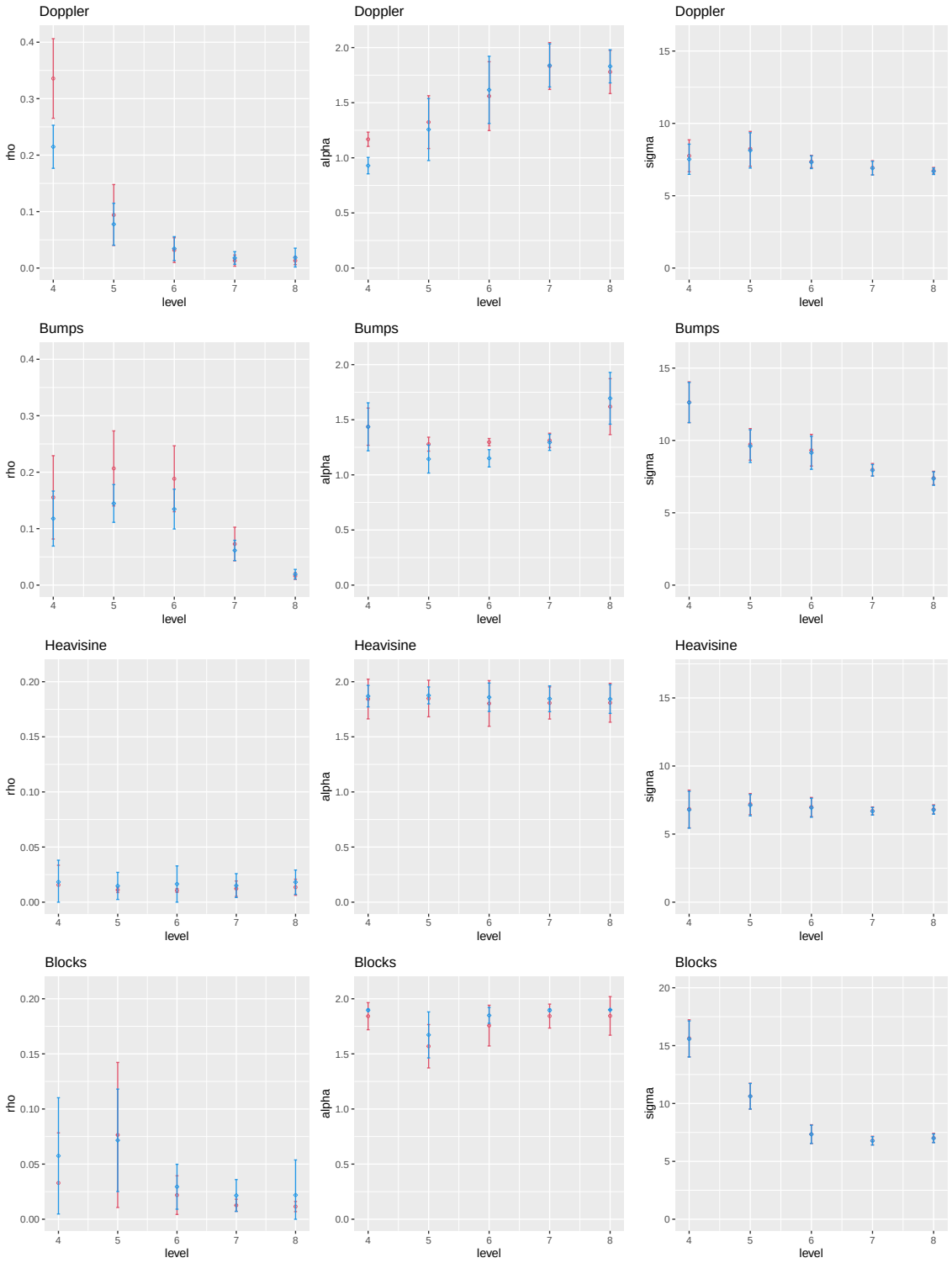


Figure 6.3: Plots of the estimates (averages with standard error bars) of the level-specific sparsity parameters obtained from maximizing $\bar{l}(\rho, \alpha, \sigma^2)$. The blue points and segments refer to the estimation using the log likelihood in (6.12) and will be used for shrinking; while the red figures refer to the estimation using the log likelihood in (6.13) and will be used for soft-thresholding.

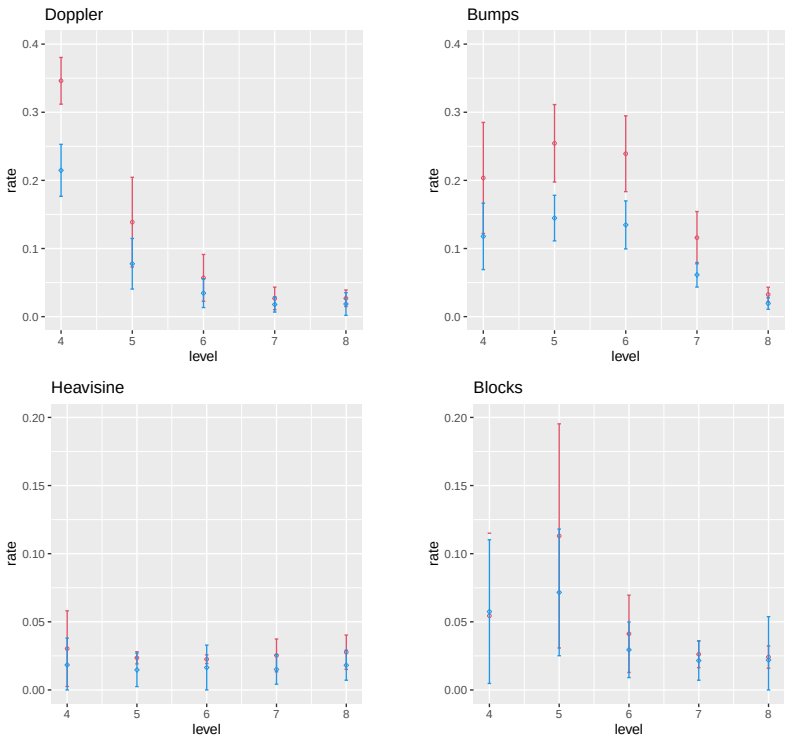


Figure 6.4: Plots of the estimates (averages with standard error bars) of the level-specific marginal mixture weights. The blue points and segments refer to the estimation of ρ in (6.12); the red figures refer to the estimation of $\rho\epsilon_\nu^{-\alpha}$ obtained from the estimates of ρ and α in (6.13).

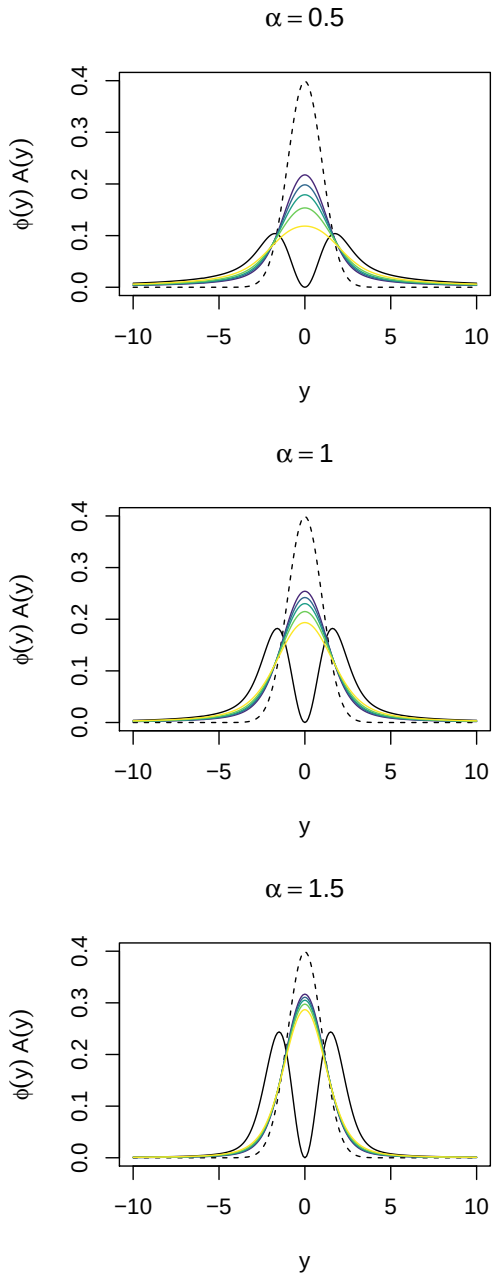


Figure 6.5: Comparison between $\phi(y)\zeta(y)$ (black curve) and $\phi(y)A(y)$ (colored curves). Different colors for $\phi(y)A(y)$ functions correspond to different values of ρ which imply different values of $\epsilon_\nu = (\log(1/\rho))^{-1/2\alpha}$. From dark blue to light yellow, the ρ parameter is $0.01, 0.025, 0.05, 0.10, 0.20$. The dashed black line depicts the standard Gaussian density.

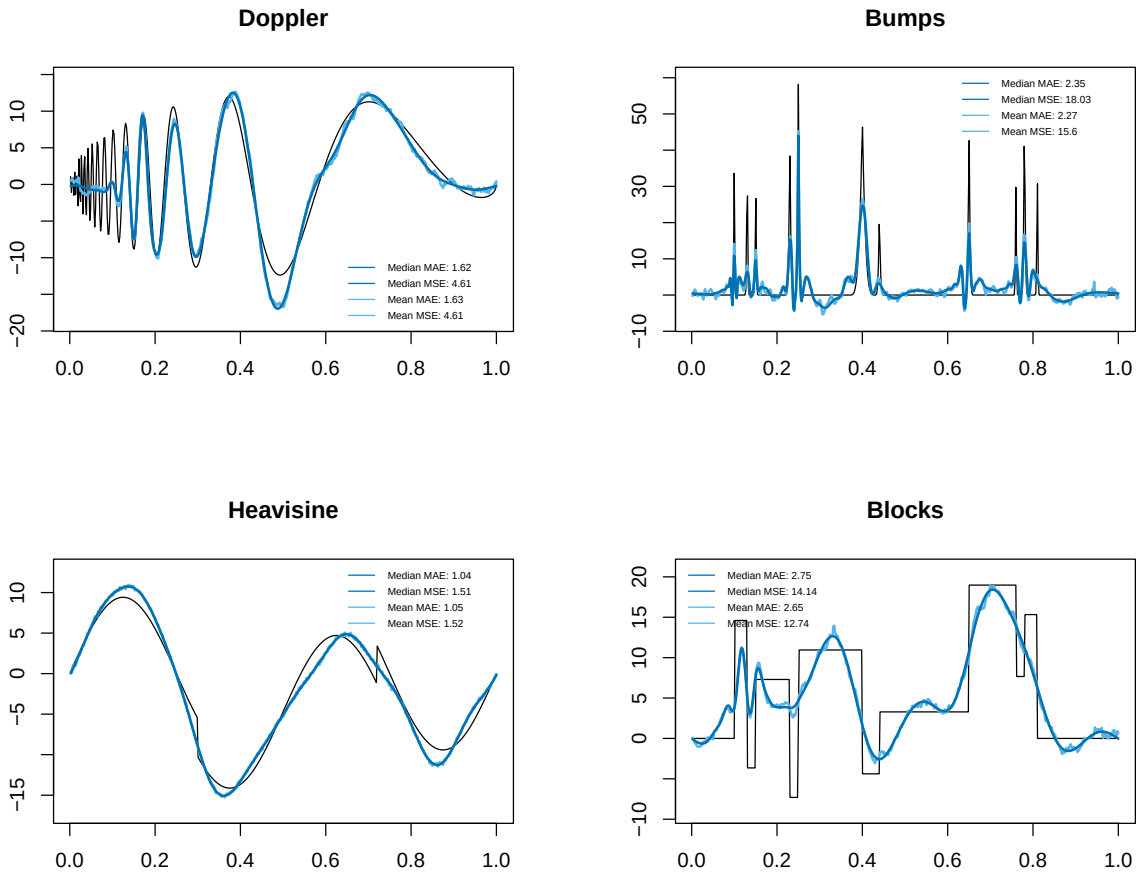


Figure 6.6: Plots of the estimated signal for the four functions: Doppler, Bumps, Heavisine and Blocks. The dark blue lines refer to the estimation using the conditional median for soft-thresholding the TIDWT coefficients while the light blue lines refer to the shrinkage estimation, using the conditional mean.

6.5.1 Sparsity and EbayesThresh

We now present a comparison between the two sparsity-based approaches described in Section 6.4.1, with the soft-thresholding and shrinkage methods described in J&S, derived from adopting either the 'quasi-Cauchy' or the Laplace density as the prior for the TIDWT coefficients. These last estimations are easily implemented in R using the **EbayesThresh** package (Johnstone and Silverman, 2005B [46]).

To see how the methods compare, Figure 6.7 shows the denoised TIDWT coefficients obtained for the Doppler function, using six different strategies. We consider the conditional median and the conditional mean obtained using: the sparse integral approximations (labelled as sparsity); the 'quasi-Cauchy' density (labelled as Cauchy); and the Laplace density (labelled as Laplace). In all cases, we only denoise the coefficients from level 4 to level 8, leaving those at coarser levels unchanged. No major differences are observed across the different distributions of the signal coefficients; while the impact of choosing to use the conditional mean or the conditional median is again observed in the number of coefficients that are estimated to be zero. In fact, when using the conditional median, no matter what distribution is chosen, the coefficients for the levels 6, 7 and 8 are thresholded to zero, so that those resolution levels do not contribute to the reconstruction of the signal vector. Accordingly, it is not surprising that the sparsity-based approaches overall perform quite similarly to the J&S's methods, both in terms of MAE and MSE. We summarize their performance for the reconstruction of the four signal functions in Figure 6.8.

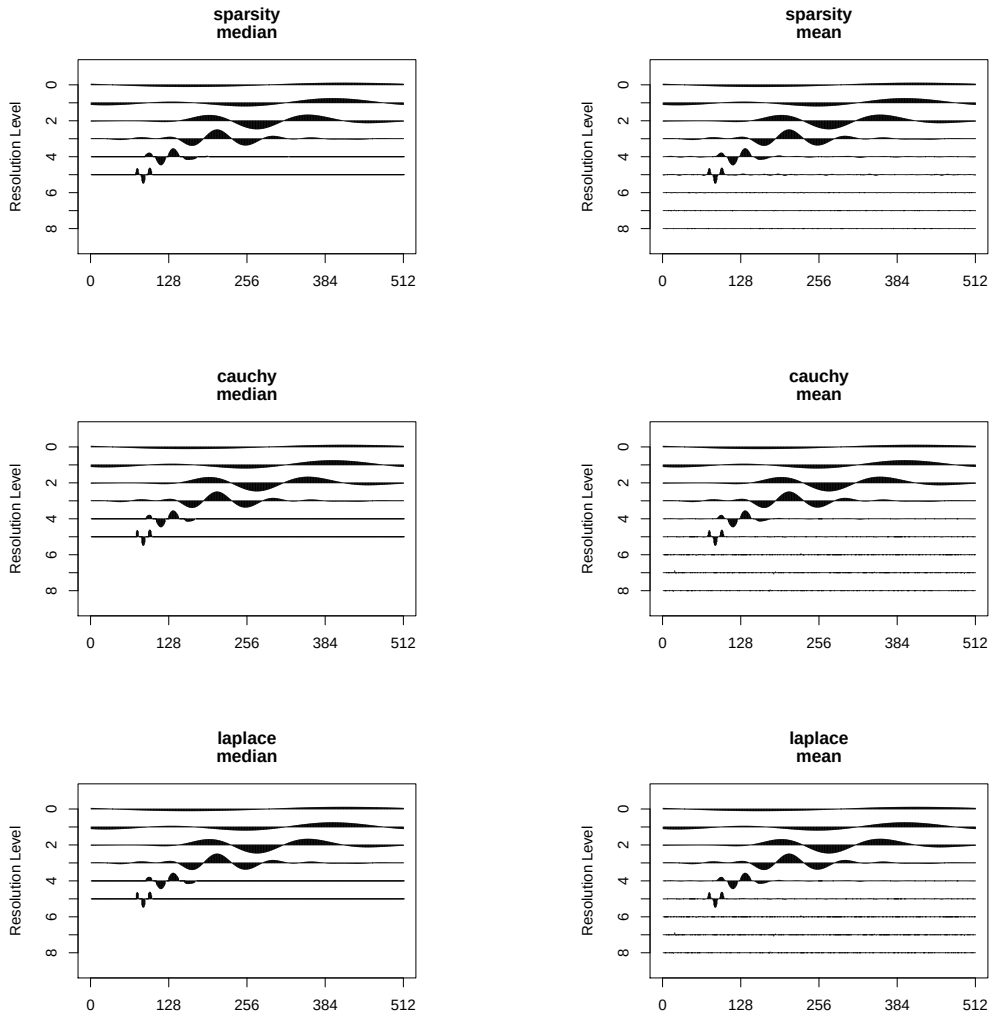


Figure 6.7: Plots of the TI wavelet coefficients for the Doppler function. These are estimated by either the conditional median or the conditional mean, under different formulations of the unconditional distribution for the signal wavelet coefficients: the sparsity models described in Section 6.4.1 (top), the 'quasi-Cauchy' prior as in (6.3) (middle), and the Laplace prior (bottom).

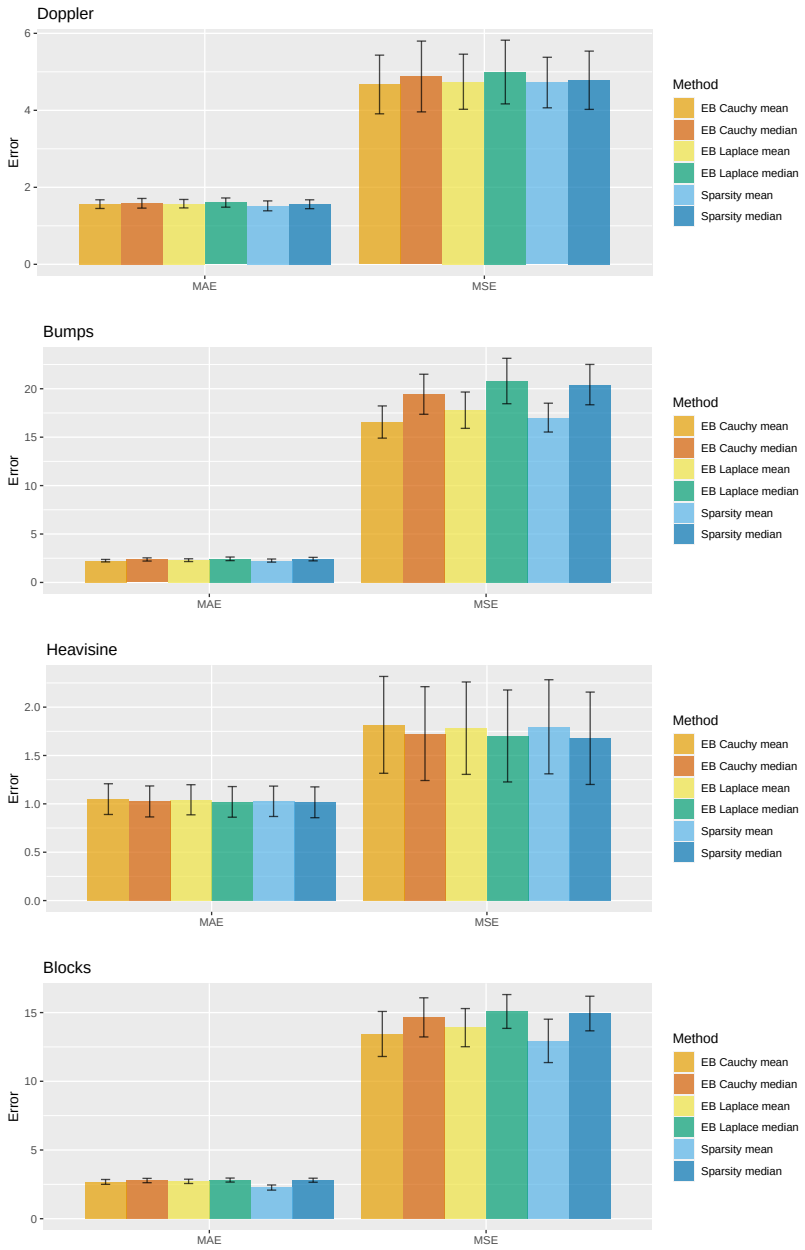


Figure 6.8: Mean square error (MSE) and mean absolute error (MAE) of different methods for reconstructing the signal of the four test functions. The height of the bars are the averages over 100 simulations, while the whiskers depict the standard errors. The scale parameter is let vary across levels. The signal to noise ratio is equal to one for all simulated functions.

6.5.2 Image smoothing

As a last example for comparison, we replicate the image smoothing exercise presented in Johnstone and Silverman (2005B) [46] (J&SB henceforth). The image is a black-and-white

photograph of Ingrid Daubechies, after whom the Daubechies wavelets are named. This image is contained in the **waveslim** package and is stored as a 256×256 matrix. After corrupting the image by adding to each entry of the matrix, a Gaussian noise having standard deviation of ten, we obtain the two-dimensional wavelet transform of the noisy image using the command `dwt.2d`. If interested in reproducing the code, the reader is referred to J&SB. We apply the sparsity soft-thresholding method to the coefficients contained in the first nine of the 13 matrices produced by `dwt.2d`. We group the nine matrices in sets of three since in a two-dimensional wavelet transform, there are three filters interacting with the original matrix for each resolution level: one for the vertical direction, one for the horizontal direction and one for the diagonal. For each of these three sets, we estimate the sparsity parameters by maximizing (6.14), with (6.13) as the approximated log likelihood. Once estimated the level-specific sparsity parameters, we estimate the coefficients of the nine matrices using the conditional median and then transform everything back to the original domain to obtain the denoised image.

As in J&SB, together with the original, noisy and sparse denoised image, in Figure 6.9, we also plot the image denoised using a kernel smoothing method. One can notice that the sparsity-based smoothed image has less background noise than the kernel smoothed image, whereas this latter, on the other hand, preserves some contrasts more faithfully.

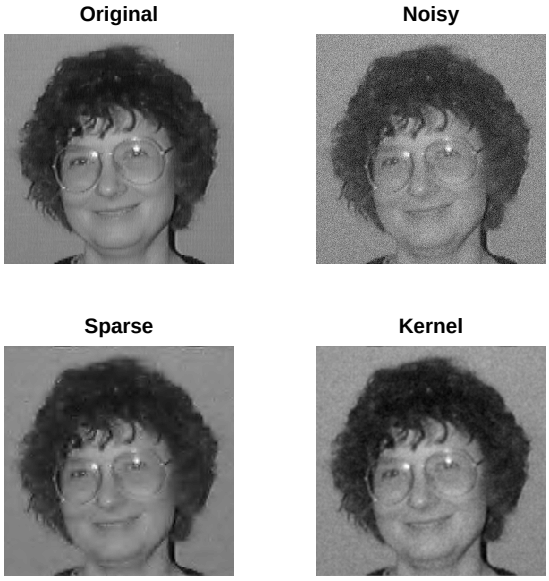


Figure 6.9: Comparison of images of Ingrid Daubechies: original (top left), noisy (top right), denoised using sparsity methods (bottom left) and denoised using kernel methods (bottom right).

Chapter 7

Sparsity for Gaussian graphical models

7.1 Introduction

Understanding the relationship between variables can sometimes be the main question of interest, especially in those settings where the number of recorded variables p is large. Indeed, data of this kind arise in many applications, such as gene array expression levels, climate data and spectroscopy, among many others.

Unfortunately, when the dimension p is large, inference on the covariance matrix Σ and on its inverse, the precision matrix Ω, becomes very problematic, if not unfeasible, and additional assumptions need to be made in order to handle the high dimensionality of the matrices. An assumption which has been very commonly considered in the recent literature on high-dimensional inference, is in fact sparsity.

Tapering, banding and thresholding (see for instance Bickel and Levina 2008A[12], 2008B [13], and the references given in the introduction of Banerjee and Ghosal, 2015 [5]) are all methods aimed at inducing a degree of sparsity in either Σ or Ω, and can be successfully applied in situations where there is some natural ordering in the underlying variables, for

instance in time series data or spatial data. When such natural ordering among the variables is not present, then it is necessary to have an estimation method that is invariant under variable permutations. In these cases, it is particularly useful to pose the problem in terms of inferring the underlying undirected graph structure. Indeed, when the data are assumed to be Gaussian, the precision matrix can be read in terms of its corresponding graphical model, where the vertices represent the variables and the presence or absence of an undirected edge between two vertices corresponds to the presence or absence of a dependence between the two variables, conditionally on all the others.

Gaussian graphical models have been extensively studied, from different perspectives. Penalized likelihood methods have been developed together with their corresponding optimization algorithms. Graphical lasso sets the regularization term to be an L_1-penalty on the entries of Ω, and the coordinate-descent approach of Friedman et al. (2008) [38] is one of the many algorithms proposed in the literature to solve the maximization problem. See the introduction of Bickel and Levina (2008B) [13] for a more comprehensive discussion.

Bayesian methods have also been developed, with different choices for the prior distribution to induce sparsity in the precision matrix Ω. Wang (2012) [74] proposed the Bayesian version of the graphical lasso by considering the Laplace prior for the off-diagonal entries and the exponential prior for the diagonal entries. This prior leads to a posterior mode of Ω coinciding with the graphical lasso estimate. However, this prior for the off-diagonal entries does not induce any sparsity in the posterior graphical structure, for which reason, Banerjee and Ghosal (2015) [5] modified it by inserting an atomic component at zero. This choice, though, leads to greater difficulties in terms of computations, which are overcome by employing the Laplace approximation.

As an alternative to these entry-wise priors, one can consider priors on the entire matrix space. After Dawid and Lauritzen (1993) [21] introduced the hyper inverse Wishart prior for Σ, Roverato (2002) [61] generalized it and defined the conjugate family of priors, named G-Wishart prior, for precision matrices Ω having G as underlying graphical structure. Since

then, the use of the G-Wishart prior has become common in this kind of settings and a lot of different techniques, such as Monte Carlo integration (Atay-Kayis and Massam, 2005 [3]) and block Gibbs sampler (Lenkoski and Dobra, 2011 [48]) among others, have been considered to simplify computations and sample from this distribution. See Wang and Li (2012) [75] for a more comprehensive review of the methods. Recently Liu and Martin (2019) [49] proposed an empirical G-Wishart prior, where the prior center hyper-parameter is estimated from the data. This in turn, allows them to make an effective use of the Laplace approximation to compute the normalization constant of the G-Wishart distribution so that computations are much faster.

Our approach is developed within the sparsity framework described earlier and the estimation of the graphical structure relies on the negligibility theory presented in Chapter 4. To go around the computational difficulty similar to that encountered by other methods mentioned in the previous paragraph, we use the Laplace approximation to approximate the joint probability of the data and the graph structure. To check the reliability of this approximation, we compare it with a Monte Carlo estimate, which we obtain following a very similar strategy of Atay-Kayis and Massam (2005) [3]. We then design a simple Metropolis-Hastings algorithm to sample from the conditional distribution of the graph given the data and estimate the median probability model, i.e., the graphical model comprising of those edges having a conditional probability of being non-negligible, larger than one half. We also discuss how the estimation of sparsity parameters can be carried out in this special context. We conclude by presenting some results which illustrate how our method works on both simulated and real data.

7.2 Model assumptions

Let $X = [x_1, \ldots, x_n]$ be a collection of n independent vectors, identically distributed as the random vector $x = [X_1, \ldots, X_p]'$ having distribution $N_p(0, \sigma^2\Omega^{-1})$. Here σ is an arbitrary scale parameter multiplying the random vectors $\tilde{x}_i \sim N_p(0, \Omega^{-1})$. Both the dimension p and the sample size n can be large, but we still assume $p \leq n$. The log likelihood for the precision matrix Ω is

$$-\tfrac{np}{2}\log 2\pi + \tfrac{n}{2}\left(\log\det\Omega - p\log\sigma^2\right) - \tfrac{n}{2\sigma^2}\mathrm{tr}(S\Omega)$$

where $S = \frac{1}{n}XX'$ is the sample covariance matrix.

To induce sparsity, we suppose that the off-diagonal elements of Ω, $\{\omega_{ij} : i < j,\ i, j \in [p]\}$ are independent and identically distributed with distribution P_ν, a univariate scale sparse measure with first order pair (ρ, H), where H is the α inverse-power exceedance measure. For the diagonal elements $\{\omega_{jj} : j \in [p]\}$, we suppose that they are independent and identically distributed with continuous distribution $P(dx) = p(x)dx$, and they are independent of the off-diagonals.

As mentioned in the previous section, we are interested in understanding the relationships among the p variables $X_1, \ldots, X_p$. This can be achieved by assuming an underlying graph structure and estimating the non-negligible conditional dependencies. So consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\} = [p]$ is the set of vertices and $E \subset \{(i, j) \in [p] \times [p] : i < j\}$ is the set of undirected edges, and let ϵ_ν be a negligibility sequence for P_ν according to Definition 4.5.1, defining the negligibility of the partial correlations $\{\omega_{ij}\}_{i<j}$. For each subset $\Gamma \subset \{(i, j) \in [p] \times [p] : i \leq j\}$, we want to compute the probability of the event

$$A_{\Gamma,\epsilon_\nu} = \{|\omega_{ij}| > \epsilon_\nu \text{ for } (i, j) \in \Gamma, |\omega_{ij}| \leq \epsilon_\nu \text{ for } (i, j) \notin \Gamma\}$$

given the observed value of X. Note that, besides the non-negligible off-diagonal indices (i,j), $i<j$ with $|\omega_{ij}|>\epsilon_\nu$, Γ will always include the diagonal indices (i,j), $i=j$. So the number of elements of Γ can be written as $p+s$, where $s\in[p(p-1)/2]$ denotes the number of edges corresponding to non-negligible partial correlations.

For a fixed set Γ, the conditional probability of A_{Γ,ϵ_ν} given X is proportional to

$$\begin{aligned}\mathbb{P}(X,A_{\Gamma,\epsilon_\nu}) &= \mathbb{P}(X\mid A_{\Gamma,\epsilon_\nu})\cdot\mathbb{P}(A_{\Gamma,\epsilon_\nu})\\ &= \int \mathbb{P}(X,\Omega\mid A_{\Gamma,\epsilon_\nu})d\Omega\cdot\mathbb{P}(A_{\Gamma,\epsilon_\nu})\\ &= \int_{\mathcal{P}_{\Gamma,\epsilon_\nu}} (2\pi)^{-\frac{np}{2}}(\det(\Omega/\sigma^2))^{\frac{n}{2}}e^{-\frac{n}{2}\,\mathrm{tr}(S\Omega/\sigma^2)}\,\mathbb{P}(d\Omega\mid A_{\Gamma,\epsilon_\nu})\cdot\mathbb{P}(A_{\Gamma,\epsilon_\nu})\,,\end{aligned} \tag{7.1}$$

where $\mathcal{P}_{\Gamma,\epsilon_\nu}$ is the set of symmetric positive definite $p\times p$ matrices having $|w_{ij}|>\epsilon_\nu$ for all $i<j$, $(i,j)\in\Gamma$ and $|w_{ij}|\le\epsilon_\nu$ for all $i<j$, $(i,j)\notin\Gamma$.

Following the negligibility theory developed for univariate sparse measures, for bounded and continuous functions of $(\omega_{ij})_{i,j}$, integrating against

$$\mathbb{P}(d\Omega\mid A_{\Gamma,\epsilon_\nu}) = \prod_{\substack{(i,j)\in\Gamma\\ i<j}} P_\nu(d\omega_{ij}\mid|\omega_{ij}|>\epsilon_\nu)\prod_{\substack{(i,j)\notin\Gamma\\ i<j}} P_\nu(d\omega_{ij}\mid|\omega_{ij}|\le\epsilon_\nu)\prod_{\substack{(i,j)\in\Gamma\\ i=j}} P(d\omega_{ij})$$

is, up to an error of order $\rho\epsilon_\nu^{-\alpha}$, asymptotically equivalent to integrating against

$$F(d\Omega) = \prod_{\substack{(i,j)\in\Gamma\\ i<j}} \tilde{H}(d\omega_{ij})\prod_{\substack{(i,j)\notin\Gamma\\ i<j}} \delta_0(d\omega_{ij})\prod_{\substack{(i,j)\in\Gamma\\ i=j}} P(d\omega_{ij})\,,$$

where $\tilde{H}(dx)=\tilde{h}(x)dx$ is the normalized exponentially-tilted exceedance measure. Notice that, now that we are integrating against F, the domain of integration can be set to be $\mathcal{P}_\Gamma$, the cone of symmetric positive definite $p\times p$ matrices having $w_{ij}=0$ for all $(i,j)\notin\Gamma$ and $w_{ij}\neq 0$ for all $(i,j)\in\Gamma$. A matrix in $\mathcal{P}_\Gamma$ is guaranteed to have exactly $p+2s$ non-zero

entries. Thus, the sparse approximation to the integral appearing in (7.1) is

$$
\begin{aligned}
&(2\pi)^{-\frac{np}{2}} \int_{\mathcal{P}_{\Gamma,\epsilon_\nu}} (\det(\Omega/\sigma^2))^{\frac{n}{2}} e^{-\frac{n}{2}\,\mathrm{tr}(S\Omega/\sigma^2)}\, F(d\Omega) = \\
&(2\pi)^{-\frac{np}{2}} \int_{\mathcal{P}_{\Gamma}} (\det(\Omega/\sigma^2))^{\frac{n}{2}} e^{-\frac{n}{2}\,\mathrm{tr}(S\Omega/\sigma^2)} \prod_{\substack{(i,j)\in\Gamma \\ i<j}} \tilde{H}(d\omega_{ij}) \prod_{\substack{(i,j)\in\Gamma \\ i=j}} P(d\omega_{ij}) = \\
&(2\pi)^{-\frac{np}{2}} \int_{\mathcal{P}_{\Gamma}} e^{\frac{n}{2}\log\det(\Omega/\sigma^2) - \frac{n}{2}\,\mathrm{tr}(S\Omega/\sigma^2) + \sum_{(i,j)\in\Gamma}\log\tilde{h}(\omega_{ij}) + \sum_i \log p(\omega_{ii})} \prod_{(i,j)\in\Gamma} d\omega_{ij} \prod_i d\omega_{ii}\,,
\end{aligned}
$$

where the notation $(i,j)\in\Gamma$ is the short version of $(i,j)\in\Gamma, i<j$, since the other case $(i,j)\in\Gamma, i=j$ can simply be denoted by i.

On the other hand, under the independence and sparsity assumptions,

$$
\mathbb{P}(A_{\Gamma,\epsilon_\nu}) = (1-\rho\epsilon_\nu^{-\alpha})^{p(p-1)/2-s}(\rho\epsilon_\nu^{-\alpha})^s + o(\rho\epsilon_\nu^{-\alpha})\,.
$$

Hence, the sparse approximation to the conditional probability of A_{Γ,ϵ_ν} given X is proportional to

$$
(1-\rho\epsilon_\nu^{-\alpha})^{p(p-1)/2-s}(\rho\epsilon_\nu^{-\alpha})^s (2\pi)^{-\frac{np}{2}} \int_{\mathcal{P}_\Gamma} e^{\frac{n}{2}\psi(\Omega)}\, d\Omega\,, \tag{7.2}
$$

where

$$
\psi(\Omega) = \log\det(\Omega/\sigma^2) - \tfrac{1}{\sigma^2}\mathrm{tr}(S\Omega/\sigma^2) + \tfrac{2}{n}\sum_{(i,j)\in\Gamma}\log\tilde{h}(\omega_{ij}) + \tfrac{2}{n}\sum_i \log p(\omega_{ii})\,.
$$

To compute the integral appearing in (7.2), we will use the Laplace approximation. The following section is devoted to describe this integral approximation and compare it with a Monte Carlo integration method to verify its validity numerically. In the appendix we also provide some guarantee on the size of the error of the Laplace approximation when the sample size grows large.

7.3 Laplace Approximation

In this section we treat the sparsity parameters ρ, α, σ as known quantities. Moreover, to slightly simplify the notation, from now on, instead of referring to the event A_{Γ,ϵ_ν}, we will just denote by Γ the set of non-negligible edges. The integral that we want to compute using the Laplace approximation is

$$\int_{\mathcal{P}_\Gamma} e^{\frac{n}{2}\psi(\Omega)}\, d\Omega\,.$$

To this end, it is helpful to write it as

$$\mathcal{I} = \int_{\mathcal{P}_\Gamma} e^{\log L(\Omega/\sigma^2)+\log F(\Omega)}\, d\Omega \;= \sigma^{2p} \int_{\mathcal{P}_\Gamma} e^{\log L(\Omega)+\log F(\sigma^2\Omega)}\, d\Omega\,, \tag{7.3}$$

where

$$\log L(\Omega) = \tfrac{n}{2}\log\det\Omega - \tfrac{n}{2}\mathrm{tr}(S\Omega)$$

is the log likelihood function while

$$\log F(\Omega) = \sum_{(i,j)\in\Gamma\,, i<j} \log\tilde{h}(\omega_{ij}) + \sum_{(i,j)\in\Gamma\,, i=j} \log p(\omega_{ij})$$

is the log prior density of the entries of Ω given Γ. Here $p(\omega) = e^{-\omega}\mathbf{1}_{\omega>0}$, while $H(d\omega) = h(\omega)d\omega$ is the inverse-power exceedance measure with exponent α, defined on $\mathbb{R}\setminus\{0\}$. So,

$$\begin{aligned}
\log\tilde{h}(\omega) &= \log\frac{(1-e^{-\omega^2/2\epsilon_\nu^2})h(\omega)}{\int(1-e^{-\omega^2/2\epsilon_\nu^2})h(\omega)\,d\omega} \\
&= \log(k_\alpha(1-e^{-\omega^2/2\epsilon_\nu^2})|\omega|^{-\alpha-1}) \\
&= \log(k_\alpha) + \log(1-e^{-\omega^2/2\epsilon_\nu^2}) - (\alpha+1)\log|\omega|\,,
\end{aligned}$$

where $1/k_\alpha = \int(1-e^{-\omega^2/2\epsilon_\nu^2})|\omega|^{-\alpha-1}\,d\omega$. In Figure 7.1, we plot the functions $\tilde{h}(x)$, $\log\tilde{h}(x)$ and its first derivative $(\log\tilde{h}(x))'$, for $\alpha = 0.5, 1, 1.5$.

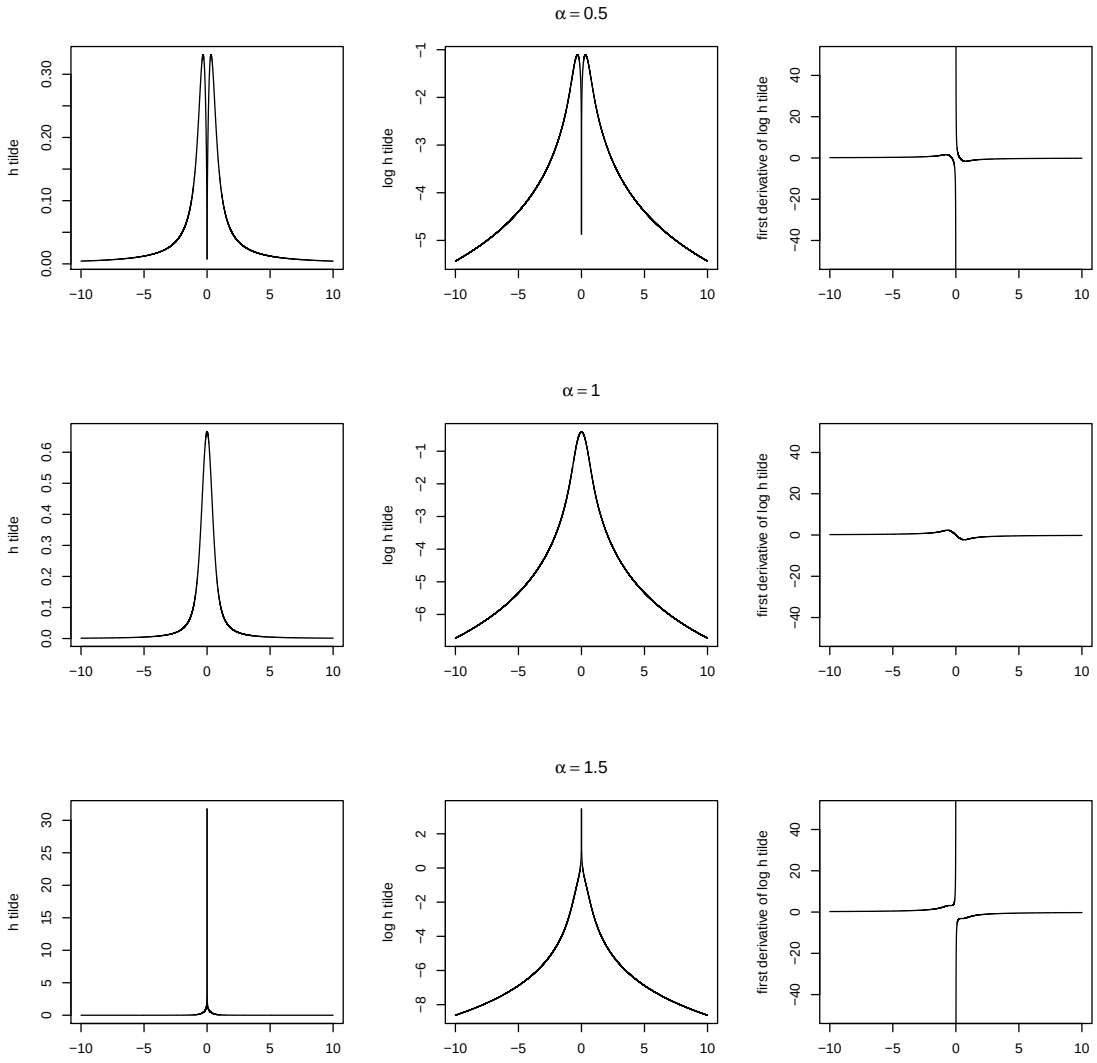


Figure 7.1: Plots of the functions $\tilde{h}(x)$, $\log \tilde{h}(x)$ and $(\log \tilde{h}(x))'$ for different values of α, $\alpha = 0.5, 1, 1.5$, while $\epsilon_\nu = 0.3$. The limit of $\tilde{h}(x)$ as $x \to 0$ is zero for $\alpha < 1$, finite and different from zero for $\alpha = 1$, and is infinite for $\alpha > 1$. However, $\tilde{h}(x)$ is defined on $\mathbb{R} \setminus \{0\}$ so that there is no trouble in taking the logarithm. The first derivative of $\log \tilde{h}(x)$ is finite for any $|x| > t$, for $t > 0$.

Let

$$\hat{\Omega} = \underset{\Omega \in \Theta_\Gamma}{\arg\max} \log L(\Omega) \tag{7.4}$$

denote the maximum likelihood estimator of Ω over the set

$$\Theta_\Gamma = \left\{ \Omega \in \mathcal{P}_\Gamma : \xi_n^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq \xi_n \right\},$$

where ξ_n is a deterministic sequence converging to infinity as $n \to \infty$.

Then the Laplace approximation for (7.3) can be written as

$$\sigma^{2p} e^{\log L(\hat{\Omega}) + \log F(\sigma^2 \hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))} \right)^{1/2}, \tag{7.5}$$

where $H(\Omega) = n\,(\Omega^{-1} \otimes \Omega^{-1})$ is two times the negative Hessian of the function $\log L(\Omega)$. In the appendix, we show that, under some mild assumptions, as $n \to \infty$, the relative error between the exact integral and the Laplace approximation goes to zero.

For finite n, we can assess the accuracy of the Laplace approximation by comparing it to the result obtained using a Monte Carlo integration technique. To obtain such an alternative for computing (7.3), we follow the procedure which Atay-Kayis and Massam (2005) [3] proposed to compute the G-Wishart normalizing constant. In the appendix, we report the main passages presented in that paper, with the necessary adjustments to our setting. For simplicity, we make this comparison fixing the scale parameter σ to be one.

The Monte Carlo method aims at expressing the integral as an expected value of a certain function g of some random variables which are relatively easy to simulate. In this context, the random variables are $U_{ii}^2 \sim \chi^2_{n+1+\nu_i}$, where $\nu_i = \#\{j > i : (i,j) \in \Gamma\}$ for all $i = 1, \ldots, p$ and $Z_{ij} \sim N(0,1)$ for all $i \neq j : (i,j) \in \Gamma$, all independent of each other. Then the integral in (7.3) can be written as

$$\mathcal{I} = C_{T,\delta,\Gamma}\ \mathbb{E}(g(U_{ii}, Z_{ij}))\,,$$

where $g(u,z)$ is as defined in (7.26) and $C_{T,\delta,\Gamma}$ is a constant dependent of the sample size $n = \delta - 1$, the set Γ and the Cholesky decomposition $(X'X)^{-1} = T'T$. The Monte Carlo technique estimates the expected value $\mathbb{E}(g(U_{ii}, Z_{ij}))$ by first generating the collection of random variables $\{U_{ii}^t, Z_{ij}^t\}$ for $t = 1, \ldots, N_{\text{sim}}$ simulations, and then computing the average

of $g(U_{ii}, Z_{ij})$ over the simulations

$$\hat{\mathbb{E}}(g(U_{ii}, Z_{ij})) = \frac{1}{N_{\text{sim}}} \sum_{t=1}^{N_{\text{sim}}} g(U_{ii}^t, Z_{ij}^t)\,.$$

Now, following the Laplace approximation method, we approximate $\log \mathcal{I}$ with

$$\log \mathcal{I}^L = \log L(\hat{\Omega}) + \log F(\hat{\Omega}) + \tfrac{p+s}{2} \log(2\pi) - \tfrac{1}{2} \log \det(H(\hat{\Omega}))\,,$$

while the Monte Carlo simulation method gives

$$\log \mathcal{I}^{MC} = \log C_{T,\delta,\Gamma} + \log \hat{\mathbb{E}}(g(U_{ii}, Z_{ij}))\,.$$

Figure 7.2 shows the relative error between the two methods, as a function of the sample size n, where each panel corresponds to one of the five scenarios considered for the true precision matrix Ω, described in detail in Section 7.6. We can see that the relative error is quite small even for $n = 100$, meaning that the Laplace approximation is reliable for relatively moderate values of n. Moreover, the error decreases as n increases in all scenarios. This should be confirming what we show in the appendix regarding the relative error between the Laplace approximation and the exact value of $\mathcal{I}$.

7.4 Metropolis-Hastings algorithm

We now describe how we design a Metropolis-Hastings algorithm in order to estimate the conditional distribution of the non-negligible set Γ given the observed matrix X. Again, we treat the sparsity parameters as known. Recall that the conditional probability that the set of non-negligible edges is Γ, is

$$\mathbb{P}(\Gamma \mid X) = c \cdot \mathbb{P}(X \mid \Gamma)\mathbb{P}(\Gamma)$$

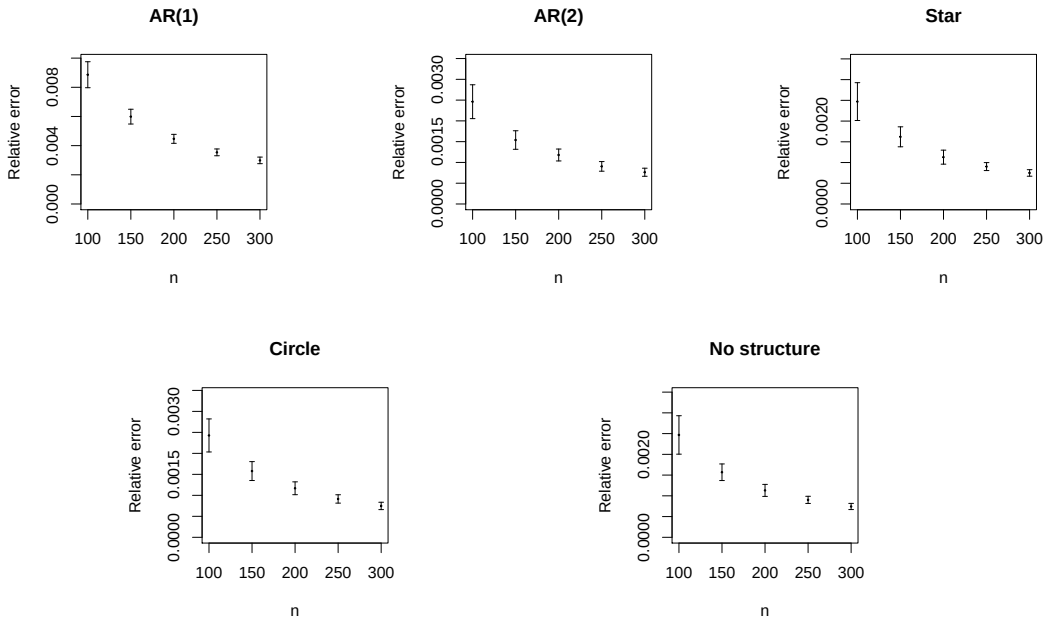


Figure 7.2: Relative error between the Laplace approximation and the Monte Carlo method for computing $\log \mathcal{I}$, as a function of the sample size n. For each sample size, we repeat the comparison 100 times and plot the average relative error $\pm$ one standard error. Each panel corresponds to a different model for generating Ω, as described in Section 7.6, with $p = 30$. The Monte Carlo computations are based on $N_{\text{sim}} = 1000$ simulations. The sparsity parameters are set to be $\rho = 0.07, \alpha = 1, \sigma = 1$.

where the normalizing constant c is just the reciprocal of

$$\mathbb{P}(X) = \sum \mathbb{P}(X \mid \Gamma)\mathbb{P}(\Gamma)\,.$$

Here the sum is taken over all possible graphs corresponding to precision matrices with the set of non-negligible partial correlations given by Γ, comprising of the p diagonal elements and $s \in [p(p-1)/2]$ off-diagonal entries. Since all graphs include the diagonal entries, what changes from graph to graph is the set of non-negligible undirected edges. Thus, the sum ranges over the power set $2^{[p(p-1)/2]}$. Clearly, this computation is prohibitive even for very small values of p. To overcome this computational infeasibility, we design a Metropolis-Hastings (M-H) algorithm to construct a Markov chain having $\mathbb{P}(\Gamma \mid X)$ as stationary distribution.

Given the estimated conditional distribution for the non-negligible Γ set, we can also estimate for each edge its marginal inclusion probability, i.e., the conditional probability of that edge belonging to the set of non-negligible edges, given the data. Then, following a common practice in the literature (see for example, Liu and Martin, 2019 and Banerjee and Ghosal, 2015), we can select the median probability model Γ_M, by classifying as non-negligible those edges whose marginal inclusion probability is greater than 0.5, and as negligible otherwise.

Following is a schematic description of our M-H algorithm.

1. Generate an initial subset S_0 from the power set $2^{[p(p-1)/2]}$ by setting the size k_0 to be the rounded integer of $0.05 \cdot p(p-1)/2$ and then drawing at random, without replacement, k_0 elements from $\{(i,j)\}_{i<j}$. Set $\Gamma_0 = \{(i,j)\}_{i=j} \cup S_0$.

2. Let $\theta = (1 - k_0/(p(p-1)/2))^{\mathbf{1}_{k_0>1}}$. Generate $\xi \sim \text{Ber}(\theta)$.

   If $\xi = 1$, add a new element x uniformly drawn from $\{(i,j)\}_{i<j} \backslash S_0$: set $S_1 = S_0 \cup \{x\}$ and $\Gamma_1 = \{(i,j)\}_{i=j} \cup S_1$.

   If $\xi = 0$, remove a current element x uniformly drawn from S_0: set $S_1 = S_0 \setminus \{x\}$ and $\Gamma_1 = \{(i,j)\}_{i=j} \cup S_1$.

3. Accept Γ_1 as the new Γ_0 with probability

$$A = \min\left\{1, \frac{p_1 \cdot q_{10}}{p_0 \cdot q_{01}}\right\}$$

   where $p_i = \mathbb{P}(\Gamma_i \mid X)$, for $i = 0, 1$, is the conditional probability that Γ_i is the set of non-negligible edges while q_{ij} is the probability of proposing set S_j given we have set S_i, for $i, j = 0, 1$.

   Otherwise, keep the current Γ_0 as the new Γ_0.

4. Repeat steps 2-3 until the target probabilities converge.

With this proposal mechanism, we add or remove a new index from the current set with probability depending on its size: the bigger the set already is, i.e., $k_0/(p(p-1)/2)$ is large, the less likely is to add a new element and the more likely is to remove one. On the other hand, given the uniform distribution for the selection of the index to add or remove, the transition probability from S_i to S_j is

$$q_{ij} = \begin{cases} \mathbb{P}(\text{add})\mathbb{P}(X = x \mid \text{add}) = (1 - k_i/(p(p-1)/2))^{\mathbf{1}_{k_i>1}} \frac{1}{p(p-1)/2 - k_i} & \text{if } k_j = k_i + 1 \\ \mathbb{P}(\text{remove})\mathbb{P}(X = x \mid \text{remove}) = (k_i/(p(p-1)/2))^{\mathbf{1}_{k_i>1}} \frac{1}{k_i} & \text{if } k_j = k_i - 1 \end{cases}$$

where $k_l = \#S_l$ is the size of S_l. Therefore, the ratio $\frac{p_1 \cdot q_{10}}{p_0 \cdot q_{01}}$ determining the acceptance probability A, is indeed proportional to the ratio of the two conditional probabilities p_1/p_0. The advantage is that this ratio can be approximated by combining (7.2) and (7.5), and be computed as

$$(\rho\epsilon_\nu^{-\alpha})^{k_1-k_0}(1-\rho\epsilon_\nu^{-\alpha})^{k_0-k_1}(4\pi)^{(k_1-k_0)/2} \frac{e^{\log L(\hat{\Omega}_{\Gamma_1}) + \log F(\sigma^2 \hat{\Omega}_{\Gamma_1})} \det(H(\hat{\Omega}_{\Gamma_1}))^{-1/2}}{e^{\log L(\hat{\Omega}_{\Gamma_0}) + \log F(\sigma^2 \hat{\Omega}_{\Gamma_0})} \det(H(\hat{\Omega}_{\Gamma_0}))^{-1/2}} .$$

In the simulation study presented in Section 7.6, we use this Metropolis-Hastings Markov chain (MHMC) to estimate the median probability model for different underlying graphical structures.

7.5 Sparsity parameter estimation

In this section, we discuss how we can estimate the sparsity parameters in this context, which is made particularly challenging by the latent graphical structure. To overcome some

of the difficulties, we make use of the Expectation-Maximization (E-M) algorithm as well as the Self-Normalizing Importance Sampling (SNIS) integration technique.

7.5.1 E-M algorithm and SNIS integration

If the graphical structure of Ω, summarized by the non-negligible Γ set, was observable, then we could write the log likelihood for the sparsity parameters based on both X and Γ

$$l(\rho, \alpha, \sigma; X, \Gamma) = \log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma)\,.$$

However, Γ is not observed so $l(\rho, \alpha, \sigma; X, \Gamma)$ is not computable. One possibility to overcome this latency is to take expectation over $2^{[p(p-1)/2]}$ with respect to some measure q,

$$\mathbb{E}_q(l(\rho, \alpha, \sigma; X, \Gamma)) = \sum_{\Gamma \in 2^{[p(p-1)/2]}} \log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma) \cdot q(\Gamma)\,, \tag{7.6}$$

and then maximize (7.6) to obtain an estimate for ρ, σ and α, which will clearly depend on the measure q. The E-M algorithm chooses q to be the conditional distribution of Γ given X, which in turn depends on the parameters ρ, α and σ. So the E-M algorithm proceeds as follows:

1. set $\theta_0 = (\rho_0, \alpha_0, \sigma_0)$ to be the parameter in $q(\Gamma) = \mathbb{P}(\Gamma \mid X; \rho_0, \alpha_0, \sigma_0)$ and compute $\mathbb{E}_q(l(\rho, \alpha, \sigma; X, \Gamma))$ as a function of $\theta = (\rho, \sigma, \alpha)$;

2. maximize $\mathbb{E}_q(l(\rho, \alpha, \sigma; X, \Gamma))$ over $\theta = (\rho, \sigma, \alpha)$ and update θ_0 to be the solution $\hat{\theta}$;

3. repeat steps 1-2 until the two parameters $\theta_0, \hat{\theta}$ are close enough.

Now, from a practical point of view, choosing the measure q to be the conditional distri-

bution of Γ given X,

$$q(\Gamma) = \mathbb{P}(\Gamma \mid X; \rho_0, \alpha_0, \sigma_0) = \frac{\mathbb{P}(X, \Gamma; \rho_0, \alpha_0, \sigma_0)}{\mathbb{P}(X; \rho_0, \alpha_0, \sigma_0)} = c_0 q^U(\Gamma)\,,$$

is problematic insofar, as discussed earlier, computing $c_0^{-1} = \mathbb{P}(X; \rho_0, \alpha_0, \sigma_0)$ is prohibitive. Thus, once more, we need to find a way to avoid computing the normalizing constants. In this framework, we take advantage of the so called self-normalizing importance sampling (SNIS) integration technique. The idea behind this method is to sample from another probability measure $\tilde{q}$, which can itself be known up to some normalizing constant $\tilde{c}$, and then, get rid of all unknown constants by taking ratios. More precisely, letting $\tilde{q}$ be this second probability measure on $2^{[p(p-1)/2]}$, rewrite (7.6) as

$$\mathbb{E}_q(\log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma)) = \mathbb{E}_{\tilde{q}}\left(\log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma) \cdot \frac{q(\Gamma)}{\tilde{q}(\Gamma)}\right)\,,$$

where the subscript in $\mathbb{E}_g$ indicates the measure g, with respect to which the expectation is taken. Now imagine drawing N times from $\tilde{q}$ to obtain $\{\Gamma_i \sim \tilde{q}\}_{i=1}^N$. By strong law of large numbers, as $N \to \infty$, almost surely,

$$\frac{1}{N}\sum_{\Gamma_i \sim \tilde{q}} \log \mathbb{P}(X, \Gamma_i; \rho, \alpha, \sigma) \frac{c_0 q^U(\Gamma_i)}{\tilde{c}\,\tilde{q}^U(\Gamma_i)} \to \mathbb{E}_{\tilde{q}}\left(\log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma) \cdot \frac{q(\Gamma)}{\tilde{q}(\Gamma)}\right)\,,$$

where $\tilde{q}(\Gamma) = \tilde{c}\,\tilde{q}^U(\Gamma)$. Similarly, almost surely,

$$\frac{1}{N}\sum_{\Gamma_i \sim \tilde{q}} \frac{c_0 q^U(\Gamma_i)}{\tilde{c}\,\tilde{q}^U(\Gamma_i)} \to \mathbb{E}_{\tilde{q}}\left(\frac{c_0 q^U(\Gamma)}{\tilde{c}\,\tilde{q}^U(\Gamma)}\right) = \sum_{\Gamma} \frac{q(\Gamma)}{\tilde{q}(\Gamma)}\tilde{q}(\Gamma) = 1\,,$$

so that, taking the ratio of the two expressions, the two normalizing constants c_0 and $\tilde{c}$ cancel out and, as $N \to \infty$,

$$\frac{\frac{1}{N}\sum_{\Gamma_i \sim \tilde{q}} \log \mathbb{P}(X, \Gamma_i; \rho, \alpha, \sigma) \cdot \frac{q^U(\Gamma_i)}{\tilde{q}(\Gamma_i)}}{\frac{1}{N}\sum_{\Gamma_i \sim \tilde{q}} \frac{q^U(\Gamma_i)}{\tilde{q}(\Gamma_i)}} \to \mathbb{E}_q(\log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma))\,, \tag{7.7}$$

with probability one. In this way, knowing the two measures q and $\tilde{q}$ up to a constant is sufficient to estimate the expected value required in the E-M algorithm.

7.5.2 Computational aspects and choice of $\tilde{q}$

Before going further in the discussion of how we implement the SNIS integration within the E-M algorithm, we first review the quantities appearing in (7.7). First of all, recall that we can compute $\log \mathbb{P}(X, \Gamma; \rho, \alpha, \sigma)$ as

$$\log \mathbb{P}(X \mid \Gamma; \rho, \alpha, \sigma) + \log \mathbb{P}(\Gamma; \rho, \alpha)\,.$$

Since $X \mid \Omega \sim N(0, \sigma^2 \Omega^{-1})$, we have

$$\mathbb{P}(X \mid \Gamma; \rho, \alpha, \sigma) = \sigma^{2p} \int_{\mathcal{P}_\Gamma} e^{\log L(\Omega) + \log F_{\alpha,\rho}(\sigma^2 \Omega)}\, d\Omega\ , \tag{7.8}$$

where

$$\log L(\Omega) = -\tfrac{np}{2} \log(2\pi) + \tfrac{n}{2} \log\det(\Omega) - \tfrac{n}{2} \mathrm{tr}(\tfrac{1}{n} X X' \Omega)\,,$$

and

$$\log F_{\alpha,\rho}(\sigma^2 \Omega) = \sum_{(i,j)\in\Gamma,\, i<j} \log \tilde{h}(\sigma^2 \omega_{ij}; \alpha, \rho) + \sum_{(i,j)\in\Gamma,\, i=j} \log p(\sigma^2 \omega_{ij})\,.$$

On the other hand,

$$\mathbb{P}(\Gamma; \rho, \alpha) = (\rho \epsilon_\nu^{-\alpha})^s (1 - \rho \epsilon_\nu^{-\alpha})^{p(p-1)/2 - s}\,.$$

Now, we approximate the integral in (7.8) using the Laplace approximation

$$\sigma^{2p} L(\hat{\Omega})\, F_{\alpha,\rho}(\sigma^2 \hat{\Omega}) (4\pi)^{\frac{p+s}{2}} \det(H(\hat{\Omega}))^{-\frac{1}{2}}\,,$$

where

$$\hat{\Omega} = \arg\max_{\Omega \in \Theta_\Gamma} \log L(\Omega) = \arg\max_{\Omega \in \Theta_\Gamma} \tfrac{n}{2} \log\det(\Omega) - \tfrac{n}{2} \mathrm{tr}(S\Omega)\,.$$

Therefore, denoting by $\text{Joint}_\theta(\Gamma) = \mathbb{P}(X, \Gamma; \theta)$, for each set Γ, we can write

$$\log \text{Joint}_\theta(\Gamma) \approx \log f(\hat{\Omega}) + \log \text{Priors}_\theta(\Gamma)\,,$$

where

$$f(\hat{\Omega}) = L(\hat{\Omega})(4\pi)^{\frac{p+s}{2}} \det(H(\hat{\Omega}))^{-\frac{1}{2}}\,,$$

while

$$\text{Priors}_\theta(\Gamma) = \sigma^{2p} F_{\alpha,\rho}(\sigma^2\hat{\Omega})\,(\rho\epsilon_\nu^{-\alpha})^s(1-\rho\epsilon_\nu^{-\alpha})^{p(p-1)/2-s}\,.$$

Going back to the E-M algorithm, since $q^U(\Gamma) = \mathbb{P}(X, \Gamma; \theta_0)$, we can rewrite the two steps of the E-M algorithm as follows

1. compute the expected value $\mathbb{E}_q(\log \mathbb{P}(X, \Gamma; \theta))$ as

$$\sum_{\Gamma_i \sim \tilde{q}} \log \text{Joint}_\theta(\Gamma_i) \cdot R_{\theta_0}(\Gamma_i)\,,$$

   where

$$R_{\theta_0}(\Gamma_i) = \frac{\text{Joint}_{\theta_0}(\Gamma_i)/\tilde{q}^U(\Gamma_i)}{\sum_{\Gamma_i \sim \tilde{q}} \text{Joint}_{\theta_0}(\Gamma_i)/\tilde{q}^U(\Gamma_i)}\,;$$

2. solve the maximization problem

$$\max_\theta \sum_{\Gamma_i \sim \tilde{q}} \log \text{Joint}_\theta(\Gamma_i) \cdot R_{\theta_0}(\Gamma_i)\,.$$

Here we make two observations.

The first observation is that $f(\hat{\Omega})$ does not depend on any sparsity parameters. Therefore,

the solution to

$$\max_{\theta} \mathbb{E}_q(\log \mathbb{P}(X, \Gamma; \theta)) \approx \max_{\theta} \sum_{\Gamma_i \sim \tilde{q}} \log \text{Joint}_{\theta}(\Gamma_i) \cdot R_{\theta_0}(\Gamma_i)$$

is the same as the solution to

$$\max_{\theta} \sum_{\Gamma_i \sim \tilde{q}} \log \text{Priors}_{\theta}(\Gamma_i) \cdot R_{\theta_0}(\Gamma_i) \,.$$

The second observation concerns the choice of $\tilde{q}$. Indeed, we decided to draw from this distribution to overcome the difficulty of drawing from $q(\Gamma) = \mathbb{P}(\Gamma \mid X; \theta_0)$. In principle, we could draw directly from q after estimating it using the MHMC as described in Section 7.4. However, this estimation would need to be repeated at every iteration of the E-M algorithm corresponding to a different value of θ_0, and this would result in a very cumbersome procedure. Alternatively, we could choose $\tilde{q}$ to be a distribution from which sampling is very simple. For example, one could first draw the size of the set Γ_i from a binomial $Bin(p(p-1)/2, r)$, for some $r \in (0, 1)$, and then, given the size, draw that number of elements uniformly at random from $[p(p-1)/2]$, without replacement. Nevertheless, despite this being a viable option for sampling the sets Γ_i, it is not a good choice because the resulting measure on $2^{[p(p-1)/2]}$ is very different from the q measure we are trying to substitute. In fact, if for all the sets $\{\Gamma_i \sim \tilde{q}\}_i$, sampled from $\tilde{q}$, $\text{Joint}_{\theta_0}(\Gamma_i)$ is on a different scale of magnitude compared to $\tilde{q}^U(\Gamma_i)$, then

$$\log \text{Joint}_{\theta_0}(\Gamma_i) - \log \tilde{q}^U(\Gamma_i)$$

can be huge, in absolute value. This means that

$$\exp\{\log \text{Joint}_{\theta_0}(\Gamma_i) - \log \tilde{q}^U(\Gamma_i)\}$$

are almost all numerically zero. The all-zeroes case can be avoided by subtracting $M =$

$\max_{i=1,\ldots,N} \log(\text{Joint}_{\theta_0}(\Gamma_i)) - \log(\tilde{q}^U(\Gamma_i))$. Yet, even doing so, the expression

$$\exp\{\log \text{Joint}_{\theta_0}(\Gamma_i) - \log \tilde{q}^U(\Gamma_i) - M\}$$

is extremely polarized, giving roughly mass equal to one to the Γ set corresponding to M and zero mass to all the other sets sampled from $\tilde{q}$. This in turn implies that the expected value in the E-M algorithm gets estimated by the average of just a single value.

This extreme example highlights the necessity of choosing $\tilde{q}$ in a non automatic manner. Indeed, taking the cue from this extreme case, in order to have q^U and $\tilde{q}^U$ on the same magnitude order, suppose we choose $\tilde{q}$ to be the conditional distribution $\mathbb{P}(\Gamma \mid X; \tilde{\theta})$, for some fixed parameter $\tilde{\theta}$, so that

$$\log \text{Joint}_{\theta_0}(\Gamma_i) - \log \tilde{q}^U(\Gamma_i) = \log \text{Joint}_{\theta_0}(\Gamma_i) - \log \text{Joint}_{\tilde{\theta}}(\Gamma_i)\,. \tag{7.9}$$

Now the advantage of this choice is that, since $\log \text{Joint}_{\theta}(\Gamma_i)$ is approximated by

$$\log f(\hat{\Omega}) + \log \text{Priors}_{\theta}(\Gamma_i)\,,$$

it is immediate to see that (7.9) can be reduced to

$$\log \text{Priors}_{\theta_0}(\Gamma_i) - \log \text{Priors}_{\tilde{\theta}}(\Gamma_i)\,,$$

where we recall that, for any $\theta = (\rho, \alpha, \sigma)$,

$$\log \text{Priors}_{\theta}(\Gamma) = p \log \sigma^2 + \log F_{\alpha,\rho}(\sigma^2 \hat{\Omega}) + s \log(\rho \epsilon_\nu^{-\alpha}) + (p(p-1)/2 - s) \log(1 - \rho \epsilon_\nu^{-\alpha})\,.$$

Thus, besides simplifying the computations as we no longer have to compute $\log f(\hat{\Omega})$, this choice leads to quantities in (7.9) which are now on comparable magnitude scales. The ratio

measure with $\tilde{q} = \mathbb{P}(\Gamma \mid X; \tilde{\theta})$ can then be computed as

$$R_{\theta_0}(\Gamma_i) = \frac{\exp\{\log \text{Priors}_{\theta_0}(\Gamma_i) - \log \text{Priors}_{\tilde{\theta}}(\Gamma_i) - M\}}{\sum_{i=1}^{N} \exp\{\log \text{Priors}_{\theta_0}(\Gamma_i) - \log \text{Priors}_{\tilde{\theta}}(\Gamma_i) - M\}}\,, \tag{7.10}$$

where now $M = \max_{i=1,\ldots,N} \log \text{Priors}_{\theta_0}(\Gamma_i) - \log \text{Priors}_{\tilde{\theta}}(\Gamma_i)$. This measure is not degenerate at just one Γ set, rather is diffuse on the whole range of sets sampled from $\tilde{q}$. Clearly this choice for $\tilde{q}$ requires running a MHMC to estimate $\mathbb{P}(\Gamma \mid X; \tilde{\theta})$, prior to the actual estimation of the sparsity parameters, but this needs to be done just one time.

In Figure 7.3, we compare the ratio measure, denoted by R.binom, obtained using the binomial-uniform sampling scheme described above as $\tilde{q}$ (left panels) with the ratio measure, denoted by R.tilde, obtained using the conditional $\mathbb{P}(\Gamma \mid X; \tilde{\theta})$ as $\tilde{q}$ (right panels). In both scenarios, the red triangles depict the values of $\tilde{q}$ for the $N = 1000$ sampled sets $\Gamma_i \sim \tilde{q}$, while the black circles depict the corresponding ratio measures for those same sampled sets. The two plots in the bottom row are the zoomed-in versions of the plots in the top row. From these, we can better appreciate the difference between the two choices for $\tilde{q}$: with the binomial-uniform sampling, the ratio measure is zero everywhere but at one single set; with the conditional $\mathbb{P}(\Gamma \mid X; \tilde{\theta})$ estimated from an MHMC, the ratio measure gives non-zero mass to all sets which were sampled from $\tilde{q}$.

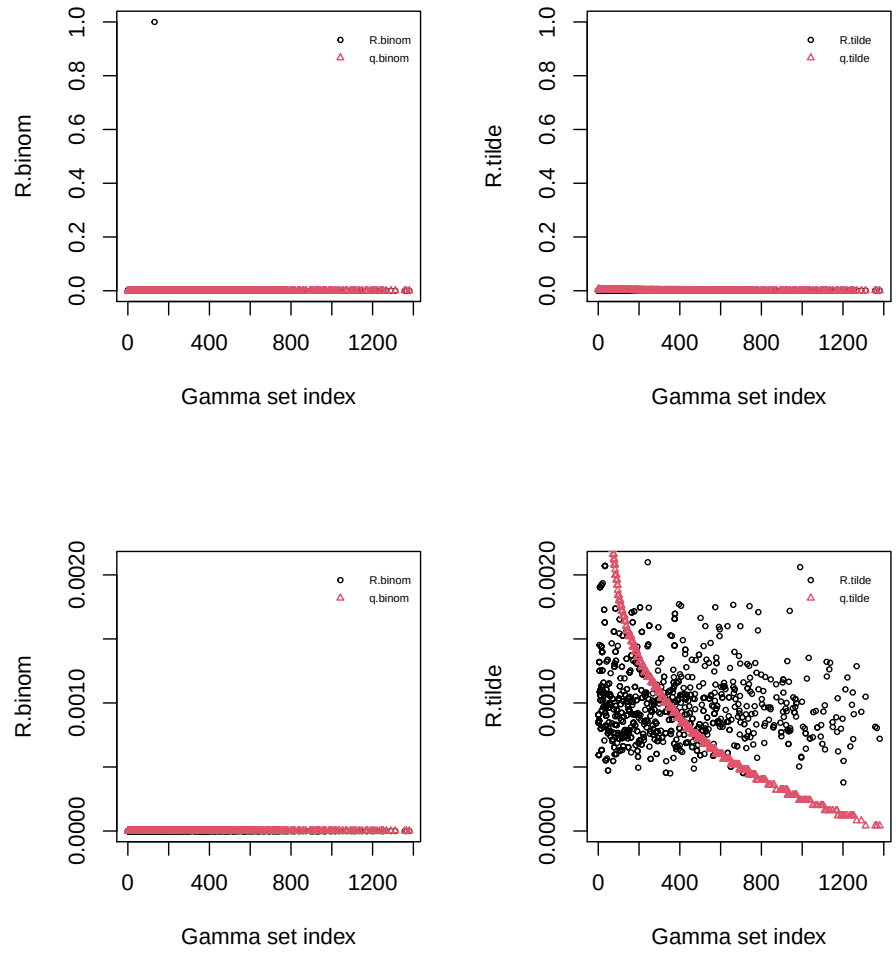


Figure 7.3: Comparison of two choices for $\tilde{q}$. Left panels depict the values of the $\tilde{q}$ measure (red triangles) and the corresponding ratio measure (black circles), when $\tilde{q}$ is the binomial-uniform sampling scheme on $2^{[p(p-1]/2}$. Right panels depict the values of the $\tilde{q}$ measure (red triangles) and the ratio measure (black circles), when $\tilde{q}$ is the conditional distribution $\mathbb{P}(\Gamma \mid X; \tilde{\theta})$, with $\tilde{\theta} = (0.07, 1, 1)$. $\mathbb{P}(\Gamma \mid X; \tilde{\theta})$ is estimated from an M-H Markov Chain with 30,000 total iterations, of which 5,000 are burn-in period.

7.5.3 Algorithm for estimating the sparsity parameters

To summarize, our final algorithm for estimating $\theta = (\rho, \alpha, \sigma)$ goes as follows:

1. fix $\tilde{\theta} = (\tilde{\rho}, \tilde{\alpha}, \tilde{\sigma})$ and run a MHMC to obtain $\tilde{q}(\Gamma) = \mathbb{P}(\Gamma \mid X; \tilde{\theta})$;

2. draw N samples $\Gamma_1, \ldots, \Gamma_N$ from $\tilde{q}(\Gamma)$ (which in fact means drawing from the M-H

Markov Chain relative frequencies);

3. fix $\theta_0 = (\rho_0, \alpha_0, \sigma_0)$, for all $i = 1, \dots, N$, compute $R_{\theta_0}(\Gamma_i)$ as in (7.10) and write

$$\eta(\theta, \theta_0) = \sum_{i=1}^{N} \log \text{Priors}_\theta(\Gamma_i) \cdot R_{\theta_0}(\Gamma_i)$$

as a function of $\theta = (\rho, \alpha, \sigma)$;

4. solve the maximization problem

$$\max_\theta \eta(\theta, \theta_0)\,;$$

5. update θ_0 to be the solution $\hat{\theta}$ and repeat steps 2 to 4, until θ_0 and $\hat{\theta}$ stabilize.

7.6 Simulation study

To illustrate how the sparse approximation described in previous paragraphs works, we perform a simulation study, with different true graph structures for Ω. To have a direct comparison with other methods previously proposed in the literature, we generate the data as in Banerjee and Ghosal (2015) [5] and Martin and Liu (2019) [49], considering five different models specified in terms of the entries of Σ or Ω as follows:

1. Model 1: AR(1) model where the entries of Σ are given by $\sigma_{ij} = 0.7^{|i-j|}$.

2. Model 2: AR(2) model where the entries of Ω are set to zero except $\omega_{ii} = 1$, $\omega_{i-1,i} = \omega_{i,i-1} = 0.5$ and $\omega_{i-2,i} = \omega_{i,i-2} = 0.25$, for all $i \in [p]$.

3. Model 3: Star model where the entries of Ω are set to zero except $\omega_{ii} = 1$, $\omega_{1,i} = \omega_{i,1} = 0.1$, for all $i \in [p]$.

4. Model 4: Circle model where the entries of Ω are set to zero except $\omega_{ii} = 2$, $\omega_{i-1,i} = \omega_{i,i-1} = 1$, for all $i \in [p]$ and $\omega_{1,p} = \omega_{p,1} = 0.9$.

5. Model 5: Sparse model with no special structure where $\Omega = \frac{1}{2}(B + B') + \tau I$, where B is a $p \times p$ matrix, in which all diagonals are set to zero while each off-diagonal entry is independently distributed like

$$\omega_{ij} \sim \begin{cases} 0.5 & \text{with probability } 0.05 \\ 1 & \text{with probability } 0.95 \end{cases} .$$

The parameter τ is chosen in such a way that the condition number of Ω is equal to p, and then Ω is standardized to have unit diagonals.

For each of these models, we generate $n = 100$ independent and identically distributed samples from $N_p(0, \Omega^{-1})$ with dimension $p = 30$ or $p = 50$. Then, we set $\tilde{\theta} = (0.07, 1, 1)$ when $p = 30$ and $\tilde{\theta} = (0.01, 1, 1)$ when $p = 50$, and follow the steps described in Section 7.5.3 to obtain some estimate of the sparsity parameters. To simplify the computations a little, we fix the α parameter to be one, and only estimate the sparsity rate ρ and the scale parameter σ. The results are reported in Table 7.1 for $p = 30$, and Table 7.2 for $p = 50$.

	$\hat{\rho} \times 10^2$	$\hat{\sigma}$
AR(1)	2.49	0.786
AR(2)	6.35	0.846
Star	2.10	1.001
Circle	2.43	0.864
No structure	5.11	0.882

Table 7.1: Estimated values for ρ and σ when $\alpha = 1$ and $p = 30$.

Given the values $\hat{\theta} = (\hat{\rho}, \hat{\alpha}, \hat{\sigma})$ for the sparsity parameters, we then run the MHMC in order to estimate the conditional distribution $\mathbb{P}(\Gamma \mid X; \hat{\theta})$. We set the total number of simulations to 24,000, with a burnin period of 4,000 iterations. At the end of the Markov chain, we select the median probability model Γ_M, by classifying as non-negligible those edges whose marginal inclusion probability is greater than 0.5. Here by marginal inclusion we mean the

	$\hat{\rho} \times 10^2$	$\hat{\sigma}$
AR(1)	1.43	0.746
AR(2)	3.75	0.859
Star	1.05	0.846
Circle	1.45	0.870
No structure	3.75	0.941

Table 7.2: Estimated values for ρ and σ when $\alpha = 1$ and $p = 50$.

event that the edge appears in the Γ set of non-negligible edges. Then, given the selected set of edges Γ_M, using the `glasso` function of the homonymous R package, we obtain the maximum likelihood estimate $\hat{\Omega}$ constrained to have Γ_M as graphical structure, and use this estimate of Ω to assess the performance of our method.

Following the literature, we compute three measures: specificity (SP), sensitivity (SE) and the Matthews Correlation Coefficient (MCC), named after the biochemist Brian W. Matthews, who introduced it in 1975. In formulae,

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\,,$$

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\,,$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}\,,$$

where TP, TN, FP, and FN denote, respectively, the number of true positives, true negatives, false positives and false negatives in the model considered.

We compare the performance of our sparsity method with the G-Wishart prior approach proposed by Liu and Martin (2019) [49] (L&M henceforth), which also makes use of a MHMC algorithm to estimate the conditional distribution of the graphical structure. For each model, we run the two different Metropolis-Hastings Markov chains 100 times and, in Table 7.4, we

display averages and standard errors of the three measures, SE, SP and MCC, the standard errors being multiplied by 10^3. Overall the two methods perform quite similarly in terms of MCC. Yet, the method based on the G-Wishart prior seems to always favor a very high level of sensitivity, whereas the sparsity approach is generally superior in terms of specificity.

	Sparsity			G-Wishart		
$p = 30$	SE	SP	MCC	SE	SP	MCC
AR(1)	0.997	**1.000**	0.985	**0.998**	**1.000**	**0.989**
	(0.262)	(0.000)	(1.267)	(0.247)	(0.000)	(1.208)
AR(2)	0.988	**0.875**	**0.885**	**0.990**	0.848	0.873
	(0.605)	(3.738)	(3.259)	(0.527)	(4.385)	(3.452)
Star	0.994	**0.369**	0.542	**0.999**	0.345	**0.564**
	(0.401)	(2.489)	(3.279)	(0.131)	(0.930)	(1.282)
Circle	**1.000**	**1.000**	**0.998**	0.999	**1.000**	0.996
	(0.111)	(0.000)	(0.543)	(0.177)	(0.000)	(0.853)
No structure	0.987	**0.862**	**0.872**	**0.992**	0.826	0.865
	(0.636)	(4.498)	(3.954)	(0.512)	(4.286)	(3.363)

Table 7.3: Averages and standard errors in parenthesis multiplied by 10^3, over 100 simulations.

	Sparsity			G-Wishart		
$p = 50$	SE	SP	MCC	SE	SP	MCC
AR(1)	0.998	**1.000**	0.984	**0.999**	**1.000**	**0.991**
	(0.145)	(0.000)	(1.151)	(0.097)	(0.000)	(0.791)
AR(2)	0.990	**0.850**	**0.863**	**0.994**	0.771	0.837
	(0.292)	(3.362)	(2.645)	(0.242)	(4.739)	(3.424)
Star	0.992	**0.381**	0.512	**0.999**	0.342	**0.571**
	(0.277)	(2.555)	(3.128)	(0.027)	(0.970)	(0.970)
Circle	**0.999**	**1.000**	**0.996**	**0.999**	**1.000**	0.995
	(0.068)	(0.000)	(0.558)	(0.0801)	(0.000)	(0.649)
No structure	0.984	**0.615**	**0.690**	**0.991**	0.551	0.674
	(0.342)	(3.412)	(3.146)	(0.292)	(4.059)	(3.368)

Table 7.4: Averages and standard errors in parenthesis multiplied by 10^3, over 100 simulations.

7.7 Gene regulatory network

In this section we apply our sparsity framework to a real data example which concerns gene regulatory network rewiring in patients having a specific type of breast cancer. This example was also analyzed by L&M, so we have a benchmark for comparison. The data was originally collected by the National Cancer Institute within The Cancer Genome Atlas (TCGA) Program, and is freely accessible through their website. However, the version of the data we use, the Agilent G450 microarray dataset, can be easily downloaded from the **DiffGraph** R package developed by Zhang et al. (2017) [80], who also provide a detailed description of the data set. As in L&M, we only consider luminal A subtype breast cancer, so the total number of patients is $n = 207$. For each patient, the p-long observed vector consists of standardized mRNA expression levels exhibited by 139 genes.

In this context, given the standardization of the mRNA expression levels, we set $\alpha = 1$ and $\sigma = 1$. And, because with real data is a little arbitrary to declare some connection to be totally absent, we take advantage of our sparsity-negligibility framework, which, per se, just makes statements on the negligibility of the connections. So we decide to choose different values for the parameter ρ, depending on the negligibility statements we would like to make about the correlations among the genes. In fact, the sequence of thresholds ϵ_ν defining the negligibility of a sparse random signal can be written as a function of the sparsity rate ρ, provided that $\rho H(\epsilon_\nu^+) = \rho\epsilon_\nu^{-\alpha} \to 0$ as $\nu \to 0$. So given $\epsilon_\nu = (\log(1/\rho_\nu))^{-1/2\alpha}$, we can choose the desired negligibility threshold to be $\epsilon_\nu = 0.33$ and set the sparsity rate to be $\rho = \exp\{-\epsilon_\nu^{-2\alpha}\}$ which is roughly $1.2 \cdot 10^{-4}$, so $\rho_\nu\epsilon_\nu^{-\alpha} = 3.6 \cdot 10^{-4}$.

With this choice of the negligibility threshold and the corresponding ρ, we run the M-H Markov chain to obtain samples from the conditional distribution of the gene regulatory network given the observed levels of expression. As before, we select the median probability model Γ_M by including an edge (i, j) between two genes if its marginal probability of being

non-negligible, i.e., of corresponding to $|\omega_{ij}| > 0.33$, given the data, is higher than 0.5. In Figure 7.4, we show the weighted graph corresponding to Γ_M, where the color intensity of each edge is proportional to its marginal probability of inclusion (which, by definition, is larger than one half). Besides the one-to-one relations, another important feature in gene network analysis is the number of connections that a gene has to other genes, which is called gene degree. In Figure 7.5, we show the unweighted graph where the different sizes and colors for the nodes reflect the different gene degrees: genes with degree greater or equal than ten are depicted in blue, those with degree between five and ten in yellow, while those with less than five connections are depicted in red. Those few genes that have largest degrees are also known as gene hubs, insofar they are important poles in the network wiring and, for this reason, are of great interest to clinicians. With our model, the top six genes with largest degree are the following: RPS6KB2 (S6K2), E2F1, AKT3, KIT, IGF1, and NCOA3. In Figure 7.6, we also plot the conditional distribution of the degree for these six genes, as derived from the estimated conditional distribution of the gene network.

All six hub genes identified by our model have strong association with breast cancer risk. Indeed, RPS6KB2 (S6K2) overexpression have proved to have prognostic and treatment predictive significance in breast cancer (Pèrez-Tenorio et al., 2011 [57]). Similarly, the E2F1 gene has been recently found by Hollern et. al (2019) [43] to be a master regulator of genes that coordinate tumor cell metastasis, while it is well known that high levels of expression of the IGF1 gene are positively associated with breast cancer (see for instance, Farabaugh et al., 2015 [34], and Monson et. al 2020 [53]). Likewise, proto-oncogene KIT (c-KIT) is frequently amplified in basal-like cancers (Nalwoga et al., 2008 [54]). As for the AKT isoforms, namely AKT1, AKT2, AKT3, these genes are known to regulate all stages of breast cancer, from initiation, prognosis, and metastasis to resistance to chemotherapy and improved hormonal therapy. We refer to Hinz and Jücker (2019) [42] for a comprehensive review on the AKT isoforms role in breast cancer.

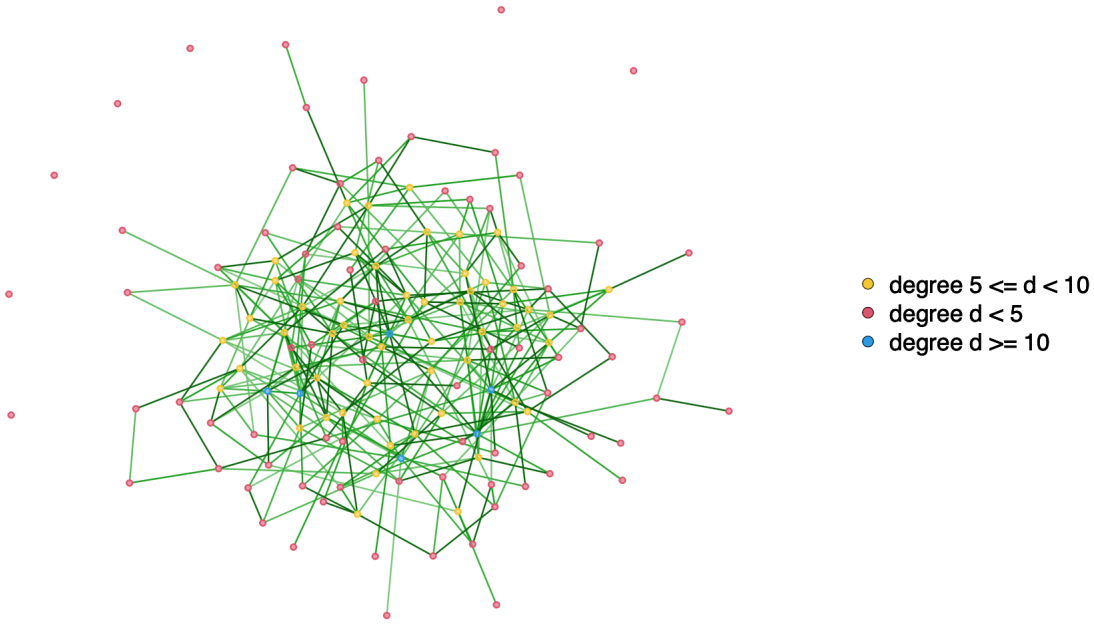


Figure 7.4: Gene regulatory network identified by the median probability model. The intensity of the color of the edges is proportional to its marginal probability of inclusion.

This selection overlaps with the top four hub genes found by L&M, where, instead of the KIT gene, they identified the EFGR gene. It is interesting to notice that both EGFR, Epidermal Growth Factor Receptor, and KIT, the gene encoding the receptor tyrosine kinase protein, are tyrosine kinase growth factor receptors (Nalwoga et al., 2008 [54]). Tyrosine kinase inhibition has become a common strategy in treatment of breast cancer, since cancer research assessed the relevance of the role that many protein kinases play during human tumorigenesis and cancer progression. L&M also identified NCOA3, whose degree is ranked sixth in the sparsity median model. Burwinkel et al. (2005) [15] found that a high percentage of primary human breast tumors shows elevated levels of expression for this gene, and overexpression of NCOA3 is correlated with worse survival rate.

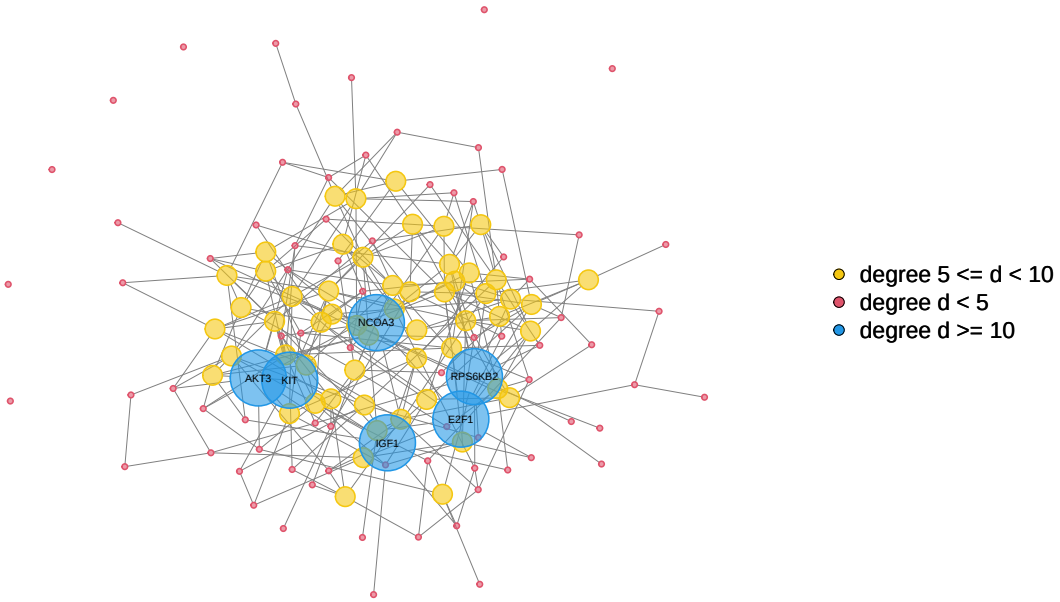


Figure 7.5: Gene regulatory network identified by the median probability model. Different sizes and colors for the nodes reflect the different gene degrees.

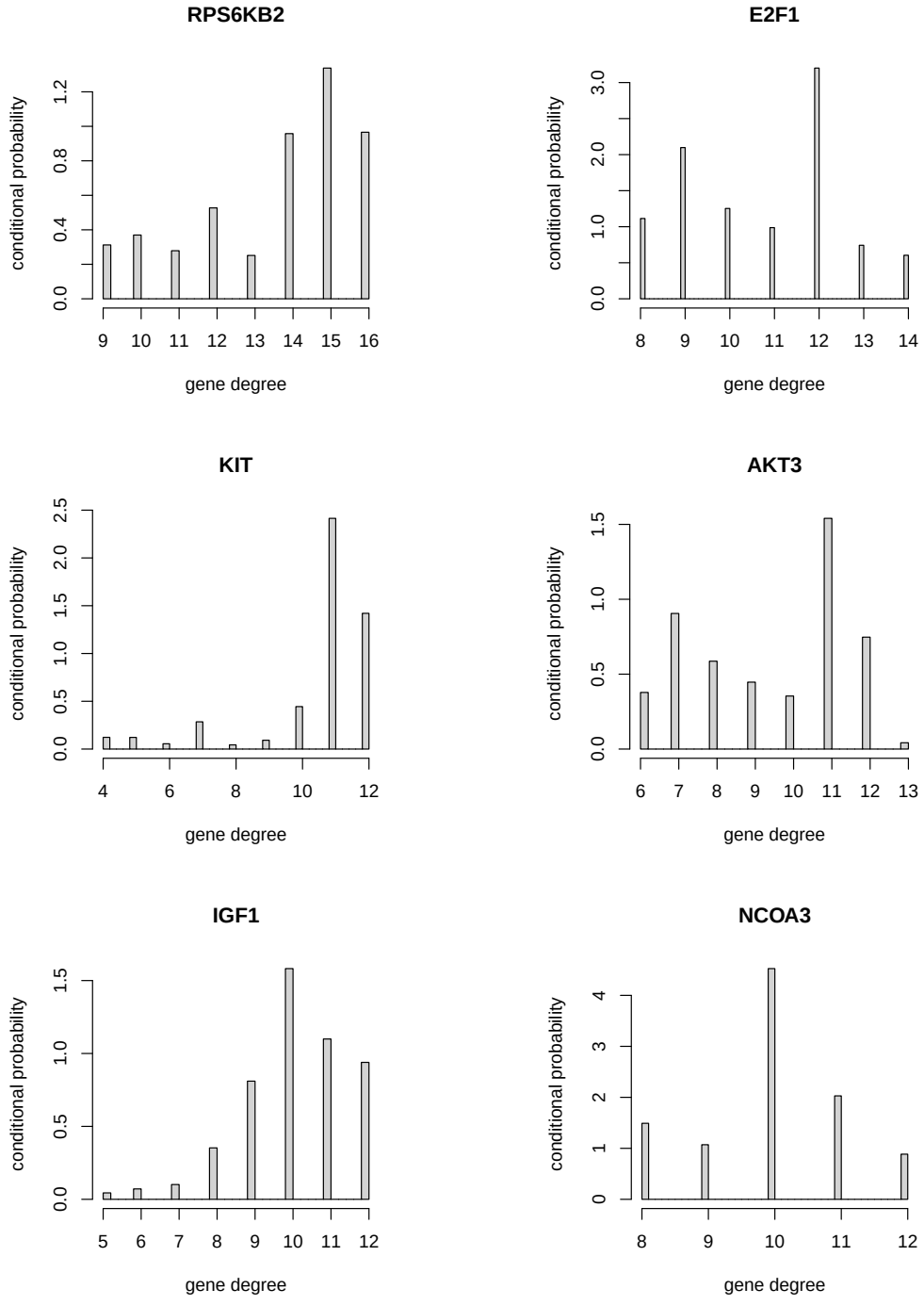


Figure 7.6: Gene degree conditional distribution for the top six hub genes: RPS6KB, E2F1, KIT, AKT3, IGF1 and NCOA3. These distributions are estimated from the M-H Markov chain used to sample from the conditional distribution of the gene network.

7.8 Appendix: Laplace approximation

7.8.1 Preliminaries

Recall that for any subset $\Gamma \subset \{(i,j) \in [p] \times [p] : i \leq j\}$, $\mathcal{P}_\Gamma$ denotes the cone of symmetric positive definite $p \times p$ matrices having $\omega_{ij} = 0$ for all $(i,j) \notin \Gamma$ and $\omega_{ij} \neq 0$ for $(i,j) \in \Gamma$. We want to compute

$$\mathcal{I} = \int_{\mathcal{P}_\Gamma} e^{\log L(\Omega) + \log F(\Omega)} \, d\Omega \,, \tag{7.11}$$

where

$$\log L(\Omega) = \tfrac{n}{2} \log\det \Omega - \tfrac{n}{2} \mathrm{tr}(S\Omega)$$

is the log likelihood function while

$$\log F(\Omega) = \sum_{(i,j)\in\Gamma, i<j} \log \tilde{h}(\omega_{ij}) + \sum_{(i,j)\in\Gamma, i=j} \log p(\omega_{ij})$$

is the log prior density of the entries of Ω given Γ. Here $p(\omega) = e^{-\omega}\mathbf{1}_{\omega>0}$, while $H(d\omega) = h(\omega)d\omega$ is the inverse-power exceedance measure with exponent α, defined on $\mathbb{R} \setminus \{0\}$. So,

$$\begin{aligned}
\log \tilde{h}(\omega) &= \log \frac{(1 - e^{-\omega^2/2\epsilon_\nu^2})h(\omega)}{\int (1 - e^{-\omega^2/2\epsilon_\nu^2})h(\omega)\, d\omega} \\
&= \log(k_\alpha (1 - e^{-\omega^2/2\epsilon_\nu^2})|\omega|^{-\alpha-1}) \\
&= \log(k_\alpha) + \log(1 - e^{-\omega^2/2\epsilon_\nu^2}) - (\alpha+1)\log|\omega| \,,
\end{aligned}$$

where $1/k_\alpha = \int (1 - e^{-\omega^2/2\epsilon_\nu^2})|\omega|^{-\alpha-1}\, d\omega$. In Figure 7.7, we plot the functions $\tilde{h}(x)$, $\log \tilde{h}(x)$ and its first derivative $(\log \tilde{h}(x))'$, for $\alpha = 0.5, 1, 1.5$.

We make the following two observations.

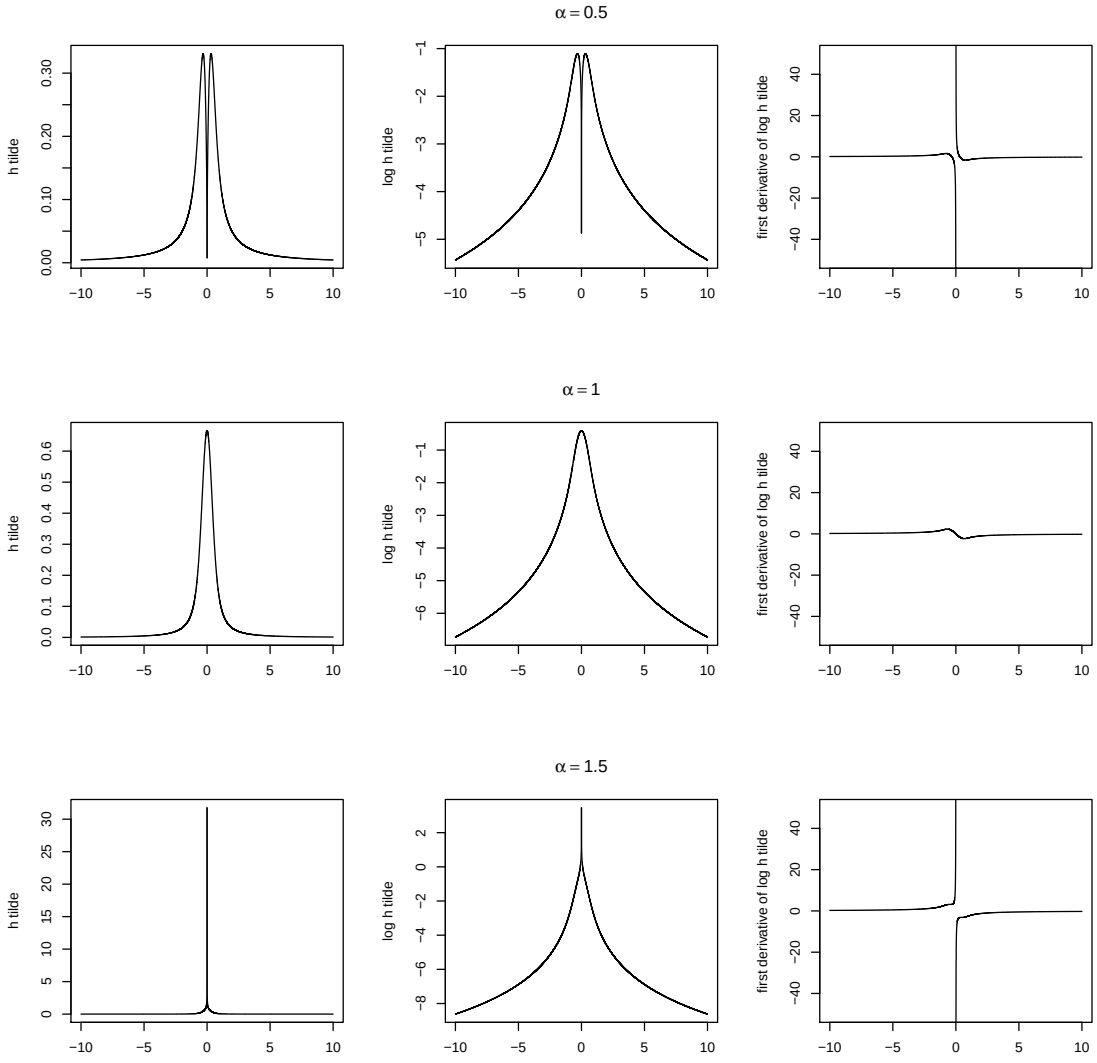


Figure 7.7: Plots of the functions $\tilde{h}(x)$, $\log \tilde{h}(x)$ and $(\log \tilde{h}(x))'$ for different values of α, $\alpha = 0.5, 1, 1.5$, while $\epsilon_\nu = 0.3$. The limit of $\tilde{h}(x)$ as $x \to 0$ is zero for $\alpha < 1$, finite and different from zero for $\alpha = 1$, and is infinite for $\alpha > 1$. However, $\tilde{h}(x)$ is defined on $\mathbb{R} \setminus \{0\}$ so that there is no trouble in taking the logarithm. The first derivative of $\log \tilde{h}(x)$ is finite for any $|x| > t$, for $t > 0$.

O1. The function $\log \tilde{h}(\omega)$ has derivative given by

$$g(\omega) = \frac{\partial}{\partial \omega} \log \tilde{h}(\omega) = \frac{\omega e^{-\omega^2/2\epsilon_\nu^2}}{\epsilon_\nu^2(1 - e^{-\omega^2/2\epsilon_\nu^2})} - \frac{\alpha + 1}{\omega} \,.$$

Now, the absolute value of g can be bounded by

$$\begin{aligned} |g(\omega)| &= \left| \frac{2}{\omega} \left(\frac{\omega^2 e^{-\omega^2/2\epsilon_\nu^2}}{2\epsilon_\nu^2(1 - e^{-\omega^2/2\epsilon_\nu^2})} - \frac{\alpha + 1}{2} \right) \right| \\ &\leq \left| \frac{2}{\omega} \right| \cdot \left(\frac{\omega^2 e^{-\omega^2/2\epsilon_\nu^2}}{2\epsilon_\nu^2(1 - e^{-\omega^2/2\epsilon_\nu^2})} + \frac{\alpha + 1}{2} \right) = b(\omega) \,. \end{aligned}$$

Both $|g(\omega)|$ and the majorating function $b(\omega)$ are symmetric around zero, so we can just look at the positive axis. On $(0, \infty)$, $b(\omega)$ is monotone decreasing, being the product of two non-negative functions that are decreasing. Therefore, given $t > 0$, $b(t) \geq b(\omega)$ for every $\omega \geq t$, so that, thanks to symmetry,

$$|g(\omega)| \leq b(t) \,,$$

for every ω such that $|\omega| \geq t$. This means that, for any $t > 0$, letting $\bar{B}_{t(0)} = \{\omega \in \mathbb{R} : |\omega| > t\}$, one has that

$$|\log \tilde{h}(\omega) - \log \tilde{h}(\omega')| \leq b(t)|\omega - \omega'| \,;$$

for all $\omega, \omega' \in \bar{B}_{t(0)}$.

Now, since we consider $p(\omega) = e^{-\omega}$, then we can bound the absolute value of the first derivative of $\log p(\omega)$ by one. So, for any given $t > 0$, let

$$F_1(t) = b(t) \vee 1 \,.$$

Then,

$$\begin{aligned}\left|\log F(\Omega) - \log F(\Omega')\right| &\leq \sum_{(i,j)\in\Gamma, i<j} \left|\log \tilde{h}(\omega_{ij}) - \log \tilde{h}(\omega'_{ij})\right| + \sum_{(i,j)\in\Gamma, i=j} \left|\log p(\omega_{ij}) - \log p(\omega'_{ij})\right| \\ &\leq \sum_{(i,j)\in\Gamma, i<j} b(t)|\omega_{ij} - \omega'_{ij}| + \sum_{(i,j)\in\Gamma, i<=j} |\omega_{ij} - \omega'_{ij}| \\ &\leq F_1(t) \sum_{(i,j)\in\Gamma} |\omega_{ij} - \omega'_{ij}|\,,\end{aligned}$$

provided $\omega_{ij}, \omega'_{ij} \in \bar{B}_{t(0)}$ for all $(i,j) \in \Gamma$.

O2. Given $t > 0$, for every ω in $\bar{B}_{t(0)} = \{\omega \in \mathbb{R} : |\omega| > t\}$, clearly

$$\sup_{|\omega|>t} \log \tilde{h}(\omega) - \log \tilde{h}(1) \geq \log \tilde{h}(\omega) - \log \tilde{h}(1)\,,$$

and similarly

$$\sup_{|\omega|>t} \log p(\omega) - \log p(1) \geq \log p(\omega) - \log p(1)\,.$$

So, let

$$F_2(t) = \left(\sup_{|\omega|>t} \log \tilde{h}(\omega) - \log \tilde{h}(1)\right) \vee \left(\sup_{|\omega|>t} \log p(\omega) - \log p(1)\right)\,,$$

and denote by $\mathbf{1}_{\mathrm{Mat}}$ the matrix that has (i,j)-entry equal to one if $(i,j) \in \Gamma$ and zero otherwise. Then

$$\begin{aligned}\log F(\Omega) - \log F(\mathbf{1}_{\mathrm{Mat}}) &= \sum_{(i,j)\in\Gamma, i<j} \left(\log \tilde{h}(\omega_{ij}) - \log \tilde{h}(1)\right) \quad + \sum_{(i,j)\in\Gamma, i=j} (\log p(\omega_{ij}) - \log p(1)) \\ &\leq \sum_{(i,j)\in\Gamma, i<j} F_2(t) \quad + \sum_{(i,j)\in\Gamma, i=j} F_2(t) \\ &\leq F_2(t)(p+2s)\,,\end{aligned}$$

provided $\omega_{ij} \in \bar{B}_{t(0)}$ for all $(i,j) \in \Gamma$.

Moreover, recall that

$$\hat{\Omega} = \underset{\Omega \in \Theta_\Gamma}{\arg\max} \log L(\Omega) \tag{7.12}$$

is the maximum likelihood estimator of Ω over the set

$$\Theta_\Gamma = \{\Omega \in \mathcal{P}_\Gamma : \xi_n^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq \xi_n\},$$

where ξ_n is a deterministic sequence converging to infinity as $n \to \infty$.

Also, in what follows, $\hat{t} > 0$ is the scalar for which $|\hat{\omega}_{ij}| > \hat{t}$ for all $(i, j) \in \Gamma$.

7.8.2 Assumptions

As $n \to \infty$, we assume

A1. $p \sim n^c$ for some $c \in (0, 1/3)$.

A2. $\xi_n \sim p^m$ where $m > 0$ such that

$$\xi_n^8((p+s)\log n)^{3/2} = o(\sqrt{n}).$$

For instance, if $s \sim p$, then $0 < m < (1-3c)/(16c)$ would suffice in order to have

$$\xi_n^8 \sqrt{\frac{(p+s)^3 \log^3 n}{n}} \to 0.$$

Moreover, since m needs be positive, then $c < 1/3$, so that $p \sim n^c = o(n^{1/3})$. This matches with Shun and McCullagh (1995), where it is suggested that the Laplace approximation for

high-dimensional integrals is reliable with no correction term, provided that the dimension of the integral is $o(n^{1/3})$.

A3. $\hat{\Omega}$ solution to

$$\max_{\Omega \in \Theta_\Gamma} \log L(\Omega)$$

satisfies

$$\sum_{ij} |1 - \hat{\omega}_{ij}| \leq \hat{k}_n \,,$$

where $\hat{k}_n$ such that, for any positive constant F_1,

$$\frac{e^{F_1 \hat{k}_n}}{n^{(p+s)/2+1}} = o\left(\frac{\xi_n^8 \zeta_n^3}{\sqrt{n}}\right) .$$

For instance, if $\hat{k}_n = \beta(p+s)\log n$ then $e^{F_1 \hat{k}_n} = n^{\beta F_1 (p+s)}$ so that the requirement

$$\frac{e^{F_1 \hat{k}_n}}{n^{(p+s)/2+1}} = \left(\frac{n^{\beta F_1}}{n^{1/2}}\right)^{(p+s)} \frac{1}{n} = o\left(\frac{\xi_n^8 \zeta_n^3}{\sqrt{n}}\right)$$

would be easily verified with $\beta \leq \frac{1}{2F_1}$.

7.8.3 Laplace approximation error

We want to show that, under assumptions A1. - A2. - A3., for n large enough, the integral in (7.11) can be sufficiently well approximated by the Laplace approximation

$$e^{\log L(\hat{\Omega}) + \log F(\hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} , \tag{7.13}$$

where $H(\Omega) = n\,(\Omega^{-1} \otimes \Omega^{-1})$ is two times the negative Hessian of $\log L(\Omega)$.

To this end, we follow the same strategy presented in Barber et al. (2016) [6], also used

by L&M. We split the integration domain $\mathcal{P}_\Gamma$ into two regions: a neighborhood of $\hat{\Omega}$

$$\mathcal{N} = \{\Omega \in \mathcal{P}_\Gamma : ||H(\hat{\Omega})^{1/2}\text{vec}(\Omega - \hat{\Omega})||_2 \leq \zeta_n\}$$

and its complement

$$\mathcal{N}^c = \{\Omega \in \mathcal{P}_\Gamma : ||H(\hat{\Omega})^{1/2}\text{vec}(\Omega - \hat{\Omega})||_2 > \zeta_n\}\,.$$

Here vec(A) stands for the vectorization of the matrix A, so if $A \in \mathbb{R}^{p\times p}$ then vec$(A) \in \mathbb{R}^{p^2}$. We can then write (7.11) as the sum of the following two integrals

$$\mathcal{I}_1 = \int_{\mathcal{N}} e^{\log L(\Omega)+\log F(\Omega)}\, d\Omega\,,$$

$$\mathcal{I}_2 = \int_{\mathcal{N}^c} e^{\log L(\Omega)+\log F(\Omega)}\, d\Omega\,.$$

(i) We start by approximating $\mathcal{I}_1$. Since this integral is computed over a neighborhood of $\hat{\Omega}$, we can approximate the log likelihood function with its Taylor expansion centered at $\hat{\Omega}$ and write

$$\log L(\Omega) = \log L(\hat{\Omega}) - \tfrac{1}{4}\Delta^T H(\hat{\Omega})\Delta + R_n^L(\Delta)\,,$$

where $\Delta = \text{vec}(\Omega - \hat{\Omega}) \in \mathbb{R}^{p^2}$. Yet Δ has only $p + 2s$ non-zero entries since both Ω and $\hat{\Omega}$ are supposed to be in $\mathcal{P}_\Gamma$ and only $p + s$ free entries since both Ω and $\hat{\Omega}$ are symmetric.

By Lemma 3 in L&M, since $\hat{\Omega}$ defined in (7.12) is such that

$$\xi_n^{-1} \leq \lambda_{\min}(\hat{\Omega}) \leq \lambda_{\max}(\hat{\Omega}) \leq \xi_n\,,$$

the approximation error in the Taylor series expansion of the log likelihood function can be bounded by

$$|R_n^L(\Delta)| \leq \tfrac{n}{2}(c_1\xi_n^5 \|\Delta\|_2^3 + c_2\xi_n^4 \|\Delta\|_2^4)\,, \tag{7.14}$$

where c_1 and c_2 are positive constants.

Let us now look at $\log F$. Consider $t \in (0, \hat{t}\,]$ arbitrarily small. Then, by O1.,

$$\left|\log F(\Omega) - \log F(\hat{\Omega})\right| \leq F_1(t) \sum_{(i,j)\in\Gamma} |\omega_{ij} - \hat{\omega}_{ij}|\,,$$

provided that $|\omega_{ij}| > t$ for all $(i,j) \in \Gamma$. So using the vectorized notation, we have

$$\left|\log F(\Omega) - \log F(\hat{\Omega})\right| \leq F_1(t) \|\Delta\|_1 \leq F_1(t)\sqrt{p+2s}\ \|\Delta\|_2\ , \tag{7.15}$$

where the last inequality follows from the known norm inequality $\|a\|_1 \leq \sqrt{n}\,\|a\|_2$ for any vector $a \in \mathbb{R}^n$.

Putting (7.14) and (7.15) together, we get that, for $\Omega \in \mathcal{N}$ and any arbitrary $t \in (0, \hat{t}\,]$, $\log L(\Omega) + \log F(\Omega)$ can be lower and upper bounded by

$$\log L(\hat{\Omega}) + \log F(\hat{\Omega}) - \tfrac{1}{4}\Delta^T H(\hat{\Omega})\Delta \pm \tfrac{n}{2}(c_1\xi_n^5 \|\Delta\|_2^3 + c_2\xi_n^4 \|\Delta\|_2^4) \pm F_1(t)\sqrt{p+2s}\ \|\Delta\|_2\ .$$

Now, on $\mathcal{N}$ we have $||H(\hat{\Omega})^{1/2}\Delta||_2 \leq \zeta_n$ so that

$$\begin{aligned}
\|\Delta\|_2 &\leq ||H(\hat{\Omega})^{-1/2}||_2\, ||H(\hat{\Omega})^{1/2}\Delta||_2 \\
&\leq n^{-1/2}||\hat{\Omega}^{-1} \otimes \hat{\Omega}^{-1}||_2^{-1/2}\, \zeta_n \\
&\leq n^{-1/2}||\hat{\Omega}||_2\, \zeta_n \\
&\leq \frac{\xi_n \zeta_n}{\sqrt{n}}\,,
\end{aligned}$$

where the last inequality follows from the fact that $\lambda_{\max}(\hat{\Omega}) \leq \xi_n$. Therefore, denoting by $\theta_n = \xi_n \zeta_n / \sqrt{n}$,

$$\tfrac{n}{2}(c_1 \xi_n^5 \|\Delta\|_2^3 + c_2 \xi_n^4 \|\Delta\|_2^4) + F_1(t) \sqrt{p+2s}\, \|\Delta\|_2 \leq \tfrac{n}{2}(c_1 \xi_n^5 \theta_n^3 + c_2 \xi_n^4 \theta_n^4) + F_1(t) \sqrt{p+2s}\, \theta_n$$

and so

$$\mathcal{I}_1 = e^{\log L(\hat{\Omega}) + \log F(\hat{\Omega})} \int_{\mathcal{N}} e^{-\frac{1}{4}\Delta^T H(\hat{\Omega}) \Delta}\, d\Delta \cdot e^{\pm\left(\frac{n}{2}(c_1 \xi_n^5 \theta_n^3 + c_2 \xi_n^4 \theta_n^4) + F_1(t)\sqrt{p+2s}\,\theta_n\right)},$$

where $a = b \cdot e^{\pm c}$ denotes $a \in [b \cdot e^{-c}, b \cdot e^{c}]$. By making the change of variable $\eta = H(\hat{\Omega})^{1/2} \Delta$, we have that the integral in the above expression is

$$\int_{\|\eta\| \leq \zeta_n} e^{-\frac{1}{4}\|\eta\|_2^2}\, d\eta \cdot \det(H(\hat{\Omega}))^{-1/2} \leq (4\pi)^{\frac{p+s}{2}} \det(H(\hat{\Omega}))^{-1/2} \cdot \Pr(\chi^2_{p+s} \leq \zeta_n^2)\,.$$

By Lemma A.1 in Barber et al. (2016), if $\zeta_n = \sqrt{5(p+s)\log n}$, then $\Pr(\chi^2_{p+s} \leq \zeta_n^2)$ can be bounded by $e^{\pm 1/\sqrt{n}}$ so that

$$\mathcal{I}_1 = e^{\log L(\hat{\Omega}) + \log F(\hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))} \right)^{1/2} \cdot e^{\pm\left(\frac{n}{2}(c_1 \xi_n^5 \theta_n^3 + c_2 \xi_n^4 \theta_n^4) + F_1(t)\sqrt{p+2s}\,\theta_n + 1/\sqrt{n}\right)}.$$

Let us examine the term in the exponential appearing in the RHS. Recalling that $\theta_n = \xi_n \zeta_n / \sqrt{n}$, we can rewrite it as

$$\frac{1}{2}\left(c_1 \xi_n^8 \frac{\zeta_n^3}{\sqrt{n}} + c_2 \xi_n^8 \frac{\zeta_n^4}{n} \right) + F_1(t) \sqrt{p+2s}\, \frac{\xi_n \zeta_n}{\sqrt{n}} + \frac{1}{\sqrt{n}}\,.$$

So choosing $\zeta_n = \sqrt{5(p+s)\log n}$, with the appropriate assumptions A1. and A2. on the growth of p and ξ_n, we have that

$$\xi_n^8 \sqrt{\frac{125(p+s)^3 \log^3 n}{n}} \to 0$$

as $n \to \infty$. Thus, for n large enough and t arbitrarily small, denoting by $F_1 = F_1(t)$, the term in the exponential is smaller than one. So we can use the fact that $e^{-x} \geq 1 - 2x$ and $e^x \leq 1 + 2x$ for all $0 \leq x \leq 1$, to find

$$\begin{aligned}\mathcal{I}_1 =& e^{\log L(\hat{\Omega}) + \log F(\hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))} \right)^{1/2} \cdot \\ & \cdot \left(1 \pm \xi_n^8 \sqrt{\frac{125(p+s)^3 \log^3 n}{n}} \left(c_1 + c_2 \sqrt{\frac{5(p+s)\log n}{n}} + 2F_1 + 2 \right) \right) .\end{aligned} \tag{7.16}$$

(ii) We now pass to approximate $\mathcal{I}_2$. As before we start analyzing $\log L(\Omega)$. As the integration region is now further away from $\hat{\Omega}$, the quadratic approximation of the log likelihood might not be accurate anymore. However, we can exploit the concavity of $\log L$ to bound the difference $\log L(\Omega) - \log L(\hat{\Omega})$. To this end, given $\Omega \in \mathcal{N}^c$ and $\Delta = \mathrm{vec}(\Omega - \hat{\Omega})$, define

$$\Omega' = \hat{\Omega} + \frac{\zeta_n}{||H(\hat{\Omega})^{1/2}\Delta||_2}(\Omega - \hat{\Omega}) .$$

It is easy to see that Ω' is on the boundary of $\mathcal{N}$ as $\Delta' = \mathrm{vec}(\Omega' - \hat{\Omega})$ is such that $||H(\hat{\Omega})^{1/2}\Delta'||_2 = \zeta_n$. Then by the concavity of $\log L$, we find that

$$\log L(\Omega') \geq \lambda \log L(\Omega) + (1 - \lambda) \log L(\hat{\Omega}) ,$$

where $\lambda = \zeta_n / ||H(\hat{\Omega})^{1/2}\Delta||_2$, so that

$$\log L(\Omega) - \log L(\hat{\Omega}) \leq \frac{1}{\lambda}(\log L(\Omega') - \log L(\hat{\Omega})) .$$

Using the bound on the Taylor expansion for $\log L(\Omega')$ derived in the previous part and the

fact that $\|\Delta'\|_2 \leq \theta_n = \xi_n \zeta_n / \sqrt{n}$, we get

$$\begin{aligned}
\log L(\Omega) - \log L(\hat{\Omega}) &\leq \frac{1}{\lambda}(-\tfrac{1}{4}\Delta'^T H(\hat{\Omega})\Delta' + \tfrac{n}{2}(c_1\xi_n^5 \|\Delta'\|_2^3 + c_2\xi_n^4 \|\Delta'\|_2^4)) \\
&\leq \frac{||H(\hat{\Omega})^{1/2}\Delta||_2}{\zeta_n}(-\tfrac{1}{4}\zeta_n^2 + \tfrac{n}{2}(c_1\xi_n^5\theta_n^3 + c_2\xi_n^4\theta_n^4)) \\
&\leq ||H(\hat{\Omega})^{1/2}\Delta||_2 \left(-\tfrac{1}{4}\zeta_n + \tfrac{1}{2}\xi_n^8\left(c_1\frac{\zeta_n^2}{\sqrt{n}} + c_2\frac{\zeta_n^3}{n}\right)\right) .
\end{aligned} \tag{7.17}$$

Let us now look at $\log F$. Since $\log F(\Omega)$ is not concave, we cannot utilize the same expedient. Yet, to find a bound on $\log F(\Omega)$, we first use O2. and consider the same $t \in (0, \hat{t}\,]$ as in the previous part. Then $F_2(t) > 0$ is such that

$$\log F(\Omega) - \log F(\mathbf{1}_{\mathrm{Mat}}) \leq F_2(t)(p + 2s)\,. \tag{7.18}$$

Moreover, given $t \in (0, \hat{t}\,]$, since $\hat{\omega}_{ij} \in \bar{B}_{t(0)}$ for all $(i,j) \in \Gamma$, using O1. we also have that

$$\left|\log F(\hat{\Omega}) - \log F(\mathbf{1}_{\mathrm{Mat}})\right| \leq F_1(t) \sum_{(i,j)\in\Gamma} |1 - \hat{\omega}_{ij}|\,. \tag{7.19}$$

Putting (7.18) and (7.19) together,

$$\begin{aligned}
\log F(\Omega) - \log F(\hat{\Omega}) &= \log F(\Omega) - \log F(\mathbf{1}_{\mathrm{Mat}}) + \log F(\mathbf{1}_{\mathrm{Mat}}) - \log F(\hat{\Omega}) \\
&\leq F_2(t)(p+2s) + \left|\log F(\mathbf{1}_{\mathrm{Mat}}) - \log F(\hat{\Omega})\right| \\
&\leq F_2(t)(p+2s) + F_1(t) \sum_{(i,j)\in\Gamma} |1 - \hat{\omega}_{ij}|\,.
\end{aligned}$$

Now, since by A3., we are furthermore assuming that $\sum_{i\leq j} |1 - \hat{\omega}_{ij}| \leq \hat{k}_n$, then we have

$$\log F(\Omega) \leq \log F(\hat{\Omega}) + F_2(t)(p+2s) + F_1(t)\hat{k}_n\,, \tag{7.20}$$

provided that $|\omega_{ij}| > t$ for all $(i,j) \in \Gamma$.

Combining (7.17) and (7.20),

$$\log L(\Omega) + \log F(\Omega) \leq \log L(\hat{\Omega}) + ||H(\hat{\Omega})^{1/2}\Delta||_2 \left(-\tfrac{1}{4}\zeta_n + \tfrac{1}{2}\xi_n^8 \left(c_1 \frac{\zeta_n^2}{\sqrt{n}} + c_2 \frac{\zeta_n^3}{n} \right) \right) + \log F(\hat{\Omega}) + F_1(t)\hat{k}_n + F_2(t)(p+2s)\,.$$

So we find that the integral $\mathcal{I}_2$ can be bounded as

$$\mathcal{I}_2 \leq e^{\log L(\hat{\Omega}) + \log F(\hat{\Omega})} e^{F_1(t)\hat{k}_n + F_2(t)(p+2s)} \int_{\mathcal{N}^c} e^{||H(\hat{\Omega})^{1/2}\Delta||_2 \left(-\frac{1}{4}\zeta_n(1-w_n)\right)}\, d\Delta\,, \tag{7.21}$$

where

$$w_n = 2\,\xi_n^8 \left(c_1 \frac{\zeta_n}{\sqrt{n}} + c_2 \frac{\zeta_n^2}{n} \right)\,.$$

The integral in the inequality above can be further bounded by changing variable to $\eta = H(\hat{\Omega})^{1/2}\Delta$ and applying Lemma A.2 in Barber et al. (2016)

$$\begin{aligned}
&\int_{\mathcal{N}^c} e^{\left\|H(\hat{\Omega})^{1/2}\Delta\right\|_2 \left(-\frac{1}{4}\zeta_n(1-w_n)\right)}\, d\Delta \\
&\leq \det(H(\hat{\Omega}))^{-1/2} \int_{\|\eta\|_2 > \zeta_n} e^{\|\eta\|_2 \left(-\frac{1}{4}\zeta_n(1-w_n)\right)}\, d\eta \\
&\leq \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))} \right)^{1/2} \frac{2^{-(p+s)/2}\zeta_n^{p+s-1}}{\Gamma\left(\frac{p+s}{2}\right)\zeta_n(1-w_n)/4} e^{-\frac{1}{4}\zeta_n^2(1-w_n)}\,,
\end{aligned}$$

where n is large enough so that $1 - w_n > 0$. Indeed, recalling that $\zeta_n = \sqrt{5(p+s)\log n}$, we have that

$$w_n = 2\,\xi_n^8 \left(c_1 \sqrt{\frac{(p+s)\log n}{n}} + c_2 \frac{(p+s)\log n}{n} \right)$$

converges to zero as $n \to \infty$. Therefore choosing n large enough, say to guarantee $1 - w_n \in (4/5, 1)$, with some further algebraic passages we can get that the last expression can be

bounded by

$$\begin{aligned}
&\left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{\frac{1}{2}} \frac{2^{-\frac{p+s}{2}+1}(5(p+s)\log n)^{\frac{p+s}{2}-1}}{\Gamma\left(\frac{p+s}{2}\right)(1-w_n)/4}\, e^{-\frac{5}{4}(p+s)\log n\cdot(1-w_n)} \\
&\leq \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{\left(\frac{p+s}{2}\right)^{\frac{p+s}{2}-1}}{\Gamma\left(\frac{p+s}{2}\right)} \cdot \frac{(5\log n)^{\frac{p+s}{2}-1}}{(1-w_n)/4} \cdot n^{-\frac{5(p+s)}{4}\cdot(1-w_n)} \\
&\leq \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{\left(\frac{p+s}{2}\right)^{\frac{p+s}{2}-1}}{\Gamma\left(\frac{p+s}{2}\right)} \cdot \frac{(5\log n)^{\frac{p+s}{2}-1}}{\frac{4}{5}\cdot\frac{1}{4}} \cdot n^{-\frac{5(p+s)}{4}\cdot\frac{4}{5}} \\
&\leq \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{e^{\frac{p+s}{2}}}{\sqrt{\pi(p+s)}} \cdot \frac{(5\log n)^{\frac{p+s}{2}-1}}{\frac{1}{5}} \cdot n^{-(p+s)}\,,
\end{aligned}$$

where the last inequality follows from Stirling's lower bound on the Gamma function. Going back to (7.21), incorporating these last passages, we find that

$$\begin{aligned}
\mathcal{I}_2 &\leq e^{\log L(\hat{\Omega})+\log F(\hat{\Omega})} e^{F_1(t)\hat{k}_n+F_2(t)(p+2s)} \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{e^{\frac{p+s}{2}}}{\sqrt{\pi(p+s)}}\, (5\log n)^{\frac{p+s}{2}-1} \cdot n^{-(p+s)} \\
&\leq e^{\log L(\hat{\Omega})+\log F(\hat{\Omega})} \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{e^{(4F_2(t)+1)\frac{p+s}{2}}}{\sqrt{\pi(p+s)}} \left(\frac{5\log n}{n}\right)^{\frac{p+s}{2}-1} e^{F_1(t)\hat{k}_n} \cdot n^{-\frac{(p+s)}{2}-1} \\
&\leq e^{\log L(\hat{\Omega})+\log F(\hat{\Omega})} \left(\frac{(2\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{e^{4F_2(t)+1}}{\sqrt{\pi(p+s)}} \left(\frac{5e^{2F_2(t)+1}\log n}{n}\right)^{\frac{p+s}{2}-1} \frac{e^{F_1(t)\hat{k}_n}}{n^{(p+s)/2+1}}\,.
\end{aligned}$$

If we assume that $\hat{k}_n = \beta(p+s)\log n$ with $\beta \leq 1/(2F_1(t))$ as specified in A3., then

$$\frac{e^{F_1(t)\hat{k}_n}}{n^{(p+s)/2+1}} = n^{F_1(t)\beta(p+s)-(p+s)/2-1} = o\left(\frac{\xi_n^8\zeta_n^3}{\sqrt{n}}\right)\,.$$

So, for sufficiently large n and arbitrarily small t, denoting by $F_2 = F_2(t)$, $\mathcal{I}_2$ can be very loosely bounded by

$$\mathcal{I}_2 \leq e^{\log L(\hat{\Omega})+\log F(\hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))}\right)^{1/2} \frac{e^{4F_2+1}}{\sqrt{\pi}} \frac{\xi_n^8\zeta_n^3}{\sqrt{n}}\,. \tag{7.22}$$

Combining the two bounds for $\mathcal{I}_1$ and $\mathcal{I}_2$ in (7.16) and (7.22), we can finally derive

$$\mathcal{I} = e^{\log L(\hat{\Omega})+\log F(\hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))} \right)^{1/2} \cdot \left(1 \pm \xi_n^8 \sqrt{\frac{125(p+s)^3 \log^3 n}{n}} \left(c_1 + c_2 \sqrt{\frac{5(p+s)\log n}{n}} + 2F_1 + 2 + \frac{e^{4F_2+1}}{\sqrt{\pi}} \right) \right) .$$

Since under our assumptions A1. and A2., as $n \to \infty$,

$$\xi_n^8 \sqrt{\frac{125(p+s)^3 \log^3 n}{n}} \to 0 \,,$$

we have that, as n gets large, the relative error between $\mathcal{I}$ and the Laplace approximation to $\mathcal{I}$,

$$\mathcal{I}^{\mathcal{L}} = e^{\log L(\hat{\Omega})+\log F(\hat{\Omega})} \left(\frac{(4\pi)^{p+s}}{\det(H(\hat{\Omega}))} \right)^{1/2} ,$$

becomes negligible.

7.8.4 Monte Carlo computation

In order to assess the accuracy of the Laplace approximation to

$$\mathcal{I} = \int_{\mathcal{P}_\Gamma} e^{\log L(\Omega)+\log F(\sigma^2\Omega)} \, d\Omega \,,$$

we compute this integral using a Monte Carlo integration technique. To this end, we follow the procedure proposed by Atay-Kayis and Massam (2005) [3] (A-K&M from now on) for computing the G-Wishart normalizing constant, making some small adjustments to our case. We report here the main passages appearing in that paper with the necessary modifications to our setting. For simplify the exposition, we fix the scale parameter σ to be one.

The integrand can be written as

$$e^{\log L(\Omega)+\log F(\Omega)} = \det(\Omega)^{\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}(X'X\Omega)+\log F(\Omega)}\,.$$

So letting $\delta = n+1$ and $D = X'X$, the integral in (7.11) can be rewritten as

$$\int_{\mathcal{P}_\Gamma} \det(\Omega)^{\frac{\delta-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(D\Omega)+\log F(\Omega)}\, d\Omega\,.$$

Now let $D^{-1} = T'T$ be the Choleski decomposition of $D^{-1} = (X'X)^{-1}$ and let $\Omega = \phi'\phi$ be the Choleski decomposition of Ω. If $\Omega \in \mathcal{P}_\Gamma$ then the entries of ϕ are such that, for $(i,j) \in \Gamma$

$$\phi_{ij} = \frac{\omega_{ij} - \sum_{k=1}^{i-1} \phi_{ki}\phi_{kj}}{\phi_{ii}}\,,$$

while for $(i,j) \notin \Gamma$,

$$\phi_{1k} = 0 \qquad \text{for } k = 2,\ldots,p\,,$$

$$\phi_{ij} = -\frac{\sum_{k=1}^{i-1} \phi_{ki}\phi_{kj}}{\phi_{ii}} \qquad \text{for } 1 < i \le j \le p\,.$$

These expressions show that the entries ϕ_{ij}, $(i,j) \notin \Gamma$, are functions of ϕ_{ij}, $(i,j) \in \Gamma$. So let ϕ_Γ be the projection of ϕ on to $\mathcal{P}_\Gamma^T$, the set of Γ-incomplete upper-triangular matrices with positive diagonal elements and completion ϕ. That is, ϕ_Γ has specified entries for $(i,j) \in \Gamma$ such that $\phi_{\Gamma,ij} = \phi_{ij}$ and has empty entries for $(i,j) \notin \Gamma$. Thus we can make the change of variable $\Omega \in \mathcal{P}_\Gamma \to \phi_\Gamma \in \mathcal{P}_\Gamma^T$. The determinant of this transformation, given in Roverato (2000), is

$$J_1 = 2^p \prod_i (\phi_{ii}^2)^{\nu_i+1}$$

where, for each $i = 1, \ldots, p$, $\nu_i = \#\{j > i : (i,j) \in \Gamma\}$. So we have

$$\begin{aligned}
\mathcal{I} &= \int_{\mathcal{P}_\Gamma} \det(\Omega)^{\frac{\delta-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(D\Omega)+\log F(\Omega)}\, d\Omega \\
&= \int_{\mathcal{P}_\Gamma^T} 2^p \prod_i (\phi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} e^{-\frac{1}{2}\mathrm{tr}((T'T)^{-1}\phi'\phi)+\log F_\phi(\phi)}\, d\phi_\Gamma \\
&= \int_{(\mathbb{R}^+)^p \times \mathbb{R}^s} 2^p \prod_i (\phi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} e^{-\frac{1}{2}\mathrm{tr}((T'T)^{-1}\phi'\phi)+\log F_\phi(\phi)} \prod_i d\phi_{ii} \prod_{i\neq j:(i,j)\in\Gamma} d\phi_{ij}\,.
\end{aligned} \tag{7.23}$$

Here the function F_ϕ is such that

$$\log F_\phi(\phi) = \log F(\Omega) = \sum_{i,j} f_\Gamma(\omega_{ij}) = \sum_{(i,j)\in\Gamma} f_\Gamma(\omega_{ij})$$

and it can be found by expressing $\{\omega_{ij}\}_{(i,j)\in\Gamma}$ as a function of $\{\phi_{ij}\}$. In fact, for all $(i,j) \in \Gamma$ $\omega_{ij} = \sum_{k=1}^i \phi_{ki}\phi_{kj}$, so that

$$\log F_\phi(\phi) = \sum_{(i,j)\in\Gamma} f_\Gamma\left(\sum_{k=1}^i \phi_{ki}\phi_{kj}\right).$$

Now the strategy is to make a second change of variable $\phi_\Gamma \in \mathcal{P}_\Gamma^T \to \psi_\Gamma \in \mathcal{P}_\Gamma^T$, where the completion of ψ_Γ is $\psi = \phi T^{-1}$. Then,

$$\psi_{rs} = \frac{1}{t_{ss}}\left(\phi_{rs} - \sum_{l=r}^{s-1} \psi_{rl} t_{ls}\right),$$

and the Jacobian of this second transformation, given in Lemma 3 of A-K&M, is

$$J_2 = \prod_i t_{ii}^{k_i+1},$$

where, for each $i = 1, \ldots, p$, $k_i = \#\{j < i : (i,j) \in \Gamma\}$. The integral in (7.23) can be

rewritten as

$$\begin{aligned}\mathcal{I} &= \int_{(\mathbb{R}^+)^p\times\mathbb{R}^s} 2^p \prod_i (\phi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} e^{-\frac{1}{2}\mathrm{tr}((T'T)^{-1}\phi'\phi)+\log F_\phi(\phi)} \prod_i d\phi_{ii} \prod_{i\neq j:(i,j)\in\Gamma} d\phi_{ij} \\ &= \prod_i t_{ii}^{k_i+1} \int_{(\mathbb{R}^+)^p\times\mathbb{R}^s} 2^p \prod_i (t_{ii}^2\psi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} e^{-\frac{1}{2}\mathrm{tr}(\psi'\psi)+\log F_\psi(\psi)} \prod_i d\psi_{ii} \prod_{i\neq j:(i,j)\in\Gamma} d\psi_{ij}\,.\end{aligned} \tag{7.24}$$

Similarly to before, the function F_ψ is such that

$$\log F_\psi(\psi) = \log F_\phi(\phi)$$

and it can be found by expressing $\{\phi_{ij}\}$ as a function of $\{\psi_{ij}\}$. In fact, as $\phi_{rs} = \sum_{l=r}^s \psi_{rl}t_{ls}$,

$$\log F_\psi(\psi) = \sum_{(i,j)\in\Gamma} f_\Gamma\left(\sum_{k=1}^i\left(\sum_{l=k}^i \psi_{rl}t_{ls}\sum_{l=k}^j \psi_{rl}t_{ls}\right)\right)\,.$$

Lemma 2 in A-K&M shows that it is possible to express ψ_{ij}, with $i\neq j : (i,j)\notin\Gamma$ as a function of $\{\psi_{ij}\}_{(i,j)\in\Gamma}$, so again, one can rewrite the integrand as a function of only $\{\psi_{ij}\}_{(i,j)\in\Gamma}$. To keep the notation lighter, from now on, when writing $\{\psi_{ij}\}_{i\neq j:(i,j)\notin\Gamma}$, we will implicitly refer to the expressions (31) and (32) appearing in A-K&M.

Since

$$-\tfrac{1}{2}\mathrm{tr}(\psi'\psi) = -\frac{1}{2}\left(\sum_i \psi_{ii}^2 + \sum_{i\neq j:(i,j)\in\Gamma}\psi_{ij}^2 + \sum_{i\neq j:(i,j)\notin\Gamma}\psi_{ij}^2\right)\,,$$

it is easy to express the integral in (7.24) as a constant times the expected value of a function of independent standard Normal and Chi-squared random variables. In fact,

$$\begin{aligned}&\prod_i (\psi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} e^{-\frac{1}{2}\mathrm{tr}(\psi'\psi)+\log F_\psi(\psi)} = \\ &\prod_i (\psi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} \exp\{-\tfrac{1}{2}\sum_i \psi_{ii}^2 - \tfrac{1}{2}\sum_{i\neq j:(i,j)\in\Gamma}\psi_{ij}^2\}\exp\{-\tfrac{1}{2}\sum_{i\neq j:(i,j)\notin\Gamma}\psi_{ij}^2 + \log F_\psi(\psi)\}\,.\end{aligned}$$

So the integrand in (7.24) can be written as

$$g(\psi_{ii}, \psi_{ij}) \prod_i (\psi_{ii}^2)^{\frac{\delta-1+\nu_i}{2}} e^{-\frac{1}{2}\psi_{ii}^2} \prod_{i\neq j:(i,j)\in\Gamma} e^{-\frac{1}{2}\psi_{ij}^2} \tag{7.25}$$

where

$$g(\psi_{ii}, \psi_{ij}) = \exp\{-\tfrac{1}{2} \sum_{i\neq j:(i,j)\notin\Gamma} \psi_{ij}^2 + \log F_\psi(\psi)\}\,. \tag{7.26}$$

Here we use (ψ_{ii}, ψ_{ij}) as a short notation for indicating $(\psi_{ii}, \psi_{i\neq j:(i,j)\in\Gamma})$.

From the expression in (7.25), one can see that the integral can be computed as the expected value of the function g of the random variables $\psi_{ii}^2 \sim \chi^2_{\delta+\nu_i}$ and $\psi_{ij} \sim N(0,1)$, all independent of each other. That is,

$$\mathcal{I} = C_{T,\delta,\Gamma}\ \mathbb{E}(g(U_{ii}, Z_{ij}))\,,$$

where

$$C_{T,\delta,\Gamma} = \prod_i t_{ii}^{\delta+\nu_i+k_i} (2\pi)^{\nu_i/2} \Gamma\left(\tfrac{\delta+\nu_i}{2}\right) 2^{(\delta+\nu_i)/2}$$

while $U_{ii}^2 \sim \chi^2_{\delta+\nu_i}$ for all $i = 1, \ldots, p$ and $Z_{ij} \sim N(0,1)$ for all $i \neq j : (i,j) \in \Gamma$, all independent of each other.

Part III

Vector sparsity

Chapter 8

Vector sparsity

8.1 Introduction

At the beginning of Chapter 4, we proposed an extension of the mathematical definition of sparsity to sequences of distributions defined on $\mathbb{R}^d$, $d > 1$. Given this general definition, in part II, we focused our attention on d-dimensional sparse measures which are product of d scalar sparse measures. Assuming this kind of multivariate sparsity, and combining it with the negligibility theory developed for univariate sparse measures, we looked at how some statistical problems can be formulated within the sparsity-negligibility framework.

In this part of the thesis instead, we study another kind of multivariate sparsity. We consider those d-dimensional measures which are rotationally invariant with respect to the inner product defining the metric on $\mathbb{R}^d$, and the sparsity of the vector is induced by the sparsity of its radius. We call such measures *vector-sparse measures*. In this context of rotational invariance, we introduce the d-dimensional $\cosh_d$ function, and study some asymptotic properties of the corresponding ζ_d function.

In this chapter, we derive the sparse approximations to some relevant functionals arising

in the signal-plus-noise model, assuming that the signal has a vector-sparse distribution. In particular, we derive the conditional distribution, given the observed vector, for both the signal direction and the signal magnitude. In the next chapter on the other hand, we study the Gaussian linear regression problem, and in that framework, the random vector assumed to have a vector-sparse distribution will be the coefficient vector.

8.2 Vector sparsity

We start by recalling the definition of multivariate sparsity given in Chapter 4.

Definition 8.2.1. A sequence of probability distributions $\{P_{\nu,d}\}_\nu$, defined on $(\mathbb{R}^d; \|\cdot\|)$, is said to have a sparse limit with rate ρ_ν if there exists a unit exceedance measure H_d such that

$$\lim_{\nu\to 0} \rho_\nu^{-1} \int_{\mathbb{R}^d} w(x)\, P_{\nu,d}(dx) = \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} w(x)\, H_d(dx),$$

for every function $w \in \mathcal{W}_d^{\#}$.

Component-wise sparsity assumes that the d components of the vector $X \sim P_{\nu,d}$, are independent, and usually identically distributed. If instead of the *i.i.d.* assumption, which requires the exchangeability of the components, but it is made with respect to a specific coordinate system given by the Cartesian axes, one is willing to treat any linear combination of the d components on equal footing, then one can assume the sparse measure $P_{\nu,d}$ on $\mathbb{R}^d$ to be invariant under the group of rotations and reflections. This means that, for all Borel sets B,

$$P_{\nu,d}(B) = P_{\nu,d}(OB)$$

for every transformation $O : \mathbb{R}^d \to \mathbb{R}^d$ which is orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$ chosen on $\mathbb{R}^d$, inducing the metric $d(x,y) = \|x-y\| = \langle x-y, x-y\rangle^{1/2}$. Since orthogonal transformations are isometries, they preserve the distance between any two points:

$d(x,y) = d(Ox,Oy)$, for every $x, y \in \mathbb{R}^d$. Therefore, given a Euclidean inner product defined by some positive definite matrix $A \in \mathbb{R}^{d\times d}$, $\langle x, y\rangle = x'AA'y$, then O must be such that $O'AA'O = AA'$ and $\det(O) = \pm 1$.

The probability of the set B, $P_{\nu,d}(B)$, does not change after rotating and/or reflecting B, for any arbitrary rotation and/or reflection, i.e., the probability measure $P_{\nu,d}$ on $(\mathbb{R}^d, \langle\cdot,\cdot\rangle)$ is invariant under the orthogonal group $\mathcal{O}(d)$, if and only if it is possible to factorize $P_{\nu,d}$ into two components:

$$\Gamma(d\,\tilde{x})P_\nu^R(d\,\|x\|)\,, \tag{8.1}$$

the uniform measure Γ for the direction vector $\tilde{x} = x/\|x\|$ on the unit sphere $\mathcal{S}^d = \{z : \|z\|^2 = 1\}$, and a radial measure P_ν^R for the radius of the vector $\|x\|$. See Theorem 2.5 in Fang, Kotz and Ng (1990) [33]. Therefore, in order to have $P_{\nu,d}$ converging to the Dirac delta measure at the origin, as $\nu \to 0$, it is necessary that the radial measure P_ν^R on $[0,\infty)$ converges to the Dirac delta measure at zero. This can be easily achieved by assuming that P_ν^R is two times the positive part of some sparse measure P_ν on $\mathbb{R}$.

With this geometric intuition in mind, we now give a more formal definition of a vector-sparse probability distribution.

Definition 8.2.2. Given an inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$, defining the norm $\|u\| = \sqrt{\langle u,u\rangle}$ and the corresponding unit sphere $\mathcal{S}^d = \{z : \|z\| = 1\}$, let $O : \mathbb{R}^d \to \mathbb{R}^d$ be an orthogonal operator such that $\langle Ou, Ov\rangle = \langle u, v\rangle$, for all $u, v \in \mathbb{R}^d$. If $P_{\nu,d}$ is a sequence of probability distributions defined on $(\mathbb{R}^d; \|\cdot\|)$, which is a sparse according to Definition 8.2.1, and is such that, for any Borel set B,

$$P_{\nu,d}(B) = P_{\nu,d}(OB)\,,$$

where $OB = \{z \in \mathbb{R}^d : O^{-1}z \in B\}$, then we say that $P_{\nu,d}$ is vector sparse.

Notice that if $P_{\nu,d}$ is vector sparse, i.e., it can be written as in (8.1), then also its exceedance measure H_d has the same structure, and can be factorized as

$$\Gamma(d\,\tilde{x})H^R(d\,\|x\|)\,, \tag{8.2}$$

where Γ is the uniform measure on $\mathcal{S}^d$, while H^R is the one dimensional exceedance measure on $\mathbb{R}_+ \setminus \{0\}$ corresponding to the radial sparse measure P^R_ν. If this latter is two times the positive part of any symmetric distribution P_ν, analogously H^R is two times the positive part of the symmetric exceedance measure H corresponding to P_ν. Any such exceedance measure H_d is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, so it gives zero mass to any proper subspace of $\mathbb{R}^d$.

We can easily check that H_d is a Lévy measure on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ by computing the integral

$$\begin{aligned}\int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\|x\|^2 \wedge 1)\, H_d(dx) &= \int_{\mathcal{S}^d} \Gamma(d\tilde{x}) \int_{(0,\infty)} (R^2 \wedge 1)\, H^R(dR) \\ &= \frac{1}{2}\int_{\mathbb{R}\setminus\{0\}} (R^2 \wedge 1)\, H(dR)\,,\end{aligned}$$

which is finite as long as H is a Lévy measure on $\mathbb{R} \setminus \{0\}$.

Before proceeding, we would like to conclude this section by highlighting that spherical symmetry, which we also call rotational invariance, is a statement on the probability measure as a function defined on the Euclidean vector space $(\mathbb{R}^d, \langle\cdot,\cdot\rangle)$, seen as an inner product space. The orthogonal group $\mathcal{O}(d)$ is the set of $d\times d$ orthogonal matrices, with group operation being the usual matrix multiplication $AB = \sum_i \sum_j A_{ij}B_{ij}$, where $A, B \in \mathbb{R}^{d\times d}$. When the group $\mathcal{O}(d)$ acts on $\mathbb{R}^d$, the function $f : \mathbb{R}^d \to [0,\infty)$ defined by $f(x) = \|x\|$, is maximal invariant, since $f(x) = f(Ox)$ for all $O \in \mathcal{O}(d)$, and $f(x_1) = f(x_2)$ implies that there is some group element O such that $x_1 = Ox_2$. The first requirement follows immediately. For the second

one instead, suppose that $\|x_1\| = \|x_2\|$. For any x, we can find O such that $Ox = \|x\|\, e_1$, where $e_1' = (1, 0, \ldots, 0)$. So let O_1, O_2 be such that $O_1 x_1 = \|x_1\|\, e_1 = \|x_2\|\, e_1 = O_2 x_2$. Then $x_1 = O_1^{-1} O_2 x_2$. The maximal invariance of $f(x) = \|x\|$ can be used to prove that a probability distribution is spherically symmetric if and only if its characteristic function, $\psi(t)$, can in fact be written as a function of only the norm of its argument. Indeed, $\psi(O't) = \psi(t)$ for all $O \in \mathcal{O}(d)$ means that $\psi(t)$ is an invariant function with respect to the group $\mathcal{O}(d)$, which in turn means that $\psi(t)$ needs to be a function of the maximal invariant $\|t\|$. For an introduction to the incredibly fascinating subject of group invariance and its applications in Statistics, we refer to the series of lectures given by Morris Eaton in 1987 at the University of Michigan, and gathered in Eaton (1989) [25].

8.2.1 Rotationally invariant inverse-power exceedance

If the inner product imposed on $\mathbb{R}^d$ is the standard Euclidean inner product $\langle u, v \rangle_2 = u'v = \sum_{i=1}^d u_i v_i$, then the orthogonal operator on $\mathbb{R}^d$ is such that $O'O = I_d$, where I_d is the identity matrix, while the unit sphere is $\mathcal{S}^d = \{z : z'z = 1\}$. Suppose that the radial exceedance measure H^R is the inverse-power measure

$$H^R(dx) = K_\alpha\, x^{-\alpha-1} dx\,,$$

where $K_\alpha = \frac{\alpha 2^{\alpha/2}}{\Gamma(1-\alpha/2)}$, and $\alpha \in (0, 2)$. Then

$$H_d(dx) = K_{d,\alpha}\, \|x\|_2^{-\alpha-d}\, dx\,, \tag{8.3}$$

is the rotationally invariant inverse-power exceedance measure, where

$$K_{d,\alpha} = K_\alpha / \text{Area}(\mathcal{S}^d) = \frac{\alpha 2^{\alpha/2}}{\Gamma(1-\alpha/2)} \frac{\Gamma(d/2)}{2\pi^{d/2}}$$

is the scalar such that H_d is a unit exceedance measure.

The measure in (8.3) is, up to a multiplicative constant, the Lévy measure μ associated with the rotationally invariant, symmetric α-stable ($S\alpha S$) process, whose characteristic function is the exponential of

$$\int_{\mathcal{S}^d}\int_0^\infty (e^{ir\theta' s}-1)\frac{dr}{r^{\alpha+1}}\frac{ds}{\text{Area}(\mathcal{S}^d)} = -\int_{\mathcal{S}^d}|\theta' s|^\alpha \frac{ds}{\text{Area}(\mathcal{S}^d)} = -c\,\|\theta\|^\alpha\,,$$

for some $c > 0$ (Samorodnitsky and Taqqu, 1994 [62]). In particular, when $\alpha = 1$, H_d is proportional to the Lévy measure of the multivariate Cauchy distribution

$$P_d(dx) = \frac{\Gamma(\frac{d+1}{2})}{\sqrt{\pi}\,\pi^{d/2}}\frac{dx}{(\|x\|_2^2+1)^{\frac{d+1}{2}}}\,.$$

In the one-dimensional case, the sparsity rate for the scaled Cauchy defining the rarity of threshold exceedance, $P_\nu(|X| > \epsilon) = \rho H(|X| > \epsilon)$, is given by $\rho = \sqrt{2/\pi}\,\nu$. When passing to the d-dimensional analog, the sparsity rate for the scaled version of P_d describes the rarity of norm threshold exceedance

$$P_{\nu,d}(\|X\| > \epsilon) = \rho_d H_d(\|X\| > \epsilon)\,.$$

So $\rho_d = \nu\,\frac{\sqrt{2}\,\Gamma(d/2+1/2)}{\Gamma(d/2)}$ can be found from

$$P_{\nu,d}(\|X\| > \epsilon) = \nu\,\frac{\Gamma(\frac{d+1}{2})}{\sqrt{\pi}\,\Gamma(d/2)}\int_\epsilon^\infty \frac{1}{\|x\|_2^2}\,d\,\|x\| + o(\nu) = \rho_d H_d(\|X\| > \epsilon) + o(\rho).$$

Therefore, the one-dimensional rate gets scaled by a factor of $\frac{\sqrt{\pi}\Gamma(d/2+1/2)}{\Gamma(d/2)}$, shown in Figure 8.1. As the dimension d becomes large, the scaling factor behaves like $\sqrt{\pi}(d/2)^{1/2}$.

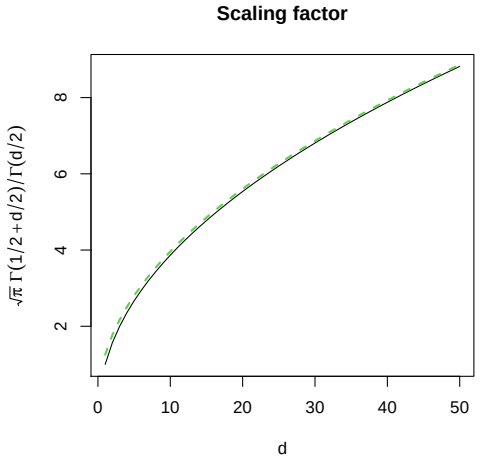


Figure 8.1: Dimension scaling effect on the one-dimensional sparsity rate. The dashed green line depicts the limiting behavior $\frac{\sqrt{\pi}\Gamma(d/2+1/2)}{\Gamma(d/2)} \sim \sqrt{\pi}(d/2)^{1/2}$ as $d \to \infty$.

8.2.2 More general rotationally-invariant exceedance

If instead of the standard Euclidean inner product, given any positive definite matrix $A \in \mathbb{R}^{d\times d}$, we consider its induced inner product $\langle u, v\rangle_A = u'AA'v$. Then the orthogonal operator O determining the rotational invariance is such that $u'O'AA'Ov = u'AA'v$, for all $u, v \in \mathbb{R}^d$, while the unit sphere is $\mathcal{S}_A^d = \{z : \|z\|_A^2 = z'AA'z = 1\}$.

Generalizing (8.3), we can define the A-rotationally invariant inverse-power exceedance measure on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ to be

$$H_{d,A}(dx) = K_{d,\alpha}\, \|x\|_A^{-\alpha-d} \det(A) dx\,. \tag{8.4}$$

Clearly, (8.4) coincides with (8.3) when A is the identity matrix. Otherwise, having the measure in (8.4) on x is equivalent to having the measure in (8.3) on the linearly transformed vector $\xi = A'x$.

Indeed, more generally, whenever $P_{\nu,d}$ is vector sparse on $(\mathbb{R}^d, \|\cdot\|_A)$, i.e., $P_{\nu,d}$ is rotation-

ally invariant with respect to $\langle \cdot, \cdot \rangle_A$, then its exceedance measure is

$$H_{d,A}(dx) = H_d(d(A'x)) \,,$$

where H_d is rotationally invariant with respect to $\langle \cdot, \cdot \rangle_2$.

8.3 Activity thresholds

From the normalization chosen in Definition 4.2.1, for defining a unitary exceedance measure, $\int (1 - e^{-\|x\|^2/2}) H_d(dx) = 1$, we can interpret the sparsity rate ρ as the unitary soft-threshold exceedance rate under $P_{\nu,d}$,

$$\int (1 - e^{-\|x\|^2/2}) P_{\nu,d}(dx) = \rho + o(\rho) \,.$$

Alternatively, for any given H_d, one can find the threshold $\epsilon_1 > 0$ for which the hard-threshold exceedance probability is one,

$$H_d(\epsilon_1^+) = \int_{\mathbb{R}^d} \chi_{\epsilon_1^+}(x) H_d(dx) = 1 \,,$$

where for any positive threshold $\epsilon > 0$, we denote $\epsilon^+ = \{\|x\| > \epsilon\}$. Then, calling such ϵ_1 the standard activity threshold, the sparsity rate ρ can be seen as the standard hard-threshold exceedance rate under $P_{\nu,d}$,

$$P_{\nu,d}(\epsilon_1^+) = \rho + o(\rho) \,.$$

For vector-sparse measures,

$$H_d(\epsilon_1^+) = \int_0^\infty \int_{\mathcal{S}^d} \chi_{\epsilon_1^+}(x) \frac{d\,\tilde{x}}{\text{Area}(\mathcal{S}^d)} \, H^R(d\,\|x\|) = \int_{\epsilon_1}^\infty H^R(d\,\|x\|) \,.$$

Thus, for instance, when H^R is the inverse-power measure,

$$\epsilon_1 = \frac{\sqrt{2}}{\Gamma(1-\alpha/2)^{1/\alpha}},$$

so the standard threshold is 0.94 for $\alpha = 0.5$, 0.8 for $\alpha = 1$, and 0.60 for $\alpha = 1.5$. Notice that these thresholds are independent of d, and this is because the exceedance event is in terms of the norm of the signal. If one prefers the standard activity thresholds to scale with the dimension, then one can express the exceedance event in terms of the root mean square (RMS) of the signal: $\tilde{\epsilon}^+ = \{\|x\|/\sqrt{d} > \tilde{\epsilon}\}$. Then clearly, $\tilde{\epsilon}_1$ such that $H_d(\tilde{\epsilon}_1^+) = 1$, is just given by $\tilde{\epsilon}_1 = \epsilon_1/\sqrt{d}$. Thus, the RMS-standard threshold scales the standard threshold by a factor of $1/\sqrt{d}$.

8.4 ζ_d and $\cosh_d$ functions

We now investigate the zeta function and zeta measure associated to rotationally invariant exceedance measures. In this setting, we introduce a d-dimensional analog of the ordinary cosh function, which appears inside the integral of the zeta function. This $\cosh_d$ function is defined to be the same exponential function as in the unidimensional case, but now it is uniformly averaged over the d-dimensional unit sphere. The derivation of some of the formulas and facts presented in this section, can be found in the appendix.

Recall that for any exceedance measure H_d defined on $(\mathbb{R}^d \setminus \{\mathbf{0}\}, \|\cdot\|)$, its zeta transform is

$$\zeta_d(y) = \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(\langle y, x\rangle) - 1)e^{-\|x\|^2/2}\, H_d(dx),$$

and its corresponding d-dimensional zeta measure is the integrand

$$\zeta_d(du; y) = (\cosh(\langle y, x\rangle) - 1)e^{-\|x\|^2/2} H_d(dx)\,.$$

For vector-sparse measures, the factorization of H_d into the uniform spectral measure on the unit sphere $\mathcal{S}^d = \{z : \|z\| = 1\}$, and the sparse radial measure on $(0, \infty)$, leads to

$$\zeta_d(y) = \int_0^\infty \int_{\mathcal{S}^d} (\cosh(\|x\| \langle y, \tilde{x} \rangle) - 1) e^{-\|x\|^2/2} \frac{d\,\tilde{x}}{\text{Area}(\mathcal{S}^d)}\, H^R(d\,\|x\|)\,.$$

This suggests defining the function $\cosh_d$ on $(\mathbb{R}^d, \|\cdot\|)$,

$$\cosh_d(y) = \int_{\mathcal{S}^d} e^{\langle y, \tilde{x} \rangle} \frac{d\tilde{x}}{\text{Area}(\mathcal{S}^d)}\,. \tag{8.5}$$

When $\langle y, x \rangle = y'x$ and $d = 1$, $\cosh_d(y)$ coincides with the usual $\cosh(y) = \frac{e^y + e^{-y}}{2}$. The $\cosh_d$ function is the exponential function $g(\tilde{x}) = e^{\langle y, \tilde{x} \rangle}$, uniformly averaged over the unit sphere. It is analytic with Taylor expansion

$$\cosh_d(y) = \sum_{r=0}^{\infty} \frac{\|y\|^{2r}}{(d/2)^{\uparrow r} 2^{2r} r!}\,, \tag{8.6}$$

where $\alpha^{\uparrow r}$ denotes the ascending factorial function $\alpha^{\uparrow r} = \Gamma(\alpha + r)/\Gamma(\alpha)$. For $d \geq 2$, the integrand in (8.5) is the kernel of the von Mises-Fisher density over the unit sphere with polar direction $y/\|y\|$ and concentration parameter $\|y\|$. So one can also write $\cosh_d$ as

$$\cosh_d(y) = \frac{(2\pi)^{d/2} I_{d/2-1}(\|y\|)}{\|y\|^{d/2-1}} \cdot \frac{1}{\text{Area}(\mathcal{S}^d)},$$

where I_ν denotes the modified Bessel function of the first kind of order ν.

Exploiting the symmetry of $\cosh_d$, the zeta transform of any rotationally invariant exceedance measure can be written as

$$\zeta_d(y) = \int_0^\infty (\cosh_d(\|x\|\, y) - 1)\, e^{-\|x\|^2/2}\, H^R(d\,\|x\|)\,. \tag{8.7}$$

In particular, if H^R is the inverse-power measure, then $\zeta_d(y)$ has Taylor expansion given by

$$\zeta_d(y) = \sum_{r=1}^{\infty} \frac{\|y\|^{2r}}{(d/2)^{\uparrow r} 2^r r!} \frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} \,. \tag{8.8}$$

Notice that, thanks to rotational invariance of H_d, not just $\cosh_d(y)$, but also $\zeta_d(y)$ only depends on the norm of the argument $y \in \mathbb{R}^d$. So from now on, we will either use the vector argument or the norm argument depending on the convenience of the context.

8.4.1 Limiting behavior of ζ_d and $\cosh_d$ for d large

In this section, we investigate the behavior of the functions $\cosh_d$ and ζ_d when the dimension d gets large. We restrict our analysis to the zeta function of the inverse-power radial measure. We consider vectors $y \in \mathbb{R}^d$ having unit root mean square, $\text{RMS} = \|y\| / \sqrt{d} = 1$ and compare the d-dimensional $\cosh_d$ and ζ_d to the one-dimensional versions, cosh and ζ respectively. In Figure 8.2, we show the scaling factor by which the one-dimensional functions are multiplied, when the dimension increases. We can see that for the zeta function, the impact of the dimension is larger when the inverse power α is smaller.

In Figure 8.3 we show the limiting behavior of $\cosh_d(\sqrt{d})$ and $\zeta_d(\sqrt{d})$ as the dimension goes to infinity. From the left panel, it appears quite clearly that, as $d \to \infty$, $\cosh_d(\sqrt{d})$ approaches $\sqrt{e}$, indicated by the dashed line, while looking at the right panel, we can see that, even if slowly, $\zeta_d(\sqrt{d}) \to 1$ as $d \to \infty$.

Indeed, exploiting the Taylor series in (8.6), one can write

$$\cosh_d(\sqrt{d}) = \sum_{r=0}^{\infty} \frac{(d/2)^r}{(d/2)^{\uparrow r} 2^r r!} = \sum_{r=0}^{\infty} f_d(r) \,,$$

where $f_d(r) = \frac{(d/2)^r}{(d/2)^{\uparrow r} 2^r r!}$ is a sequence, indexed by d, of non-negative functions such that, for

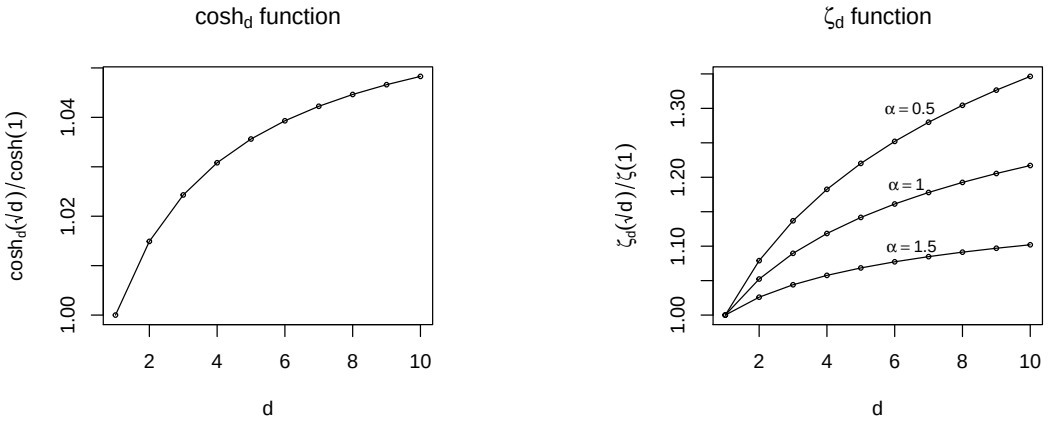


Figure 8.2: Left panel: dimension scaling effect on the one-dimensional cosh function: $\cosh_d(\sqrt{d})/\cosh(1)$. Right panel: dimension scaling effect on the one-dimensional ζ function: $\zeta_d(\sqrt{d})/\zeta(1)$, for $\alpha = 0.5, 1, 1.5$.

all $d \geq 1$, $f_d(r) \leq f_{d+1}(r)$ for all $r \geq 0$, as

$$\frac{f_d(r)}{f_{d+1}(r)} = \frac{d^r}{(d+1)^r} \frac{(d/2+1/2)^{\uparrow r}}{(d/2)^{\uparrow r}} \leq 1\,.$$

Since the sequence $f_d(r)$, as $d \to \infty$, has a limit

$$\lim_{d\to\infty} f_d(r) = \lim_{d\to\infty} \frac{(d/2)^r}{(d/2)^{\uparrow r} 2^r r!} = \frac{1}{2^r r!} = f(r)\,,$$

which is summable, $\sum_{r=0}^{\infty} f(r) = \sum_{r=0}^{\infty} \frac{(1/2)^r}{r!} = \sqrt{e}$, then by monotone convergence theorem,

$$\lim_{d\to\infty} \sum_{r=0}^{\infty} f_d(r) = \sum_{r=0}^{\infty} \lim_{d\to\infty} f_d(r)\,.$$

So $\lim_{d\to\infty} \cosh_d(\sqrt{d}) = \sqrt{e}$.

With a similar argument, one can show that

$$\lim_{d\to\infty} \zeta_d(\sqrt{d}) = 1\,.$$

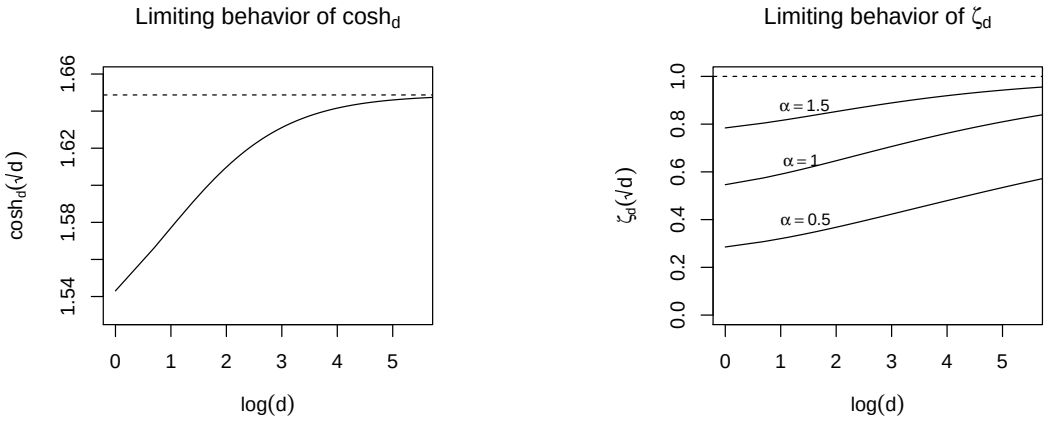


Figure 8.3: Left panel: limiting behavior of $\cosh_d(\sqrt{d})$ as $d \to \infty$. Right panel: limiting behavior of $\zeta_d(\sqrt{d})$ as $d \to \infty$, for $\alpha = 0.5, 1, 1.5$.

More generally, if we consider $\zeta_d(\sqrt{d}\,x)$, then we have

$$\lim_{d\to\infty} \zeta_d\left(\sqrt{d}\,x\right) = \begin{cases} 1-(1-x^2)^{\alpha/2} \text{ for } |x|<1 \\ 1 \qquad\qquad\qquad \text{for } |x|=1 \\ \infty \qquad\qquad\qquad \text{for } |x|>1\,. \end{cases}$$

See the appendix for all derivations. In Figure 8.4, we show $\zeta_d(\sqrt{d}\,x)$ slowly approaching the limit function $1-(1-x^2)^{\frac{\alpha}{2}}$ as d gets large, for α equal 0.5, 1, and 1.5.

8.4.2 Limiting behavior of ζ_d for $\|y\|$ large

It is also interesting to investigate the limiting behavior of $\zeta_d(\|y\|)$, with fixed dimension d, while $\|y\| \to \infty$. Once again, we study the case when H_d is as in (8.3). For simplicity of exposition, here we consider the standard Euclidean inner product, but the same holds for any Euclidean inner product. Now,

$$\begin{aligned} \zeta_d(y) &= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(y'x)-1)e^{-\|x\|^2/2}\frac{K_{d,\alpha}}{\|x\|^{\alpha+d}}\,dx \\ &= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (e^{y'x}-1)e^{-\|x\|^2/2}\frac{K_{d,\alpha}}{\|x\|^{\alpha+d}}\,dx\,. \end{aligned}$$

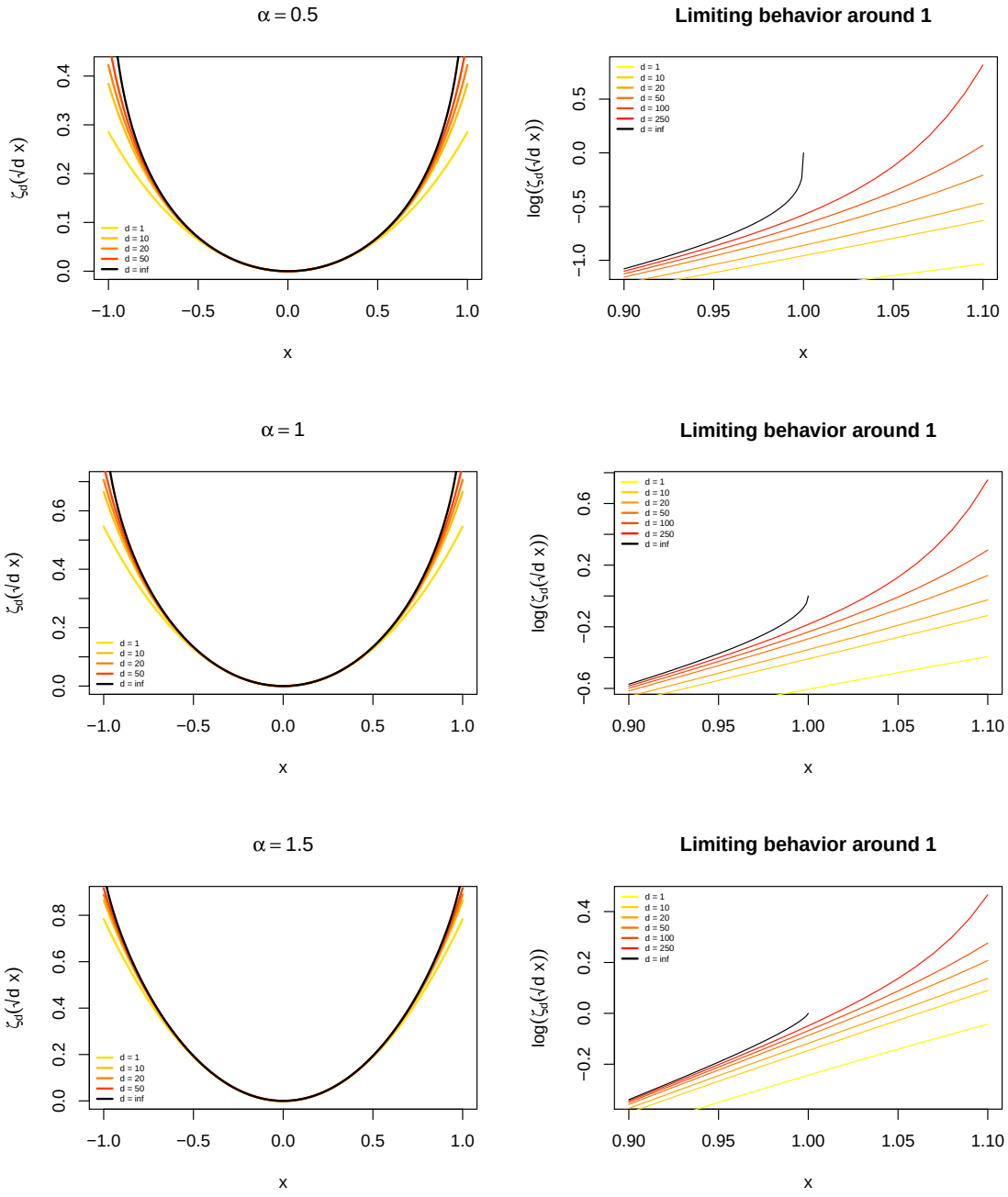


Figure 8.4: Limiting behavior of $\zeta_d(\sqrt{d}\,x)$, as a function of x, when $d \to \infty$. The black curves depict the limit function $1-(1-x^2)^{\frac{\alpha}{2}}$. Left panels: colored curves depict $\zeta_d(\sqrt{d}\,x)$ on $|x| \leq 1$ for $d = 1, 10, 20, 50$. Right panels: behaviour of $\log(\zeta_d(\sqrt{d}\,x))$ around one, with d ranging from 1 to 250.

When $\|y\|$ is large, $(e^{y'x} - 1) \approx e^{y'x}$, so that in this scenario, we can approximate $\zeta_d(y)$ with

$$\int_{\mathbb{R}^d \setminus \{\mathbf{0}\}} e^{-\phi(x)} f(x)\, dx\,,$$

where $\phi(x) = -y'x + \|x\|^2/2$ and $f(x) = K_{d,\alpha}/\|x\|^{\alpha+d}$. Since f is such that $f(x)^{-1} = o(e^{\|x\|^2/2})$, the function $-y'x + \|x\|^2/2 - \log f(x)$ is dominated by $-y'x + \|x\|^2/2$. Therefore, expanding ϕ around its point of maximum $\hat{x} = y$, the Laplace approximation gives

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus \{\mathbf{0}\}} e^{-\phi(x)} f(x)\, dx &\approx \int_{\mathbb{R}^d \setminus \{\mathbf{0}\}} e^{-\phi(\hat{x}) - \frac{1}{2}(x-\hat{x})'\phi''(\hat{x})(x-\hat{x})} f(\hat{x})\, dx \\
&= \int_{\mathbb{R}^d \setminus \{\mathbf{0}\}} e^{\frac{y'y}{2} - \frac{1}{2}(x-y)'(x-y)} f(y)\, dx \\
&= e^{\frac{y'y}{2}} f(y)(2\pi)^{d/2}\,.
\end{aligned}$$

Therefore, for $\|y\|$ large, the zeta function of the rotationally invariant inverse-power measure can be approximated by

$$\frac{e^{\frac{\|y\|^2}{2}}}{\|y\|^{\alpha+d}} \cdot \frac{\alpha\, \Gamma(d/2)\, 2^{\frac{\alpha+d}{2}}}{2\, \Gamma(1-\alpha/2)}\,. \tag{8.9}$$

In Figure 8.5, we compare, on the log scale, the exact ζ_d function with the Laplace approximation in (8.9) for $d = 2$ and $d = 6$, while $\alpha = 1$. We can see that the approximation works quite well in both cases.

8.5 Vector-sparse signal plus noise

In this section, we study the signal-plus-noise model where the observation vector $Y \in \mathbb{R}^d$ is a sum of two independent unobserved random vectors

$$Y = \mu + \eta\,.$$

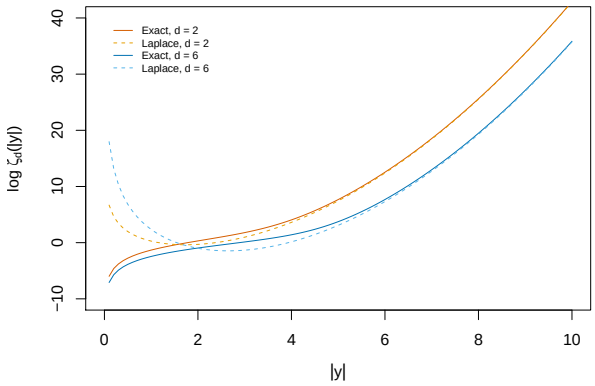


Figure 8.5: Limiting behavior of the d-dimensional zeta function as $\|y\| \to \infty$. The α parameter is fixed to 1, whereas $d = 2$ for the orange curves, and $d = 6$ for the blue curves. The solid darker lines show the exact zeta functions, while the dashed lighter lines depict the corresponding Laplace approximations, both are plotted on the log scale.

Here we assume that the signal and the noise distributions are rotationally invariant with respect to the standard Euclidean inner product. So $\mu \sim P_{\nu,d}$, vector sparse with unit exceedance measure H_d and rate ρ, and it is independent of the Gaussian noise $\eta \sim N_d(0, I_d)$. However, the derivations presented below do not rely on this specific choice of inner product. If for instance $\eta \sim N_d(0, \Sigma)$, where Σ is a known positive definite matrix, then one can consider rotational invariance with respect to $\langle u, v\rangle_{\Sigma^{-1/2}} = u'\Sigma^{-1}v$, as long as the geometry on signal vector space is believed to be the same as the geometry on the error vector space.

The first-order sparse approximation to the marginal distribution for Y at y is

$$\begin{aligned} m_\nu(y) &= \int_{\mathbb{R}^d} \phi_d(y - x)\, P_{\nu,d}(dx) \\ &= \phi_d(y)\Big(\rho \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(\langle y, x\rangle) - 1)e^{-\|x\|^2/2}\, H_d(dx) + \\ &\quad + 1 - \rho \int_{\mathbb{R}^d} (1 - e^{-\|x\|^2/2})\, H_d(dx)\Big) + o(\rho) \\ &= \phi_d(y)\,(\rho\zeta_d(y) + 1 - \rho) + o(\rho)\,. \end{aligned}$$

Since both ϕ_d and ζ_d are spherically symmetric with respect to $\langle\cdot,\cdot\rangle$, so is the sparse approx-

imation to m_ν. Similarly to the univariate sparsity setting, this latter is a two-component mixture

$$m_\nu(y) = \rho\psi_d(y) + (1-\rho)\phi_d(y) + o(\rho)\,,$$

where the function

$$\psi_d(y) = \phi_d(y)\zeta_d(y)$$

is the multivariate analog of the ψ function introduced in McC&P. Indeed, thanks to the normalization of the unitary H_d, $\psi_d(y)$ is a probability density function on $\mathbb{R}^d$, and its characteristic function is given by

$$\int e^{i\langle z,y\rangle}\psi_d(y)\,dy = e^{-\|z\|^2/2}\left(1 - \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} \big(1-\cos(\langle z,u\rangle)\big)\,H_d(dx)\right)\,. \tag{8.10}$$

See the appendix for derivation. When the radial measure is the inverse power, the characteristic function of ψ_d is

$$e^{-\|z\|^2/2}\left(1 - \frac{\|z\|^\alpha}{2^{\alpha/2}}\cdot\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{\alpha+d}{2})}\right)\,.$$

Indeed, for $d=1$, we obtain $e^{-z^2/2}\left(1-|z|^\alpha\cdot\frac{\sqrt{\pi}}{2^{\alpha/2}\Gamma(\frac{\alpha+1}{2})}\right)$, which agrees with the characteristic function of the univariate ψ function derived in McC&P.

8.5.1 Signal conditional distribution

The parallelism with the univariate case carries on to the conditional distribution of the signal μ given the observed vector y. This distribution is proportional to the joint distribution of

(Y, μ)

$$\begin{aligned}\mathbb{P}(\mu \in dx, Y \in dy) &= \phi_d(y) e^{\langle y,x\rangle} e^{-\|x\|^2/2} P_{\nu,d}(dx) \\ &= \phi_d(y) \left((\cosh(\langle y, x\rangle) - 1) e^{-\|x\|^2/2} P_{\nu,d}(dx) + e^{-\|x\|^2/2} P_{\nu,d}(dx) \right) \\ &= \phi_d(y) \left(\rho \zeta_d(dx; y) + e^{-\|x\|^2/2} P_{\nu,d}(dx) \right) + o(\rho)\,,\end{aligned}$$

where the last equality holds in the sense of integrals of functions in $\mathcal{W}_d^{\#}$. Once normalized by $m_\nu(y)$, the first-order sparse approximation to the conditional distribution of the signal can be written as

$$P_{\nu,d}(dx \mid y) = \frac{\rho \zeta_d(dx; y) + e^{-\|x\|^2/2} P_{\nu,d}(dx)}{\rho \zeta_d(y) + 1 - \rho} + o(\rho)\,.$$

This expression is just a generalization of Equation (12) of McC&P. However, because we assume $P_{\nu,d}$ factorizes into a spectral and a radial measure, it is interesting to derive the conditional distributions for the two separate components of the signal, i.e., its direction $\tilde{\mu} = \mu / \|\mu\|$ and its magnitude $\|\mu\|$.

Signal direction conditional distribution

We start by expressing the joint distribution of (μ, Y) in spherical polar coordinates,

$$\mathbb{P}(\tilde{\mu} \in d\tilde{x}, \|\mu\| \in d\,\|x\|\,, Y) = \frac{e^{-\|y\|^2/2}}{(2\pi)^{d/2}} e^{\|x\| \cdot \langle y, \tilde{x}\rangle} e^{-\|x\|^2/2}\, \Gamma(d\tilde{x})\, P_\nu^R(d\,\|x\|)\,. \tag{8.11}$$

Integrating (8.11) over $\mathcal{S}^d$, one obtains

$$\frac{1}{(2\pi)^{d/2}} e^{-\|y\|^2/2} \cosh_d(\|y\|\,\|x\|)\, e^{-\|x\|^2/2}\, P_\nu^R(d\,\|x\|)\,,$$

so that, conditionally on the magnitude $\|\mu\|$ and on y, the signal direction has distribution

$$\frac{e^{\|y\|\|x\|\cdot\langle\tilde{y},\tilde{x}\rangle}\Gamma(d\tilde{x})}{\cosh_d(\|y\|\,\|x\|)}\,.$$

This is the von Mises-Fisher distribution on $\mathcal{S}^d$ with polar direction $\tilde{y} = y/\|y\|$ and concentration parameter $\|y\|\,\|x\|$. On the other hand, the distribution of the signal direction conditional only on y is found after integrating out the radial component

$$\int_0^\infty \mathbb{P}(\tilde{\mu} \in d\tilde{x}, \|\mu\| \in d\,\|x\|\,, Y \in dy)\,.$$

So, after a few passages (see the appendix), we find

$$\mathbb{P}(\tilde{\mu} \in d\tilde{x} \mid Y \in dy) = \Gamma(d\tilde{x})\,\frac{\rho\zeta_1(\|y\|\,\langle\tilde{y},\tilde{x}\rangle) + 1 - \rho}{\rho\zeta_d(\|y\|) + 1 - \rho}\,.$$

This density can be written as a mixture of two components: the uniform measure with relative weight $1-\rho$, and a zeta-tilted uniform measure $\zeta_1(\|y\|\,\langle\tilde{y},\tilde{x}\rangle)\Gamma(d\tilde{x})/\zeta_d(\|y\|)$ with relative weight $\rho\zeta_d(\|y\|)$,

$$\frac{\rho\zeta_d(\|y\|)}{\rho\zeta_d(\|y\|) + 1 - \rho}\,\frac{\zeta_1(\|y\|\,\langle\tilde{y},\tilde{x}\rangle)\Gamma(d\tilde{x})}{\zeta_d(\|y\|)} + \frac{1-\rho}{\rho\zeta_d(\|y\|) + 1 - \rho}\,\Gamma(d\tilde{x})\,. \tag{8.12}$$

Signal magnitude conditional distribution

We now derive the conditional distribution of the signal magnitude. Integrating (8.11) over $(0,\infty)$, one obtains

$$\Gamma(d\tilde{x})\,(\rho\zeta_1(\langle y,\tilde{x}\rangle) + 1 - \rho)\;,$$

so that, conditionally on the direction $\tilde{\mu}$ and on y, the signal magnitude has distribution

$$\frac{\rho\zeta_1(d\,\|x\|\,;\langle y,\tilde{x}\rangle) + e^{-\|x\|^2/2}P_\nu^R(d\,\|x\|)}{\rho\zeta_1(\langle y,\tilde{x}\rangle) + 1 - \rho}\,.$$

Unconditionally of $\tilde{\mu}$, instead, the distribution of $\|\mu\|$ given $Y = y$ is

$$\frac{\rho(\cosh_d(\|y\| \, \|x\|) - 1)e^{-\|x\|^2/2} H^R(d \, \|x\|) + e^{-\|x\|^2/2} P_\nu^R(d \, \|x\|)}{\rho\zeta_d(\|y\|) + 1 - \rho} \, .$$

This also can be written as a mixture of two measures with same weights as for (8.12): the central spike density

$$\mathbb{P}(\|\mu\| \in d \, \|x\| \mid Y = 0) = e^{-\|x\|^2/2} P_\nu^R(d \, \|x\|)/(1 - \rho) \, ,$$

with weight proportional to $1 - \rho$, and the non-central component given by the normalized zeta measure

$$(\cosh_d(\|y\| \, \|x\|) - 1)e^{-\|x\|^2/2} H^R(d \, \|x\|)/\zeta_d(\|y\|) \, ,$$

with weight proportional to $\rho\zeta_d(\|y\|)$.

Signal conditional moments

Given the sparse approximation for the signal conditional distribution, we can derive its moment generating function,

$$\begin{aligned} \int_{\mathbb{R}^d} e^{\langle t,x \rangle} \, P_{\nu,d}(dx \mid y) &= \frac{\phi_d(y)}{m_\nu(y)} \int_{\mathbb{R}^d} e^{\langle t+y,x \rangle} e^{-\|x\|^2/2} \, P_{\nu,d}(d \, \|x\|) \\ &= \frac{\rho\zeta_d(t+y) + 1 - \rho}{\rho\zeta_d(y) + 1 - \rho} \, . \end{aligned}$$

Then the conditional expected value of μ given y is simply

$$\begin{aligned} \mathbb{E}(\mu \mid y) &= \nabla_t \, \frac{\rho\zeta_d(t+y) + 1 - \rho}{\rho\zeta_d(y) + 1 - \rho} \bigg|_{t=0} = \frac{\rho}{\rho\zeta_d(y) + 1 - \rho} \left(\frac{\partial}{\partial x} \zeta_d(x) \bigg|_{x=\|y\|} \nabla_t \, \|t+y\| \, \Big|_{t=0} \right) \\ &= \frac{\rho\zeta_d'(\|y\|)}{\rho\zeta_d(\|y\|) + 1 - \rho} \frac{y}{\|y\|} \, , \end{aligned}$$

where ζ_d' denotes the scalar derivative of the ζ_d function as a function on $(0,\infty)$. Similarly, the conditional moment generating function of the signal magnitude given y is

$$\begin{aligned}\int_0^\infty e^{t\|x\|}\, P_\nu(d\,\|x\| \mid y) &= \frac{\phi_d(y)}{m_\nu(y)} \int_0^\infty \cosh_d((t+\|y\|)\,\|x\|) e^{-\|x\|^2/2}\, P_\nu^R(d\,\|x\|) \\ &= \frac{\rho\zeta_d(t+\|y\|)+1-\rho}{\rho\zeta_d(\|y\|)+1-\rho}\,,\end{aligned}$$

so that

$$\mathbb{E}(\|\mu\| \mid y) = \frac{\partial}{\partial t} \frac{\rho\zeta_d(t+\|y\|)+1-\rho}{\rho\zeta_d(\|y\|)+1-\rho}\bigg|_{t=0} = \frac{\rho\zeta_d'(\|y\|)}{\rho\zeta_d(\|y\|)+1-\rho}\,.$$

Thus,

$$\mathbb{E}(\mu \mid y) = \mathbb{E}(\|\mu\| \mid y)\frac{y}{\|y\|}\,.$$

In a similar fashion, one can compute the r^{th} conditional moment for both μ and $\|\mu\|$.

8.5.2 Double limit condition (DLC)

Consider the rotationally invariant event $\epsilon^+ = \{\|\mu\| > \epsilon\}$, and the corresponding conditional activity probability $P_{\nu,d}(\epsilon^+|y)$, given the observation y. As in the univariate setting, instead of computing the integral for the hard-threshold function $\chi_{\epsilon^+}(dx)$, we consider the soft-threshold function $w_\epsilon(x) = 1 - e^{-\|x\|^2/2\epsilon^2}$ in $\mathcal{W}_d^\#$. Then

$$P_{\nu,d}(\|\mu\| > \epsilon \mid y) = \frac{\rho \int w_\epsilon(x)\, \zeta_d(dx;y) + \int w_\epsilon(x) e^{-\|x\|^2/2}\, P_{\nu,d}(dx)}{\rho\zeta_d(y)+1-\rho} + o(\rho)\,.$$

For any fixed $y \neq 0$, this expression tends to zero as $\rho \to 0$. Indeed, the signal conditional distribution itself converges to the Dirac delta measure at zero as $\rho \to 0$, regardless of the observed vector y.

So, as already done in multiple occasions, in order to get a nontrivial sparse limit for the

conditional probability of the signal activity, we need to let $\|y\| \to \infty$ in such a way that

$$\lim_{\nu \to 0} \lim_{\|y\| \to \infty} \rho \zeta_d(y) \to \lambda \,, \tag{8.13}$$

for some $\lambda > 0$.

Before deriving the sparse approximation for $P_{\nu,d}(\epsilon^+|y)$ under this double limit regime, in this section, we investigate what extra conditions are necessary for this regime to be, in some sense, consistent over all sparse measures having sparsity pair (ρ, H_d). In what follows, we consider d-dimensional sparse measures which are rotationally invariant. However, as we briefly mentioned in the introduction, the same exact reasoning holds for univariate sparse measures as well, with the appropriate change of dimension. So whenever in the thesis, the double limit regime is invoked, one should have in mind the following discussion. This latter, in fact, leads us to identify an extra condition on the exceedance measure, which we call *double limit condition* (DLC), that needs to be verified, under the double limit regime.

Given a pair (ρ, H_d) for which (8.13) holds, we want to establish whether there exists some sparse family $P'_{\nu,d}$ having exceedance measure H_d and rate ρ, for which the equality

$$\lim_{\nu \to 0} \lim_{\|y\| \to \infty} \int_{\mathbb{R}^d} (\cosh(y'u) - 1) e^{-\|u\|^2/2} \, P'_{\nu,d}(du) = \lim_{\nu \to 0} \lim_{\|y\| \to \infty} \rho \zeta_d(y) \tag{8.14}$$

does not hold.

To this end, consider the ϵ-perturbed family

$$P^{\epsilon}_{\nu,d}(du) = (1 - \rho^{1+\epsilon}) P_{\nu,d}(du) + \rho^{1+\epsilon} \tilde{P}_d(du),$$

where $P_{\nu,d}$ is sparse with (ρ, H_d) as rate and exceedance measure, while $\tilde{P}_d$ is an arbitrary

measure. Then $P^{\epsilon}_{\nu,d}$ has same exceedance measure H_d and rate ρ as $P_{\nu,d}$, at least in first order sparsity, as long as $\epsilon > 0$. Since we assume $\rho\zeta_d(y) \to \lambda$, then as $\nu \to 0$ and $\|y\| \to \infty$

$$\begin{aligned}
&\int_{\mathbb{R}^d} (\cosh(y'u) - 1) e^{-\|u\|^2/2} \, P^{\epsilon}_{\nu,d}(du) \\
&\sim (1 - \rho^{1+\epsilon})\rho\zeta_d(y) + \rho^{1+\epsilon} \int_{\mathbb{R}^d} (\cosh(y'u) - 1) e^{-\|u\|^2/2} \, \tilde{P}_d(du) + o(\rho) \\
&\sim \lambda - \rho^{1+\epsilon}\lambda + \rho^{1+\epsilon} \int_{\mathbb{R}^d} (\cosh(y'u) - 1) e^{-\|u\|^2/2} \, \tilde{P}_d(du) + o(\rho) \, .
\end{aligned}$$

Since $\lim_{\nu \to 0} \rho^{1+\epsilon}\lambda = 0$, then for (8.14) not to hold, we need the arbitrary measure $\tilde{P}_d(du)$ to be such that

$$\lim_{\nu \to 0} \lim_{\|y\| \to \infty} \rho^{1+\epsilon} \int_{\mathbb{R}^d} (\cosh(y'u) - 1) e^{-\|u\|^2/2} \, \tilde{P}_d(du) = c \, , \tag{8.15}$$

for some non zero constant c.

To balance $\rho^{1+\epsilon}$, we need to choose the arbitrary measure $\tilde{P}_d(du)$ in such a way that the integral appearing in (8.15) grows at least as fast as $\rho^{-1-\epsilon}$, as $\|y\| \to \infty$. Now, this can happen when $\tilde{P}_d(du)$ puts all of its mass at the maximum of the integrand; in other words, when the arbitrary measure $\tilde{P}_d(du)$ is chosen to be the Dirac delta measure at $\arg\max_u (\cosh(y'u) - 1) e^{-\|u\|^2/2}$. Now, as already observed in the derivation of the Laplace approximation in Section 8.4.2, when $\|y\|$ is large, $(e^{y'u} - 1) \approx e^{y'u}$, so that in this scenario, $\arg\max_u (\cosh(y'u) - 1) e^{-\|u\|^2/2} \approx \arg\max_u e^{y'u} e^{-\|u\|^2/2}$. Thus, for $\|y\|$ large,

$$\max_{\tilde{P}_d} \int_{\mathbb{R}^d} (\cosh(y'u) - 1) e^{-\|u\|^2/2} \, \tilde{P}_d(du) \approx \max_u e^{y'u} e^{-\|u\|^2/2} = e^{\|y\|^2/2} \, .$$

In this way, choosing $\tilde{P}_d(du) = \delta_y(du)$, the limit in the LHS of (8.15) becomes

$$\lim_{\nu \to 0} \lim_{\|y\| \to \infty} \rho^{1+\epsilon} e^{\|y\|^2/2} \, ,$$

so that, for $P^{\epsilon}_{\nu,d}(du) = (1-\rho^{1+\epsilon})P_{\nu,d}(du) + \rho^{1+\epsilon}\delta_y(du)$,

$$\int_{\mathbb{R}^d} (\cosh(y'u) - 1)e^{-\|u\|^2/2} P^{\epsilon}_{\nu,d}(du) \sim \lambda + \rho^{1+\epsilon} e^{\|y\|^2/2} + o(\rho)\,.$$

Going back to our original question, we have that, under the double limit regime (8.13), for the ϵ-perturbed family $P^{\epsilon}_{\nu,d}$, (8.14) holds if and only if the additional condition

$$\lim_{\nu\to 0}\lim_{y\to\infty} \rho^{1+\epsilon} e^{\|y\|^2/2} = 0\,, \tag{8.16}$$

also holds for all $\epsilon > 0$. Therefore, the question becomes for which exceedance measures the two conditions (8.13) and (8.16) cannot hold together. We start by looking at a couple of explicative examples.

1. Let H_d be the rotationally invariant inverse-power measure. Then, as $\|y\| \to \infty$, $\zeta_d(y) \sim e^{\|y\|^2/2}\|y\|^{-\alpha-d}$ so the requirement $\rho\zeta_d(y) \to \lambda$ implies $\rho \sim e^{-\|y\|^2/2}\|y\|^{\alpha+d}$. Therefore,

   $$\rho^{1+\epsilon} e^{\|y\|^2/2} \sim \left(e^{-\|y\|^2/2}\|y\|^{\alpha+d}\right)^{1+\epsilon} e^{\|y\|^2/2} \sim e^{-\epsilon\|y\|^2/2}\,,$$

   which means that, under double limit regime, $\rho^{1+\epsilon}e^{\|y\|^2/2} \to 0$.

2. Let H_d be rotationally invariant with a radial measure having bounded support $[0, M]$. Then $\zeta_d(y) \le e^{M\|y\|}$ so the requirement $\rho\zeta_d(y) \to \lambda$ at most implies $\rho \sim e^{-M\|y\|}$. Therefore,

   $$\rho^{1+\epsilon} e^{\|y\|^2/2} \sim \left(e^{-M\|y\|}\right)^{1+\epsilon} e^{\|y\|^2/2} \sim e^{\|y\|^2/2-(1+\epsilon)M\|y\|}\,,$$

   which means that, in this case, under double limit regime, $\rho^{1+\epsilon}e^{\|y\|^2/2} \to \infty$.

These two examples suggest that some kind of conditions on the tail of the density of H^R is needed. Indeed, since for vector-sparse measures, H_d is rotationally invariant, then its

density h_d can be written as

$$h_d(du) \propto \|u\|^{-(d-1)}\, h^R(\|u\|)du\,,$$

where h^R is the exceedance radial density. So if $h^R(\|u\|) = o(e^{-\|u\|^2/2})$ as $\|u\| \to \infty$, then, using the Laplace approximation as in Section 8.4.2, one obtains

$$\int_{\mathbb{R}^d\setminus\{0\}} e^{y'u-\|u\|^2/2-(d-1)\log\|u\|+\log h^R(\|u\|)}\, du \sim\ e^{\|y\|^2/2}\, h_d(y)(2\pi)^{d/2}\,.$$

Since $\rho\zeta_d(y) \to \lambda$ requires $\rho \sim \frac{1}{\zeta_d(y)} \sim \frac{e^{-\|y\|^2/2}}{h_d(y)}$, then

$$\rho^{1+\epsilon}e^{\|y\|^2/2} \sim \left(\frac{e^{-\|y\|^2/2}}{h_d(y)}\right)^{1+\epsilon} e^{\|y\|^2/2} \sim e^{-\epsilon\|y\|^2/2}\, h^R(\|y\|)^{-1-\epsilon}\,. \tag{8.17}$$

Thus, if indeed $h^R(\|y\|) = o(e^{-\|y\|^2/2})$ as $\|y\| \to \infty$ then $e^{-\epsilon\|y\|^2/2}\, h^R(\|y\|)^{-1-\epsilon} \to 0$, so that (8.13) and (8.16) do hold together. On the contrary, if $h^R(\|y\|) = O(e^{-\|y\|^2/2})$, then

$$\zeta_d(y) \sim e^{\|y\|^2/2\,\frac{1}{1+K}}$$

for some positive K. Then, there exists some $\epsilon > 0$ such that $\frac{1+\epsilon}{1+K} - 1 < 0$, and this implies for such ϵ,

$$\rho^{1+\epsilon}e^{\|y\|^2/2} \sim e^{-\|y\|^2/2\left(\frac{1+\epsilon}{1+K}-1\right)}$$

diverges to infinity.

In conclusion, given a pair (ρ, H_d), H_d rotationally invariant, in order to have

$$\lim_{\nu\to 0}\lim_{\|y\|\to\infty}\int_{\mathbb{R}^d}(\cosh(y'u)-1)e^{-\|u\|^2/2}\, P'_{\nu,d}(du) = \lim_{\nu\to 0}\lim_{\|y\|\to\infty}\rho\zeta_d(y) = \lambda\,,$$

for any $P'_{\nu,d}$ having (ρ, H_d) as sparsity pair, the tail of the radial exceedance density cannot be Gaussian nor sub Gaussian. We refer to this requirement on H^R as double limit condition (DLC).

8.5.3 Bayes exceedance factor

Having investigated the conditions for which we can consider the double limit regime, we now return to its initial motivation. Suppose that DLC holds. Then, the conditional probability of signal activity,

$$P_{\nu,d}(\|\mu\| > \epsilon \mid y) = \frac{\rho \int w_\epsilon(x)\, \zeta_d(dx; y) + \int w_\epsilon(x) e^{-\|x\|^2/2}\, P_{\nu,d}(dx)}{\rho\zeta_d(y) + 1 - \rho} + o(\rho)\,,$$

under the double limit regime (8.13), up to an error $o(1)$, behaves like

$$\frac{\rho\zeta_d(y)}{\rho\zeta_d(y) + 1}\,.$$

This means that, in the double limit, the conditional odds ratio for the exceedance event $\epsilon^+ = \{\|\mu\| > \epsilon\}$, reduces to $\rho\zeta_d(y)$. From Section 8.3, we know that the unconditional probability of ϵ^+ is $\rho H_d(\epsilon^+)$, so the Bayes factor for signal activity is

$$\mathrm{BF}_{\epsilon^+}(y) = \frac{\mathrm{odds}(\|\mu\| > \epsilon \mid Y)}{\mathrm{odds}(\|\mu\| > \epsilon)} = \frac{\rho\zeta_d(y)}{\rho H_d(\epsilon^+)} = \frac{\zeta_d(y)}{H_d(\epsilon^+)}\,.$$

Therefore, if we consider the standard activity threshold ϵ such that $H_d(\epsilon^+) = 1$, then the Bayes factor for this event reduces to $\zeta_d(y)$. For instance, if H_d is the rotationally invariant inverse-power measure with $\alpha = 1$, the standard threshold is roughly 0.8; so the observation vector for which the conditional odds for the event $\|\mu\| > 0.8$ equal the marginal odds, has magnitude $\|y\| = 1.778$ when $y \in \mathbb{R}^2$ ($\mathrm{RMS}(y) = 1.25$), and magnitude $\|y\| = 3.618$ when $y \in \mathbb{R}^{10}$ ($\mathrm{RMS}(y) = 1.14$).

In Table 8.1 and Table 8.2, we report the values for the norm and the root mean square (RMS) of the observation vector $y \in \mathbb{R}^d$, which lead to a Bayes exceedance factor of 1, 10, and 100, for the signal activity event $\|\mu\| > 0.8$. The values in the second and third rows, being such that $\zeta_d(y) = 10$ and $\zeta_d(y) = 100$, respectively, are also the values that give a conditional odds ratio of one, when the sparsity rate is $\rho = 10\%$ and $\rho = 1\%$ respectively.

	$d = 1$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
$\mathrm{BF}_{\epsilon^+}(y) = 1$	1.31	1.78	2.12	2.66	3.62
$\mathrm{BF}_{\epsilon^+}(y) = 10$	2.72	3.28	3.66	4.23	5.23
$\mathrm{BF}_{\epsilon^+}(y) = 100$	3.71	4.18	4.52	5.04	5.98

Table 8.1: Values of $\|y\|$ required for having a given $\mathrm{BF}_{\epsilon^+}(y)$ for the event $\|\mu\| > 0.8$, $\alpha = 1$.

	$d = 1$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
$\mathrm{BF}_{\epsilon^+}(y) = 1$	1.307	1.257	1.226	1.189	1.144
$\mathrm{BF}_{\epsilon^+}(y) = 10$	2.721	2.318	2.114	1.893	1.655
$\mathrm{BF}_{\epsilon^+}(y) = 100$	3.713	2.956	2.609	2.254	1.891

Table 8.2: $\mathrm{RMS}(y)$ required for having a given $\mathrm{BF}_{\epsilon^+}(y)$ for the event $\|\mu\| > 0.8$, $\alpha = 1$.

Now suppose we let the dimension $d \to \infty$. Then, for the standard activity threshold for which $\mathrm{BF}_{\epsilon^+}(y) = \zeta_d(y)$, the Bayes exceedance factor in the limit is

$$\mathrm{BF}^{\infty}_{\epsilon^+}(y) = \lim_{d\to\infty} \zeta_d(\|y\|) = \lim_{d\to\infty} \zeta_d\left(\sqrt{d}\ \mathrm{RMS}(y)\right) .$$

At the end of Section 8.4.1, we derived that

$$\lim_{d\to\infty} \zeta_d\left(\sqrt{d}\, x\right) = \begin{cases} 1 - (1 - x^2)^{\alpha/2} \text{ for } |x| < 1 \\ 1 \qquad\qquad\qquad \text{for } |x| = 1 \\ \infty \qquad\qquad\qquad \text{for } |x| > 1\,. \end{cases}$$

Thus, the Bayes factor for the signal standard activity event $\{\|\mu\| > \epsilon\}$, as the dimension of the signal goes to infinity, has a different limiting behavior depending on the root mean

square of the observed vector y: it converges to a number in $[0, 1)$ if $\mathrm{RMS}(y) < 1$; it is equal to one if $\mathrm{RMS}(y) = 1$; while it diverges to infinity if $\mathrm{RMS}(y) > 1$.

8.6 Appendix

1. Taylor series for $\cosh_d(y)$, Eq. (8.6). We start by considering $\langle y, x\rangle = y'x$. Let θ_1 be the angle such that $\tilde{y}'\tilde{x} = \cos(\theta_1)$. Then, expressing $\tilde{x}$ in spherical coordinates,

$$\begin{aligned}\cosh_d(y) &= \int_{\mathcal{S}^d} e^{y'\tilde{x}} \frac{d\tilde{x}}{\text{Area}(\mathcal{S}^d)} \\ &= \int_0^\infty \int_0^{2\pi} \int_0^\pi \cdots \int_0^\pi \int_0^\pi e^{\|y\| R\cos\theta_1} \frac{R^{d-1}\sin^{d-2}\theta_1 \sin^{d-3}\theta_2 \cdots \sin\theta_{d-2}}{\text{Area}(\mathcal{S}^d)} \cdot \\ &\quad \cdot d\theta_1\, d\theta_2 \cdots d\theta_{d-1} \delta_1(dR)\,.\end{aligned}$$

Now, since

$$\int_0^\pi e^{-\|y\|_2\cos\theta_1}\sin^{d-2}\theta_1 d\theta_1 = \int_0^\pi e^{\|y\|_2\cos(\pi-\theta_1)}\sin^{d-2}(\pi-\theta_1)d\theta_1 = \int_0^\pi e^{\|y\|_2\cos\xi_1}\sin^{d-2}\xi_1\, d\xi_1\,,$$

we can write

$$\int_0^\pi e^{\|y\|_2\cos\theta_1}\sin^{d-2}\theta_1 d\theta_1 = \int_0^\pi \frac{e^{\|y\|_2\cos\theta_1}+e^{-\|y\|_2\cos\theta_1}}{2}\sin^{d-2}\theta_1\, d\theta_1\,,$$

so that

$$\begin{aligned}\cosh_d(y) &= \int_0^\pi \cosh(\|y\|_2\cos\theta_1)\sin^{d-2}\theta_1\, d\theta_1 \cdot \frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})}\frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} \\ &= \int_0^1 \cosh(\|y\|\sqrt{t})(1-t)^{\frac{d-2}{2}-\frac{1}{2}}\frac{dt}{\sqrt{t}} \cdot \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \\ &= \sum_{r=0}^\infty \|y\|_2^{2r}\frac{\Gamma(\frac{d}{2})}{\Gamma(r+\frac{d}{2})}\cdot\frac{\Gamma(r+\frac{1}{2})}{\sqrt{\pi}(2r)!} \\ &= \sum_{r=0}^\infty \frac{\|y\|_2^{2r}}{(d/2)^{\uparrow r}2^{2r}r!}\,.\end{aligned}$$

Any other inner product inducing a norm on $\mathbb{R}^d$, is of the kind $\langle y, x\rangle_A = y'AA'x$ for some positive definite matrix A. In this more general case,

$$\cosh_{d,A}(y) = \int_{\mathcal{S}_A^d} e^{\langle y,\tilde{x}\rangle_A}\frac{d\tilde{x}}{\text{Area}(\mathcal{S}_A^d)}\,,$$

is the same exponential function averaged over the unit ellipsoid defined by A. However, there is an obvious relation between $\cosh_{d,A}(y)$ and $\cosh_d(y)$, which makes the definition $\cosh_d$ the only necessary. In fact,

$$\begin{aligned}\cosh_{d,A}(y) &= \int_{\mathcal{S}_A^d} e^{y'AA'x/\|x\|_A} \frac{d(x/\|x\|_A)\det(A)}{\text{Area}(\mathcal{S}^d)} \\ &= \int_{\mathcal{S}^d} e^{y'AA'x/||A'x||_2} \frac{d(A'x/||A'x||_2)}{\text{Area}(\mathcal{S}^d)} \\ &= \cosh_d(A'y)\,.\end{aligned}$$

Now, because $\cosh_d(y)$ is indeed only a function of the norm of its argument, $\cosh_d(\|y\|_2)$,

$$\cosh_{d,A}(y) = \cosh_d(A'y) = \cosh_d(\|A'y\|_2) = \cosh_d(\|y\|_A)\,.$$

So we can simply define

$$\cosh_d(y) = \int_{\mathcal{S}^d} e^{\langle y,\tilde{x}\rangle} \frac{d\tilde{x}}{\text{Area}(\mathcal{S}^d)} = \sum_{r=0}^{\infty} \frac{\|y\|^{2r}}{(d/2)^{\uparrow r} 2^{2r} r!}\,,$$

for any inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$ and its corresponding unit sphere $\mathcal{S}^d = \{z : \langle z,z\rangle = 1\}$.

2. Expressing $\zeta_d(y)$ with $\cosh_d(y)$, Eq. (8.7).

$$\begin{aligned}\zeta_d(y) &= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(\langle y,x\rangle) - 1)\, e^{-\|x\|^2/2}\, H_d(dx) \\ &= \int_0^\infty \int_{\mathcal{S}^d} (\cosh(\langle \|x\|\, y, \tilde{x}\rangle) - 1) \frac{d\tilde{x}}{\text{Area}(\mathcal{S}^d)}\, e^{-\|x\|^2/2}\, H^R(d\,\|x\|) \\ &= \int_0^\infty \left(\int_{\mathcal{S}^d} \frac{e^{\langle \|x\| y,\tilde{x}\rangle} + e^{-\langle \|x\| y,\tilde{x}\rangle}}{2} \frac{d\tilde{x}}{\text{Area}(\mathcal{S}^d)} - \int_{\mathcal{S}^d} \frac{d\tilde{x}}{\text{Area}(\mathcal{S}^d)} \right) e^{-\|x\|^2/2}\, H^R(d\,\|x\|) \\ &= \int_0^\infty \left(\frac{\cosh_d(\|x\|\, y) + \cosh_d(-\|x\|\, y)}{2} - 1 \right) e^{-\|x\|^2/2}\, H^R(d\,\|x\|) \\ &= \int_0^\infty (\cosh_d(\|x\|\, y) - 1)\, e^{-\|x\|^2/2}\, H^R(d\,\|x\|)\,.\end{aligned}$$

3. Taylor series for $\zeta_d(y)$ when H^R is the inverse-power measure, Eq. (8.8):

$$\begin{aligned}
\zeta_d(y) &= \int_0^\infty (\cosh_d(\|u\|\, y) - 1) e^{-\|u\|^2/2} K_\alpha \, \|u\|^{-\alpha-1} \, d\,\|u\| \\
&= K_\alpha \sum_{r=1}^\infty \frac{\|y\|^{2r}}{(d/2)^{\uparrow r} 2^{2r} r!} \int_0^\infty \|u\|^{2r-\alpha-1} \, e^{-\|u\|^2/2} \, d\,\|u\| \\
&= \frac{\alpha 2^{\alpha/2}}{\Gamma(1-\alpha/2)} \sum_{r=1}^\infty \frac{\|y\|^{2r}}{(d/2)^{\uparrow r} 2^{2r} r!} \frac{\Gamma(r-\alpha/2)}{\left(\frac{1}{2}\right)^{r-\alpha/2-1}} \\
&= \sum_{r=1}^\infty \frac{\|y\|^{2r}}{(d/2)^{\uparrow r} 2^r r!} \frac{\alpha \Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} \,.
\end{aligned}$$

4. We here prove that for all $d \geq 1$,

$$\frac{f_d(r)}{f_{d+1}(r)} = \frac{d^r}{(d+1)^r} \frac{(d/2+1/2)^{\uparrow r}}{(d/2)^{\uparrow r}} \leq 1 \quad \text{for all } r \geq 0\,. \tag{8.18}$$

First we recall that $\Gamma(n/2) = (n-2)!!\sqrt{\pi}/2^{\frac{n-1}{2}}$, where $n!!$ denotes the double factorial as defined in Arfken (1985), p. 547. So,

$$\frac{(d/2+1/2)^{\uparrow r}}{(d/2)^{\uparrow r}} = \frac{(d+2r-1)!!/2^{\frac{d+2r}{2}}}{(d-1)!!/2^{\frac{d}{2}}} \frac{(d-2)!!/2^{\frac{d-1}{2}}}{(d+2r-2)!!/2^{\frac{d+2r-1}{2}}} = \frac{(d+2r-1)!!}{(d-1)!!} \frac{(d-2)!!}{(d+2r-2)!!}\,.$$

Now, for any fixed $d \geq 1$,

- (8.18) holds for $r = 0$ since $\frac{f_d(0)}{f_{d+1}(0)} = 1$;

- if (8.18) holds $r = K$, then (8.18) holds for $r = K + 1$ since

$$\frac{f_d(K+1)}{f_{d+1}(K+1)} = \frac{d^{K+1}}{(d+1)^{K+1}} \frac{(d+2(K+1)-1)!!}{(d-1)!!} \frac{(d-2)!!}{(d+2(K+1)-2)!!} = \frac{d^{K+1}}{(d+1)^{K+1}} \frac{(d+2K+2-1)(d+2K-1)!!}{(d-1)!!} \frac{(d-2)!!}{(d+2K+2-2)(d+2K-2)!!} = \frac{d}{(d+1)} \frac{(d+2K+1)}{(d+2K)} \frac{f_d(K)}{f_{d+1}(K)} \leq 1\,,$$

  where the last inequality holds because:

  1. by the inductive step, $\frac{f_d(K)}{f_{d+1}(K)} \leq 1$;
  2. if $d \geq 1$, then $\frac{d}{(d+1)} \frac{(d+2K+1)}{(d+2K)} \leq 1$ for any $K \geq 0$ since

$$\begin{aligned} \frac{d}{(d+1)} \leq \frac{(d+2K)}{(d+2K+1)} &\iff d(d+2K+1) - (d+2K)(d+1) \leq 0 \\ &\iff d^2 + 2Kd + d - d^2 - 2Kd - d - 2K \leq 0 \\ &\iff -2K \leq 0\,. \end{aligned}$$

Then, by induction, for any fixed $d \geq 1$, (8.18) holds for all $r \geq 0$.

5. Here we show that $\lim_{d\to\infty} \zeta_d(\sqrt{d}) = 1$. In fact, write

$$\zeta_d(\sqrt{d}) = \sum_{r=1}^{\infty} \frac{(d/2)^r}{(d/2)^{\uparrow r} r!} \frac{\frac{\alpha}{2}\Gamma(r - \frac{\alpha}{2})}{\Gamma(1-\alpha/2)} = \sum_{r=1}^{\infty} f_d(r)\,,$$

where the sequence of functions $f_d(r) = \frac{(d/2)^r}{(d/2)^{\uparrow r} r!} \frac{\frac{\alpha}{2}\Gamma(r-\frac{\alpha}{2})}{\Gamma(1-\alpha/2)}$, for $r \geq 1$, is again an increasing sequence of non-negative functions, as for all $d \geq 1$, for all $r \geq 1$

$$\frac{f_d(r)}{f_{d+1}(r)} = \frac{d^r}{(d+1)^r} \frac{(d/2+1/2)^{\uparrow r}}{(d/2)^{\uparrow r}} \leq 1\,.$$

Now,

$$\lim_{d\to\infty} f_d(r) = \lim_{d\to\infty} \frac{(d/2)^r}{(d/2)^{\uparrow r} r!} \frac{\frac{\alpha}{2}\Gamma(r-\frac{\alpha}{2})}{\Gamma(1-\alpha/2)} = \frac{1}{r!} \frac{\frac{\alpha}{2}\Gamma(r-\frac{\alpha}{2})}{\Gamma(1-\alpha/2)} = f(r)\,,$$

and

$$\sum_{r=1}^{\infty} f(r) = 1 - \sum_{r=0}^{\infty} \frac{\Gamma(r-\alpha/2)}{\Gamma(-\alpha/2)r!} = 1 - \sum_{r=0}^{\infty} \frac{(\alpha/2)^{\uparrow r}(-1)^r}{r!} = 1 - (1-1)^{\alpha/2} = 1\,,$$

because $\alpha/2 > 0$ so the binomial series converges absolutely on $[-1, 1]$. Then, once more, by monotone convergence theorem,

$$\lim_{d\to\infty} \zeta_d(\sqrt{d}) = 1\,.$$

6. Limit for $\zeta_d(\sqrt{d}x)$ as $d \to \infty$.

$$\begin{aligned}
\sum_{r=1}^{\infty} f(r) &= \sum_{r=1}^{\infty} \frac{x^{2r}}{r!} \frac{\frac{\alpha}{2}\Gamma(r-\frac{\alpha}{2})}{\Gamma(1-\alpha/2)} \\
&= \sum_{r=0}^{\infty} \frac{x^{2r}}{r!} \frac{\frac{\alpha}{2}\Gamma(r-\frac{\alpha}{2})}{\Gamma(1-\alpha/2)} - \frac{\frac{\alpha}{2}\Gamma(-\frac{\alpha}{2})}{\Gamma(1-\alpha/2)} \\
&= 1 + \sum_{r=0}^{\infty} \frac{\frac{\alpha}{2}\left(r-\frac{\alpha}{2}-1\right)\left(r-\frac{\alpha}{2}-2\right)\ldots\left(r-\frac{\alpha}{2}-r+1\right)\Gamma\left(1-\frac{\alpha}{2}\right)}{\Gamma\left(1-\frac{\alpha}{2}\right)} \frac{x^{2r}}{r!} \\
&= 1 + \sum_{r=0}^{\infty} \frac{\frac{\alpha}{2}(-1)\left(\frac{\alpha}{2}-r+1\right)(-1)\left(\frac{\alpha}{2}-r+2\right)\ldots(-1)\left(\frac{\alpha}{2}-1\right)}{r!} x^{2r} \\
&= 1 + \sum_{r=0}^{\infty} \frac{(-1)^{r-1}\frac{\alpha}{2}\left(\frac{\alpha}{2}-1\right)\ldots\left(\frac{\alpha}{2}-r+2\right)\left(\frac{\alpha}{2}-r+1\right)}{r!} x^{2r} \\
&= 1 - \sum_{r=0}^{\infty} \binom{\frac{\alpha}{2}}{r} (-1)^r x^{2r} \\
&= 1 - (1-x^2)^{\frac{\alpha}{2}}\,.
\end{aligned}$$

where the last equality is true if and only if $|x| \le 1$.

7. Characteristic function for $\psi_d(y) = \phi_d(y)\zeta_d(y)$, Eq. (8.10).

$$
\begin{aligned}
\int e^{iz'y}\psi_d(y)dy &= \int_{\mathbb{R}^d} e^{iz'y} \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cosh(y'u)-1)e^{-\|u\|^2/2}\, H_d(du)\, \phi_d(y)\, dy \\
&= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} \int_{\mathbb{R}^d} e^{iz'y} \left(\frac{e^{y'u}+e^{-y'u}}{2} - 1 \right) \phi_d(y)\, dy\, e^{-\|u\|^2/2}\, H_d(du) \\
&= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} \int_{\mathbb{R}^d} \left(\frac{e^{(iz+u)'y}+e^{(iz-u)'y}}{2} - 1 \right) \frac{e^{-\|y\|^2/2}}{(2\pi)^{d/2}}\, dy\, e^{-\|u\|^2/2}\, H_d(du) \\
&= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} \left(\frac{e^{\|iz+u\|^2/2}+e^{\|iz-u\|^2/2}}{2} - e^{-\|z\|^2/2} \right) e^{-\|u\|^2/2}\, H_d(du) \\
&= \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} e^{-\|z\|^2/2} \left(\frac{e^{\|u\|^2/2+iz'u}+e^{\|u\|^2/2-iz'u}}{2} - 1 \right) e^{-\|u\|^2/2}\, H_d(du) \\
&= e^{-\|z\|^2/2} \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} \left(\frac{e^{iz'u}+e^{-iz'u}}{2} - e^{-\|u\|^2/2} \right) H_d(du) \\
&= e^{-\|z\|^2/2} \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cos(z'u) - e^{-\|u\|^2/2})\, H_d(du)\,.
\end{aligned}
$$

So that, using the normalization of H_d, one can write

$$
\begin{aligned}
\int e^{iz'y}\psi_d(y)dy &= e^{-\|z\|^2/2} \left(\int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (\cos(z'u)-1)\, H_d(du) + \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (1-e^{-\|u\|^2/2})\, H_d(du) \right) \\
&= e^{-\|z\|^2/2} \left(1 - \int_{\mathbb{R}^d\setminus\{\mathbf{0}\}} (1-\cos(z'u))\, H_d(du) \right)\,.
\end{aligned}
$$

8. Passages for the conditional distribution of $\tilde{\mu}$ given y, Eq. (8.12).

$$
\begin{aligned}
&\int_0^\infty \mathbb{P}(\tilde{\mu}\in d\tilde{x}, \|\mu\|\in d\,\|x\|\,, \tilde{Y}\in d\tilde{y}, \|Y\|\in d\,\|y\|) \propto \\
&\int_0^\infty e^{\|y\|\|x\|\cdot\langle\tilde{y},\tilde{x}\rangle} e^{-\|x\|^2/2} \Gamma(d\tilde{x})\, P_\nu^R(d\,\|x\|) = \\
&\Gamma(d\tilde{x}) \int_{-\infty}^\infty \cosh(z\cdot\|y\|\,\langle\tilde{y},\tilde{x}\rangle) e^{-z^2/2}\, P_\nu(dz) = \\
&\Gamma(d\tilde{x}) \left(\rho \int_{-\infty}^\infty (\cosh(z\cdot\|y\|\,\langle\tilde{y},\tilde{x}\rangle)-1) e^{-z^2/2}\, H(dz) + 1 - \rho \int (1-e^{-z^2/2})\, H(dz) \right) = \\
&\Gamma(d\tilde{x})\, (\rho\zeta_1(\|y\|\,\langle\tilde{y},\tilde{x}\rangle) + 1 - \rho)\,.
\end{aligned}
$$

Chapter 9

Vector-sparse ANOVA

9.1 Introduction

Analysis of variance is a cornerstone of statistical practice, and its relevance in the analysis of factorial experiments can hardly be exaggerated. Given the observation space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle)$, and a nested sequence of subspaces $A_0 \subset A_1 \subset \ldots \subset A_k \subset \mathbb{R}^n$, to each pair of these subspaces, $A \subset B$, it corresponds the subspace $A^\perp \cap B$. ANOVA specifies a sequence of subspace pairs so to identify a set of mutually orthogonal subspaces, for which mean squares are calculated. For any such subspace, under the ANOVA null hypothesis, the mean square for that subspace has the same expected value as the mean square for the residual subspace.

In this chapter, we propose a vector-sparse approach to the linear regression problem, exploiting the assumption that the distribution of the coefficient vector $\beta \in \mathbb{R}^d$, comprising the effects associated with the subspace, is vector sparse. The rotational invariance on the parameter space is defined with respect to the Fisher-information metric, so that a matrix $O \in \mathbb{R}^{d\times d}$ defining the symmetry is such that $OX'XO' = X'X$. Given this choice of metric, we derive the sparse approximation to the marginal density of the response $m_\nu(y)$, as a mixture of the n-dimensional Gaussian density and its product with the d-dimensional

zeta function defined in Chapter 4. Interestingly, it turns out that $m_\nu(y)$ depends on the covariate matrix X only through its orthogonal projection matrix $P_{\mathcal{X}} = X(X'X)^{-1}X'$, so that the choice of basis in which the matrix X is expressed, is not relevant.

Given the sparse approximations to the marginal distribution of the response and to the conditional distribution for the coefficient vector magnitude, we derive a sparse approximation to the Bayes factor for the exceedance event $\|\beta\|_A > \epsilon$. This Bayes factor can be compared, to some extent, with the standard ANOVA F-test. In this attempt, we look at how the tail probability, under the ANOVA F-test, matches up with the sparse exceedance Bayes factor, and how this varies for different dimensions of the vector β.

The ANOVA null hypothesis is $\beta = 0$, while the vector-sparse analysis presented in this chapter considers the event $\|\beta\|_A \leq \epsilon$, for some $\epsilon > 0$. Having said that, with the negligibility theory developed in Chapter 4, it would be easy to extend our vector-sparse analysis to give an approximation the uncertainty of the event $\|\beta\|_A \leq \epsilon_\nu$, where ϵ_ν is a negligibility sequence for radial sparse measure P^R_ν.

9.2 Vector-sparse linear regression

Consider the linear regression model on the Euclidean space $(\mathbb{R}^n, \|\cdot\|_2)$

$$Y = \mu_0 + X\beta + \epsilon\,.$$

Here $\epsilon \sim N_n(0, \sigma^2 I_n)$ while the vector μ_0 contains the additive effects on the response, of some matrix $X_0 \in \mathbb{R}^{n\times d_0}$, which are not expected to be negligible. By contrast, the matrix $X = (x_1, \ldots, x_d) \in \mathbb{R}^{n\times d}$ is the covariate matrix whose effects on the response are expected to be sparse. For this reason, the coefficient vector $\beta \in \mathbb{R}^d$ is assumed to be random with a

vector-sparse distribution. Then, conditionally on $\beta \sim P_{\nu,d}$,

$$Y - \mu_0 \sim N(X\beta, \sigma^2 I_n)\,.$$

We assume that X has full column rank and spans $\mathcal{X} \subset \mathcal{X}_0^{\perp}$, where $\mathcal{X}_0 = \text{span}(X_0)$.

As highlighted in the previous chapter, a vector-sparse measure is a measure whose sparsity is driven by its radial component. Clearly, the inner product chosen on $\mathbb{R}^d$ determines the radius as well as the rotational invariance which we assume for $P_{\nu,d}$. In this linear regression context, we choose the inner product on the parameter space to be dictated by the matrix $A \in \mathbb{R}^{p\times p}$ in such a way that, given $u, v \in \mathbb{R}^d$,

$$\langle u, v\rangle_A = u'AA'v = u'X'Xv/\sigma^2\,.$$

This choice has multiple advantages for it makes the linear map $X/\sigma : \mathbb{R}^d \to \mathbb{R}^n$ an isometric embedding: in mapping a vector β in the parameter space $(\mathbb{R}^d, \langle\cdot,\cdot\rangle_A)$, to the vector $X\beta/\sigma$ in the observation space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_2)$, the norm is preserved

$$\|\beta\|_A^2 = \|X\beta\|_2^2/\sigma^2\,.$$

Let ρ and H^R be the sparsity pair characterizing the sparse measure for the radial part of β, $\|\beta\|_A$. For simplifying the notation, for the moment let $\mu_0 = 0$. Then, denoting by $\phi_{n,\sigma}(y) = \sigma^{-n}\phi_n(y/\sigma)$, the marginal density of the response vector at y is

$$m_\nu(y) = \phi_{n,\sigma}(y)\int_{\mathbb{R}^d} e^{y'X\beta/\sigma^2} e^{-\|X\beta\|_2^2/2\sigma^2} P_{\nu,d}(d\beta)\,.$$

Now, thanks to the vector sparsity assumption, $P_{\nu,d}(d\beta)$ can be written as

$$\frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ P_\nu^R(d\,\|\beta\|_A)\,,$$

where $\tilde{\beta} = \beta/\,\|\beta\|_A$ is the direction on the unit sphere $\mathcal{S}_A^d = \{z : \|z\|_A^2 = z'X'Xz/\sigma^2 = 1\}$. Following the theory developed for the rotationally invariant measures in Chapter 8, we write the exponent in $e^{y'X\beta/\sigma^2}$ as an inner product on the parameter space

$$y'X\beta/\sigma^2 = y'X(AA')^{-1}AA'\beta/\sigma^2 = \langle (X'X)^{-1}X'y, \beta\rangle_A = \langle \eta, \beta\rangle_A\,,$$

where $\eta = (X'X)^{-1}X'y \in \mathbb{R}^d$. Then, using the $\cosh_d$ function, we can write

$$\begin{aligned} m_\nu(y) &= \phi_{n,\sigma}(y)\int_0^\infty \int_{\mathcal{S}_A^d} e^{y'X\beta/\sigma^2}e^{-\|X\beta\|^2/2\sigma^2}\frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ P_\nu^R(d\,\|\beta\|_A) \\ &= \phi_{n,\sigma}(y)\int_0^\infty \int_{\mathcal{S}_A^d} e^{\|\beta\|_A\langle\eta,\tilde{\beta}\rangle_A}e^{-\|\beta\|_A^2/2}\frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ P_\nu^R(d\,\|\beta\|_A) \\ &= \phi_{n,\sigma}(y)\int_0^\infty \cosh_d(\|\beta\|_A\,\eta)e^{-\|\beta\|_A^2/2}\ P_\nu^R(d\,\|\beta\|_A)\,. \end{aligned}$$

So the first-order sparse approximation to the marginal density of Y at y is

$$m_\nu(y) = \phi_{n,\sigma}(y)\,(\rho\zeta_d(\eta) + (1-\rho))\ ,$$

so that, also in the case of linear regression, the marginal density of Y is approximated by a mixture of two components: the Gaussian density $\phi_{n,\sigma}(y)$ and its product with the d-dimensional zeta function $\zeta_d(\eta) = \zeta_d((X'X)^{-1}X'y)$. Now, recall that ζ_d is indeed a function only of the norm of its argument, so

$$\zeta_d(\eta) = \zeta_d(\|\eta\|_A)\,,$$

and because of our choice of norm,

$$\|\eta\|_A^2 = y'X(X'X)^{-1}X'y/\sigma^2 = \|P_{\mathcal{X}}y\|_2^2/\sigma^2\,,$$

where $P_{\mathcal{X}} = X(X'X)^{-1}X'$ is the orthogonal projection matrix onto $\mathcal{X} = \text{span}(x_1, \ldots, x_d)$. So we can interchangeably write $\zeta_d(\eta)$, $\zeta_d(\|\eta\|_A)$, and $\zeta_d(\|P_{\mathcal{X}}y\|/\sigma)$, where, for simplicity of notation, henceforth we will write $\|y\| = \|y\|_2$.

Now, if the mean vector $\mu_0 \in \mathcal{X}_0$ is non zero, then

$$m_\nu(y) = (1-\rho)\phi_{n,\sigma}(y-\mu_0) + \rho\phi_{n,\sigma}(y-\mu_0)\zeta_d(\|P_{\mathcal{X}}y\|/\sigma) + o(\rho)\,, \tag{9.1}$$

since $\mathcal{X} \subset \mathcal{X}_0^\perp$ implies $P_{\mathcal{X}}(\mu_0) = 0$. It is interesting to notice that the marginal density of the response depends on X only through the orthogonal projection matrix $P_{\mathcal{X}}$. Thus, the choice of the basis vectors for X is irrelevant insofar the image space $\mathcal{X}$ is what matters.

9.2.1 Coefficient conditional distribution

The conditional distribution of the coefficient vector β given the observed response y is proportional to the joint distribution of (Y, β). So, letting $\eta = (X'X)^{-1}X'y$,

$$\begin{aligned}
\mathbb{P}(\beta \in d\beta, Y \in dy) &= \phi_{n,\sigma}(y)\, e^{y'X\beta/\sigma^2} e^{-\|X\beta\|^2/2\sigma^2}\, P_{\nu,d}(d\beta) \\
&= \phi_{n,\sigma}(y)\left(\left(\cosh(\langle\eta,\beta\rangle_A) - 1\right) e^{-\|\beta\|_A^2/2}\, P_{\nu,d}(d\beta) + e^{-\|\beta\|_A^2/2}\, P_{\nu,d}(d\beta)\right) \\
&= \phi_{n,\sigma}(y)\left(\rho\zeta_d(d\beta;\eta) + e^{-\|\beta\|_A^2/2}\, P_{\nu,d}(d\beta)\right) + o(\rho)\,,
\end{aligned}$$

where the last equality holds in the sense of integrals of functions in $\mathcal{W}_d^{\#}$. Once normalized by $m_\nu(y)$, the first-order sparse approximation to the conditional distribution of the signal

can be written as

$$P_{\nu,d}(d\beta \mid y) = \frac{\rho\zeta_d(d\beta;\eta) + e^{-\|\beta\|_A^2/2}\, P_{\nu,d}(d\beta)}{\rho\zeta_d(\eta) + 1 - \rho} + o(\rho)\,.$$

Similarly to the signal plus noise model, we can derive the conditional distribution for the direction and magnitude of the coefficient vector.

Coefficient direction conditional distribution

Conditionally on the magnitude $\|\beta\|_A$ and on y, letting $\tilde{\eta} = \eta/\|\eta\|_A$, the coefficient direction $\tilde{\beta}$ has distribution

$$\frac{e^{\|\eta\|_A \|\beta\|_A \cdot \langle\tilde{\eta},\tilde{\beta}\rangle}\Gamma(d\tilde{\beta})}{\cosh_d(\|\eta\|_A \|\beta\|_A)}\,,$$

where $\Gamma(d\tilde{\beta}) = d\tilde{\beta}/\mathrm{Area}(\mathcal{S}_A^d)$ is the uniform measure on the unit sphere $\mathcal{S}_A^d$. This is the von Mises-Fisher distribution on $\mathcal{S}_A^d$ with polar direction $\tilde{\eta} = (X'X)^{-1}X'y/\|P_{\mathcal{X}}y/\sigma\|$ and concentration parameter $\|\eta\|_A \|\beta\|_A = \|P_{\mathcal{X}}y\|\,\|\beta\|_A$. On the other hand, the distribution of $\tilde{\beta}$ conditional only on y is

$$\mathbb{P}(\tilde{\beta} \in d\tilde{\beta} \mid Y \in dy) = \Gamma(d\tilde{\beta})\,\frac{\rho\zeta_1(\langle(X'X)^{-1}X'y,\tilde{\beta}\rangle_A) + 1 - \rho}{\rho\zeta_d(\|P_{\mathcal{X}}y/\sigma\|) + 1 - \rho}\,.$$

This density can be written as a mixture of two components: the uniform measure with relative weight $1-\rho$, and a zeta-tilted uniform measure with relative weight $\rho\zeta_d(\|P_{\mathcal{X}}y/\sigma\|)$,

$$\frac{\rho\zeta_d(\|P_{\mathcal{X}}y/\sigma\|)}{\rho\zeta_d(\|P_{\mathcal{X}}y/\sigma\|) + 1 - \rho}\,\frac{\zeta_1\left(\langle(X'X)^{-1}X'y,\tilde{\beta}\rangle_A\right)\Gamma(d\tilde{\beta})}{\zeta_d(\|P_{\mathcal{X}}y/\sigma\|)} + \frac{1-\rho}{\rho\zeta_d(\|P_{\mathcal{X}}y/\sigma\|) + 1 - \rho}\,\Gamma(d\tilde{\beta})\,. \tag{9.2}$$

Coefficient magnitude conditional distribution

We now derive the conditional distribution of the coefficient magnitude. Conditionally on the direction $\tilde{\beta}$ and on y, the coefficient magnitude has distribution

$$\frac{\rho\zeta_1(dr; \langle \eta, \tilde{\beta} \rangle_A) + e^{-r^2/2} P_\nu^R(dr)}{\rho\zeta_1(\langle \eta, \tilde{\beta} \rangle_A) + 1 - \rho} .$$

Unconditionally of $\tilde{\beta}$, instead, the distribution of $\|\beta\|_A$ given $Y = y$ is

$$\frac{\rho(\cosh_d(\|P_{\mathcal{X}} y/\sigma\| \; r) - 1) e^{-r^2/2} H^R(dr) + e^{-r^2/2} P_\nu^R(dr)}{\rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|) + 1 - \rho} . \tag{9.3}$$

This also can be written as a mixture of two measures with same weights appearing in (9.2): the central spike density

$$\mathbb{P}(\|\beta\|_A \in dr \mid Y = 0) = e^{-r^2/2} P_\nu^R(dr)/(1-\rho) ,$$

with weight proportional to $1 - \rho$, and the non-central component given by the normalized zeta measure

$$(\cosh_d(\|P_{\mathcal{X}} y/\sigma\| \, r) - 1) e^{-r^2/2} H^R(dr)/\zeta_d(\|P_{\mathcal{X}} y/\sigma\|) ,$$

with weight proportional to $\rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|)$.

9.2.2 Bayes exceedance factor

Given the sparse approximation for the conditional distribution of $\|\beta\|_A$ in (9.3), we can ask for the probability of the exceedance event $\{\|\beta\|_A \geq \epsilon\}$, where $\epsilon > 0$, given the data, and look at how it compares to the initial exceedance probability $P_\nu^R(\epsilon^+)$. As done many times already, we approximate the hard-threshold exceedance probability with the soft-threshold exceedance probability by computing the expectation of the function of $w_\epsilon(x) = 1 - e^{-x^2/2\epsilon^2}$,

so

$$P_\nu^R(\|\beta\|_A > \epsilon \mid y) = \frac{\int w_\epsilon(r)\left(\rho\big(\cosh_d(\|\eta\|_A\, r) - 1\big)e^{-r^2/2}\, H^R(dr) + e^{-r^2/2}\, P_\nu^R(dr)\right)}{\rho\zeta_d(\|\eta\|_A) + 1 - \rho} + o(\rho)\,.$$

Suppose that H^R does not have Gaussian nor sub Gaussian tail, i.e., it satisfies DLC of Section 8.5.2. Then assuming $\rho \to 0$ and $\|P_{\mathcal{X}} y/\sigma\| \to \infty$ in such a way that

$$\rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|) \to \lambda\,,$$

for some $\lambda > 0$, the conditional probability that the coefficient magnitude is above $\epsilon > 0$, $\{\|\beta\|_A \geq \epsilon\} = \{\|X\beta\|/\sigma \geq \epsilon\}$, behaves like

$$\frac{\rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|)}{\rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|) + 1} + o(1)\,. \tag{9.4}$$

In this case, the posterior odds ratio for $\{\|\beta\|_A \geq \epsilon\}$ does not depend on the threshold ϵ and can be approximated by $\rho\zeta_d(\|P_{\mathcal{X}} y\|/\sigma)$.

Therefore, if we consider the activity threshold ϵ for which the unconditional exceedance probability is

$$P_\nu^R(\epsilon^+) = \frac{\rho}{1+\rho} + o(\rho)\,,$$

then the prior odds ratio for $\{\|\beta\|_A \geq \epsilon\}$ is simply ρ. So the Bayes factor for this exceedance event is

$$\mathrm{BF}_{\epsilon^+}(y) = \frac{\mathrm{odds}(\|\beta\|_A \geq \epsilon \mid y)}{\mathrm{odds}(\|\beta\|_A \geq \epsilon)} = \zeta_d(\|P_{\mathcal{X}} y\|/\sigma)\,. \tag{9.5}$$

We can regard (9.5) as the factor by which the evidence, provided by the data, multiplies the initial odds ratio $\rho/(1+\rho)$.

Notice that, if the residual variance σ^2 is estimated by the residual mean square on

$k = n - d - d_0$ degrees of freedom, then

$$f_y = (\|P_{\mathcal{X}} y\|^2 / d) / (\|Q_{\mathcal{X} \cup \mathcal{X}_0} y\|^2 / k)$$

is the observed value of the standard F-ratio statistic. In this case, the argument of the zeta function in (9.4) and (9.5), reduces to be $\sqrt{d \cdot f_y}$, so, if k is very large, then the Bayes factor can be computed as

$$\mathrm{BF}_{\epsilon^+}(y) = \zeta_d(\sqrt{d \cdot f_y}\,). \tag{9.6}$$

9.3 Sparse Bayes factors and F-ratios

At the end of Chapter 8, we studied the asymptotic behavior of the ζ_d function as $d \to \infty$ and found that, for $f \geq 0$,

$$\lim_{d \to \infty} \zeta_d\left(\sqrt{d\, f}\right) = \begin{cases} 1 - (1-f)^{\alpha/2} & \text{for } f < 1 \\ 1 & \text{for } f = 1 \\ \infty & \text{for } f > 1\,, \end{cases} \tag{9.7}$$

where α is the inverse power of the exceedance measure. For a fixed value of the F-ratio, f, if this is less or equal than one, as the dimension of β gets larger and larger, the Bayes factor converges to a number in $[0, 1]$. On the other hand, if f is larger than one, then, as $d \to \infty$, the Bayes factor diverges, even if slowly. So, when the observed value of the F-ratio is resulting from the response being projected onto a large dimensional subspace, then the evidence-multiplicative factor (9.15) grows large if $f > 1$, but if $f \leq 1$, it does not collapse to zero, unless $f = 0$.

We now try to compare, to some extent, the sparse Bayes factor in (9.15) for the exceedance event $\|\beta\|_A > \epsilon$, to the standard ANOVA test, which employs the F-ratio f_y to

test the null hypothesis $\beta = 0$. To this end, we start by fixing the degrees of freedom for the residual sum of squared to be $k = 100$, and consider a set of tail probabilities, or p-values, as if they were obtained from observed values of

$$F_Y = (\|P_{\mathcal{X}} Y\|^2 / d) / (\|Q_{\mathcal{X} \cup \mathcal{X}_0} Y\|^2 / k), \tag{9.8}$$

which, under the null $H_0 : \beta = 0$, has Fisher's F-distribution $F_{d,k}$. So, for varying dimension d, we compute the $F_{d,k}$ quantiles which give rise, under the ANOVA null hypothesis, to each p-value: $f_{d,k} = F_{d,k}^{-1}(p\text{-value})$. Then for each of these $f_{d,k}$ values, we compute the conditional probability of $\|\beta\|_A \geq \epsilon$,

$$\frac{\rho \zeta_d \left(\sqrt{F_{d,k}^{-1}(p\text{-value})} \right)}{1 + \rho \zeta_d \left(\sqrt{F_{d,k}^{-1}(p\text{-value})} \right)}. \tag{9.9}$$

In Figure 9.2, we show this conditional probability, when $\alpha = 1$, for a range of p-values from 0.5% to 10%. Since the conditional probability depends on the sparsity rate of the radial sparse measure P_ν^R, we consider two options for this rate: in the left panel we fix $\rho = 0.1$ regardless of the dimension d of the vector; in the right panel, instead, we let $\rho = 0.1$ be the rate for when $d = 1$, and for $d > 1$, we let the rate for P_ν^R scale with the dimension, so that $\rho_d = 0.1 \cdot \frac{\sqrt{\pi}\Gamma(d/2+1/2)}{\Gamma(d/2)}$. The scaling factor, already mentioned in Chapter 8, is shown in Figure 9.1.

Looking at Figure 9.2, as expected, the conditional probability for the signal magnitude to be active gets larger as the associated level of significance, i.e. tail probability, gets smaller. This happens irrespective of both the dimension d, and the rate chosen. If instead we look at the behavior of (9.9) as a function of d, then we observe different behaviors depending on the choice for the rate. For fixed ρ, as d grows, the conditional probability of $\|\beta\|_A \geq \epsilon$ first decreases and then increases, the curvature being more prominent for very small p-values. Instead, when also the rate gets scaled by the dimension, for any given p-value, the corresponding conditional probability increases monotonically as the dimension increases.

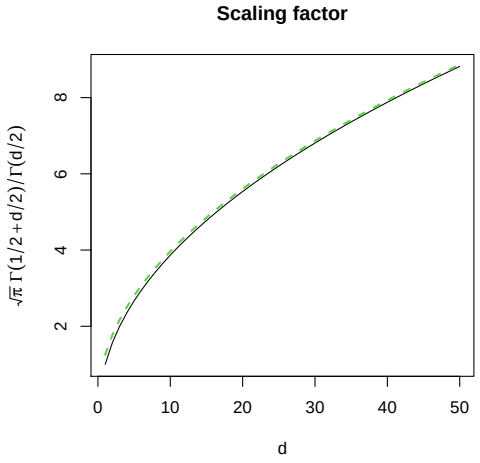


Figure 9.1: Dimension scaling effect on the one-dimensional sparsity rate. The dashed green line depicts the limiting behavior $\frac{\sqrt{\pi}\Gamma(d/2+1/2)}{\Gamma(d/2)} \sim \sqrt{\pi}(d/2)^{1/2}$ as $d \to \infty$.

In Table 9.1, we report the Bayes factor

$$\zeta_d \left(\sqrt{d \cdot F_{d,k}^{-1}(p\text{-value})} \right), \tag{9.10}$$

for different p-values and dimension d ranging from 1 to 10, while $k = 100$. Roughly speaking, moving from a p-value of 1% to 0.5% corresponds to multiplying the Bayes factor by 1.5, while when passing from 5% to 1%, the Bayes factor gets multiplied by approximately 2.6.

To look into the reasons for the curvature observed in the left panel of Figure 9.2, in Figure 9.3, we plot the Bayes factor in (9.10) as a function of $\log(d)$, under two scenarios: one with $k = 100$ (left panel), one with $k = \infty$ (right panel). This last scenario, corresponds to the hypothetical case when σ^2 is known, so the F-ratio in (9.8) reduces to a scaled quadratic form, distributed as a scaled chi-square random variable on d degrees of freedom. We can see that, for fixed $k = 100$, across all significance levels, the behavior of the Bayes factor as a function of $\log(d)$, is not monotone. By contrast, when $k = \infty$, as d gets larger, the Bayes factor decreases and, as $d \to \infty$, it seems to converge to one, for any level of

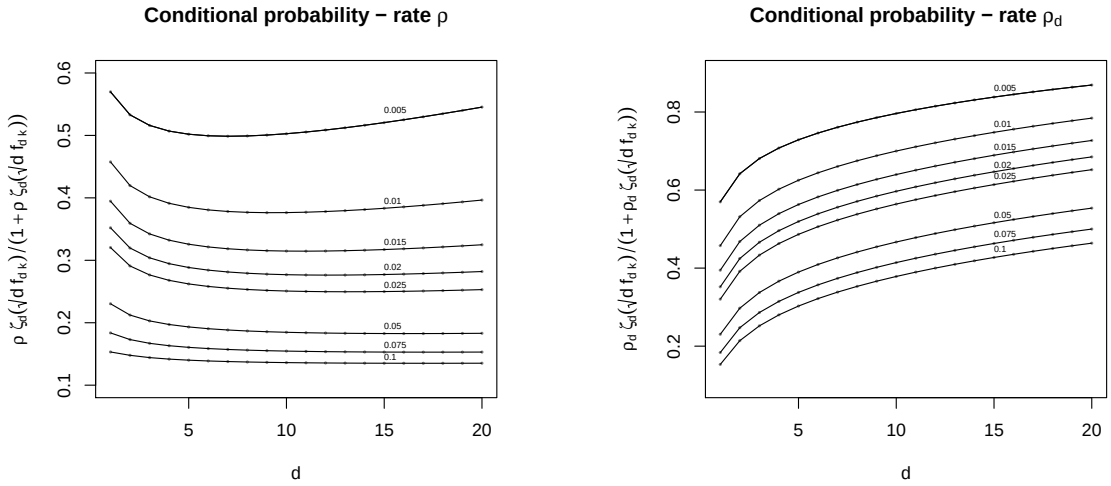


Figure 9.2: Conditional probability of the exceedance event $\{\|\beta\|_A > \epsilon\}$ as a function of the dimension d. In each plot, the different curves correspond to different p-values, ranging from 0.5% to 10%. These in turn determine the $f_{d,k}$ values used to compute (9.9), where k is fixed to 100. Left panel: $\rho = \rho = 0.1$ for all dimensions d. Right panel: $\rho = \frac{\sqrt{\pi}\Gamma(d/2+1/2)}{\Gamma(d/2)}\,\rho = \frac{0.1\,\sqrt{\pi}\Gamma(d/2+1/2)}{\Gamma(d/2)}$.

significance.

The reason why this happens is that , when $k = \infty$, the $F_{d,k}$ quantiles are indeed χ^2_d quantiles divided by d; so asymptotically as $d \to \infty$, $F^{-1}_{d,\infty}(p\text{-value}) \asymp 1$, irrespective of the significance level. This can be easily shown by using the asymptotic approximations for the χ^2_d quantiles, proposed either by Fisher (1928) [36] or by Wilson and Hilferty (1931) [76]. So from (9.7), as $d \to \infty$,

$$\zeta_d\left(\sqrt{d \cdot F^{-1}_{d,\infty}(p\text{-value})}\right) \to 1\,.$$

Whereas, when $k < \infty$, as $d \to \infty$, $F^{-1}_{d,k}(p\text{-value}) \to f_k(p\text{-value})$, and for this range of p-values from 0.5% to 10%, $f_k(p\text{-value}) > 1$. This in turn leads to divergence,

$$\zeta_d\left(\sqrt{d \cdot F^{-1}_{d,k}(p\text{-value})}\right) \to \infty\,,$$

since $\zeta_d(\sqrt{df})$ converges to the finite function $1 - \sqrt{1-f}$ only if $0 \le f \le 1$.

So in a sense, the curvature in the Bayes factor for large d is due to the fact that if the residual degrees of freedom are kept constant, as d grows larger, the $F_{d,k}$ quantiles get

p-value$\times 10^2$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$	$d=9$	$d=10$
0.5	13.25	11.41	10.66	10.28	10.08	9.98	9.95	9.97	10.02	10.11
1.0	8.44	7.23	6.71	6.43	6.25	6.15	6.08	6.05	6.03	6.04
1.5	6.52	5.61	5.20	4.97	4.83	4.74	4.67	4.63	4.61	4.59
2.0	5.43	4.70	4.37	4.18	4.05	3.97	3.91	3.87	3.85	3.83
2.5	4.72	4.10	3.82	3.66	3.55	3.48	3.43	3.39	3.37	3.35
5.0	2.99	2.69	2.55	2.46	2.40	2.35	2.32	2.30	2.28	2.27
7.5	2.25	2.09	2.01	1.95	1.91	1.89	1.87	1.85	1.84	1.83
10.0	1.81	1.73	1.68	1.65	1.63	1.61	1.60	1.59	1.58	1.58

Table 9.1: $\zeta_d\left(\sqrt{dF_{d,k}^{-1}(p\text{-value})}\right)$ for different dimensions d and different p-values, with fixed $k = 100$.

smaller, so the relative weight of the residual mean square in the F-ratio becomes larger. Indeed, if k is relatively small, one should consider $d \cdot F_\beta = \|X\beta\|^2 / (\|Q_{\mathcal{X}\cup\mathcal{X}_0} y\|^2 / k)$, rather than $\|\beta\|_A^2 = \|X\beta\|^2 / \sigma^2$, and derive the distributions of F_β and F_y taking into account the variability of the estimator for σ^2. We extend our theory to this case in Section 9.5.

9.4 Illustrative example 1

To illustrate the vector-sparse approach to the analysis of variance, we now present a real-data example, coming from a randomized experiment, described and analyzed by Villa *et al.* (2019) [72]. The aim of the study was concerned with reproductive isolation, and consequent ecological speciation, occurring in response to body-size evolution in isolated lineages of pigeon lice. The experiment started by founding 32 lice lineages from a total of 800 lice, on 32 host lice-free pigeons, and comprised observing the evolution of each lineage over 60 generations. To induce differentiability in the body-size of the evolving lice lineages, half of these were randomly assigned to be placed on normal-sized captive feral pigeons, while the other half were placed on giant runts. The body size of a giant runt pigeon is roughly three times that of a captive feral pigeon. This was done with the expectation that the host size was to cause a different evolution of the louse size. The two groups of pigeons were then further randomly assigned to four aviaries each, maintaining the two groups separate, in order to

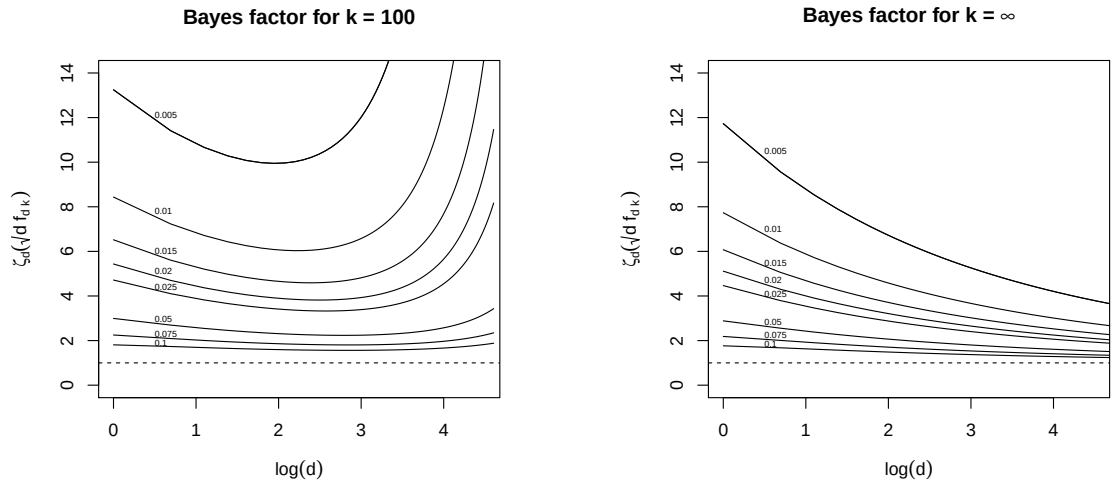


Figure 9.3: Bayes factor for the exceedance event $\{\|\beta\|_A > \epsilon\}$ as a function of the dimension d, on the logarithmic scale. In each plot, the different curves correspond to different p-values, ranging from 0.5% to 10%, which in turn determine $f_{d,k} = F_{d,k}^{-1}(p\text{-value})$ used to compute (9.10). Left panel: $k = 100$. Right panel: $k = \infty$.

	MS	dof
Host type	12	1
Aviary	290	6
Lineage	83	24
Residual	111	∞

Table 9.2: Mean squares times 10^5 of the log body length in μm, with corresponding degrees of freedom, for the three factors and residual, as measured at baseline.

avoid interference. The experiment was designed in such a way that every six months, a random sample of lice was taken from each pigeon, photographed and then returned to the bird. On each occasion, the sex and three body-size measurements were recorded for each louse.

If the randomization procedure was successful, at baseline, the expected value of the mean squares (MSs) of any of the responses, associated with the treatment factor, Host, and the two block factors, Lineage and Aviary, must be the same as the expected value of the residual mean square. However, if we look at Table 9.2, which reports the mean squares of the log body length immediately after randomization, two figures call for attention. On the

one hand, the Host MS appear to be significantly smaller than the residual MS; on the other hand, the MS for Aviary seems to be significantly bigger than the residual MS. Yet, if the former of these anomalies could be explained by an effort, in the randomization procedure, directed to minimize the baseline variability due to the treatment, the latter anomaly is a little more suspicious. The F-ratio for the Aviary factor is $290/111 = 2.61$ on $d = 6$ and $k = \infty$ degrees of freedom, which corresponds to a tail probability of roughly 1.6%. This small value could be interpreted as an indication that something went wrong in the randomization procedure. Now, if we adopt the vector-sparse approach, calling $\beta_{\text{Av}} \in \mathbb{R}^6$ the coefficients associated with the subspace Aviary $\cap$ Host$^\perp \cap \mathbf{1}^\perp$, the Bayes factor for $\|\beta_{\text{Av}}\|_A = \|X_{\text{Av}}\beta_{\text{Av}}\| / \sigma > \epsilon$ in (9.15), is $\zeta_6(\sqrt{6 \cdot 2.61}) = 3.8$. Here we are considering the threshold ϵ for which the prior odds ratio is equal to ρ. So, for $\alpha = 1$, this threshold is roughly given by $0.8 \cdot (1 + \rho)$. Clearly, for small ρ, ϵ can be simply taken to be the standard threshold 0.8 for which $H^R(\epsilon^+) = 1$. So, say that the starting odds are of 1 to 10 in favor of $\|\beta_{\text{Av}}\|_A > 0.88$, reflecting the belief about the possibility that something, in the random assignment of the lice to the pigeons, and/or of the pigeons to the aviaries, goes wrong. Then the odds become 1 to 38 after seeing the outcome of the randomization.

9.5 Unknown variance

So far, we have treated the error variance σ^2 as known, even when we estimated it with the residual mean square. In this section, we generalize the theory we derived for σ^2 known, taking into account the variability of $\|Q_{\mathcal{X} \cup \mathcal{X}_0} Y\|^2 / k$. In other words, we derive the analog of Fisher's F distribution when the coefficient vector β is random with a vector-sparse distribution. For notational convenience, we let $\mu_0 = 0$, so that the residual mean square can be written as $\|Q_{\mathcal{X}} y\|^2 / k$, where $k = n - d$. Also, for the sake of space, here we present only the main facts, and we refer to the appendix for the full derivation of each formula.

In order to obtain a conditional distribution for the coefficient vector not dependent of σ, instead of $\|\beta\|_A = \|X\beta\|/\sigma$, we derive the conditional distribution of the parameter F-ratio

$$F_\beta = \frac{\|\beta\|_A^2/d}{\|Q_{\mathcal{X}}Y/\sigma\|^2/k} = \frac{\|X\beta\|^2/d}{\|Q_{\mathcal{X}}Y\|^2/k},$$

given the observation F-ratio

$$F_Y = \frac{\|P_{\mathcal{X}}Y\|^2/d}{\|Q_{\mathcal{X}}Y\|^2/k}.$$

In fact, when two random variables are both scaled by the same scalar, considering events for their ratio, is the same as restricting the sigma algebra of each of the two random variables, to those events which are scale invariant, i.e., to those events $A \in \mathcal{F}_X$ such that $\mathbb{P}_X(\lambda A) = \mathbb{P}_X(A)$, for every $\lambda \neq 0$. For this reason, in the derivations that follow, without loss of generality, we can fix σ^2 to be one.

We start by writing the observation F-ratio as P/Q, where the sum of squares in the numerator $d \cdot P = \|P_{\mathcal{X}}Y\|^2$ has density function at p^2

$$\chi^2_d(p^2)\left(1-\rho+\rho\zeta_d\big(\sqrt{p^2}\,\big)\right), \tag{9.11}$$

whereas the sum of squares in the denominator $k \cdot Q = \|Q_{\mathcal{X}}Y\|^2 \sim \chi^2_k$. Moreover P and Q are independent so one can derive the density of F_Y at f_y to be

$$m_{d,k}(f_y) = \mathcal{F}_{d,k}(f_y)\,\big(1-\rho+\rho\,\zeta^F_{d,k}(f_y)\big) + o(\rho)\,, \tag{9.12}$$

where $\mathcal{F}_{d,k}$ denotes Fisher's F density with d and k degrees of freedom, while

$$\zeta^F_{d,k}(f_y) = \sum_{r=1}^{\infty} \frac{(df_y)^r}{(df_y+k)^r} \frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow r}}{(d/2)^{\uparrow r} r!} \frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} \tag{9.13}$$

is the zeta function for F-ratios on d and k degrees of freedom.

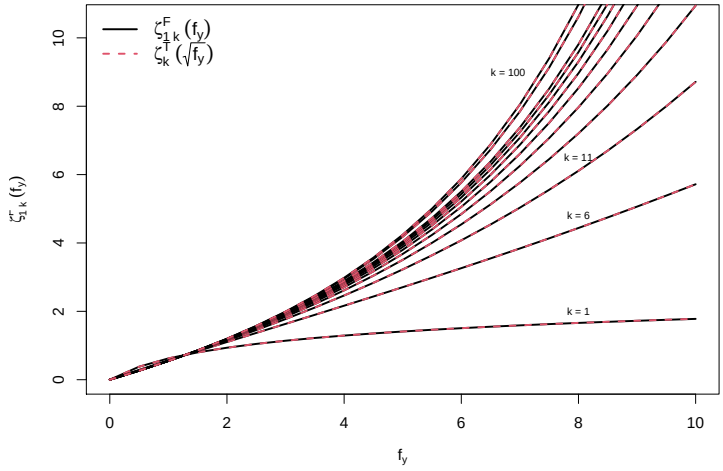


Figure 9.4: Zeta functions for F-ratios $\zeta^F_{d,k}(f_y)$ when $d = 1$ (black solid lines) and zeta functions for t-statistics $\zeta^T_k(\sqrt{f_y})$ (red dashed lines). The residual degrees of freedom k range from 1 to 100.

When $d = 1$, the zeta function for the F-ratio on $1, k$ degrees, coincides with the zeta function for the t-statistic ζ^T_k, we introduced in Chapter 1,

$$\zeta^F_{1,k}(f_y) = \zeta^T_k(\sqrt{f_y}\,)\,.$$

This is shown in Figure 9.4, for different values of k ranging from 1 to 100: the black solid curves depict $\zeta^F_{1,k}(f_y)$ while the red dashed curves depict $\zeta^T_k(\sqrt{f_y}\,)$.

On the other hand, as the residual degrees of freedom k go to infinity, $\zeta^F_{d,k}(f_y)$ converges to the d-dimensional zeta function $\zeta_d\big(\sqrt{df_y}\,\big)$. We show this convergence on the log scale, in Figure 9.5. For $d = 3$ (left panel) and $d = 10$ (right panel), the black lines show $\zeta^F_{d,k}(f_y)$ for k ranging from 1 to 500 and 5000, respectively, while the dashed red curve depicts $\zeta_d(\sqrt{df_y})$. This fact in turn, implies that, as $k \to \infty$, (9.12) converges to

$$\chi^2_d(df_y)\,\Big(1 - \rho + \rho\,\zeta_d\big(\sqrt{df_y}\,\big)\Big) + o(\rho)\,,$$

which is the same expression we would get for the sparse approximation to the density of $\frac{\|P_{\mathcal{X}}Y\|^2/d}{\sigma^2}$, if σ^2 was known. (See appendix).

The marginal density written in (9.12) is, once more, a mixture of two components: Fisher's F density and its product with the zeta function for F-ratios. It can be shown that this latter component $\psi_{d,k}(f_y) = \mathcal{F}_{d,k}(f_y)\zeta^F_{d,k}(f_y)$ is itself a probability density function. (See appendix). In Figure 9.6, we plot both the tail inflating component $\psi_{d,k}(f_y)$ (left panels) and the sparse approximation to the marginal density of F_Y, $m_{d,k}(f_y)$ (right panels), for different combinations of d and k. In each panel, as a term of comparison, we also plot the corresponding non-sparse Fisher's F density function, $\mathcal{F}_{d,k}(f_y)$.

In a similar fashion, with some more algebra, one can derive the joint probability of (F_β, F_Y) at (f_β, f_y). This can be found in the appendix. Here we directly write the conditional density for the scaled parameter F-ratio, $d \cdot F_\beta$, given the observation F-ratio f_y. To avoid notational confusion, instead of dx, we write ∂x to denote the differential form of x. So,

$$\mathbb{P}(dF_\beta \in \partial x \mid F_Y = f_y) = \frac{P_\nu^{R^2}(\partial x)\, w_{d,k}(f_y, x) + \rho\zeta^F_{d,k}(\partial x; f_y)}{1 - \rho + \rho\,\zeta^F_{d,k}(f_y)} + o(\rho)\,, \tag{9.14}$$

where

$$w_{d,k}(f_y, x) = \left(\frac{df_y + k}{df_y + k + x}\right)^{\frac{d}{2}+\frac{k}{2}} \frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow -\frac{\alpha}{2}} 2^{-\frac{\alpha}{2}}}{(df_y + k + x)^{-\frac{\alpha}{2}}}\,,$$

while

$$\zeta^F_{d,k}(\partial x; f_y) = H^{R^2}(\partial x)\, \left(\frac{df_y + k}{df_y + k + x}\right)^{\frac{d}{2}+\frac{k}{2}} \sum_{r=1}^{\infty} \frac{(df_y)^r (x)^r}{(d/2)^{\uparrow r} r!} \frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2}) 2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y + k + x)^{2r-\frac{\alpha}{2}}}$$

is the zeta measure for F-ratios on d and k degrees of freedom. One can indeed check that

$$\int \zeta^F_{d,k}(\partial x; f_y)\partial x = \zeta^F_{d,k}(f_y)\,.$$

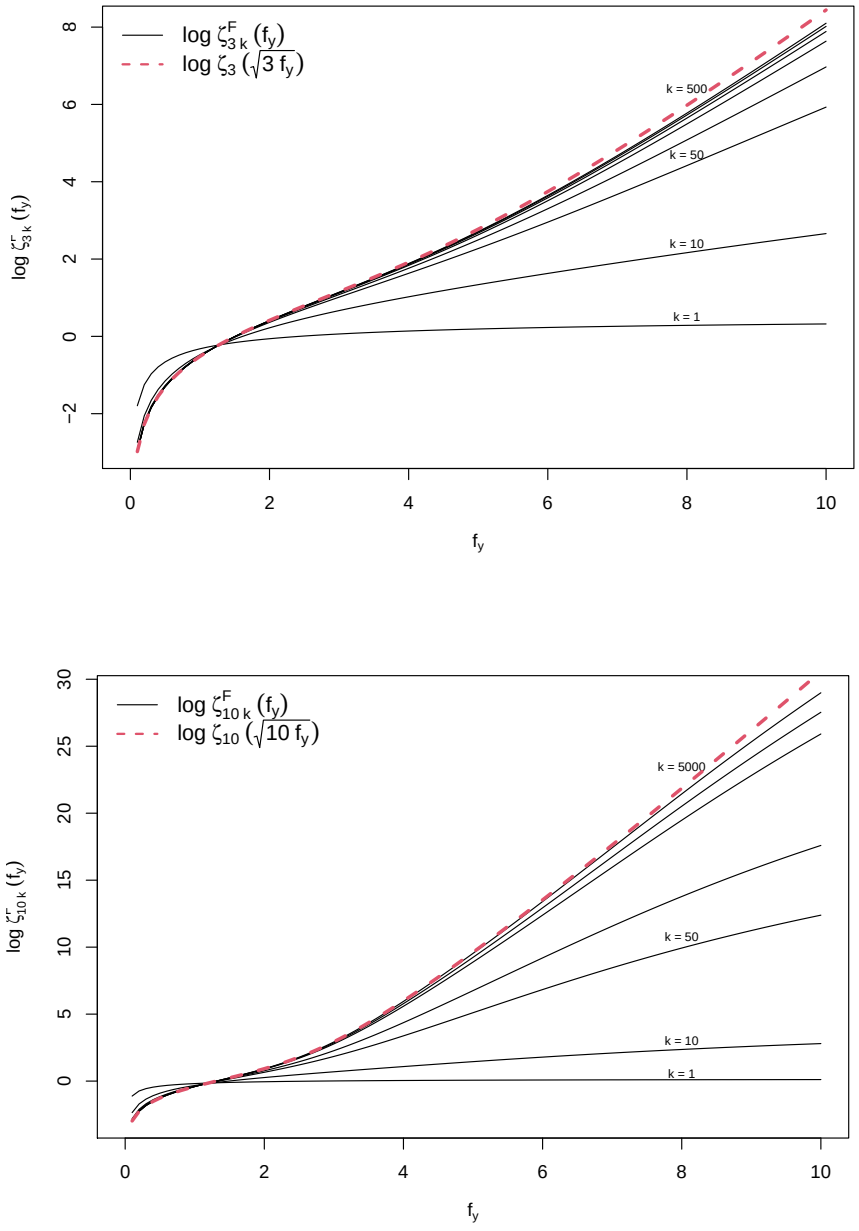


Figure 9.5: Convergence of $\zeta^F_{d,k}(f_y) \to \zeta_d(\sqrt{df_y})$ as $k \to \infty$. Left panel: $d = 3$, and k ranging from 1 to 500. Right panel: $d = 10$, and k ranging from 1 to 5000. In both panels, the red dashed curve depicts the limit function $\zeta_d(\sqrt{df_y})$.

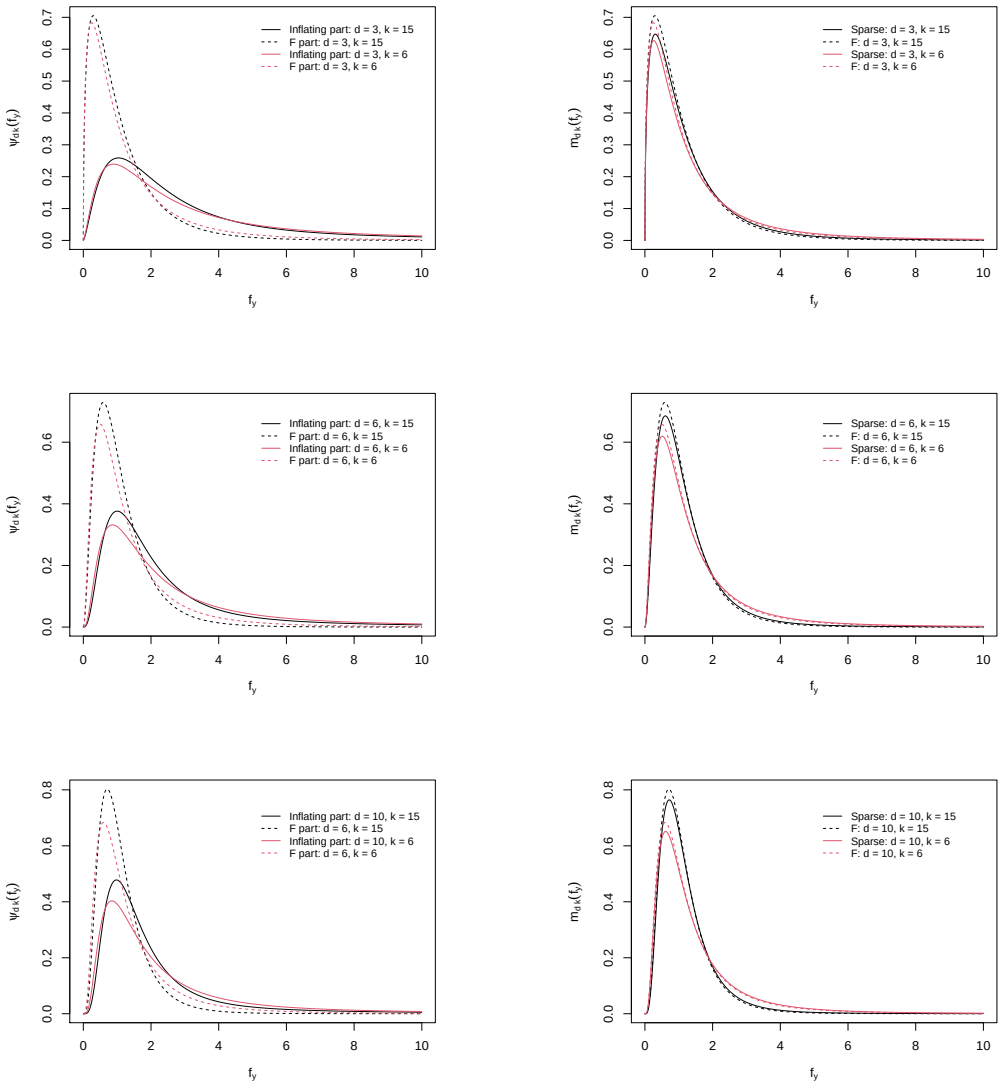


Figure 9.6: Tail inflation component $\psi_{d,k}(f_y)$ and sparse marginal density of F_Y, $m_{d,k}(f_y)$, solid curves, compared to the corresponding non-sparse Fisher's F density function, $\mathcal{F}_{d,k}$, dashed curves. The sparsity rate is $\rho = 10\%$. Top panel: $d = 3$ and $k = 6$ (red curves) and $k = 15$ (black curves). Middle panel: $d = 6$ and $k = 6$ (red curves) and $k = 15$ (black curves). Bottom panel: $d = 10$ and $k = 6$ (red curves) and $k = 15$ (black curves).

(See appendix). It is also possible to show that, as $k \to \infty$, (9.14) converges to

$$\frac{e^{-df_\beta} P_\nu^{R^2}(\partial x) + \rho\, \zeta_d\left(\partial\sqrt{x}; \sqrt{df_y}\right)}{1 - \rho + \rho\, \zeta_d(\sqrt{df_y})} + o(\rho)\,,$$

and this is the sparse approximation to the conditional distribution

$$\mathbb{P}\left(\|\beta\|_A^2 \in \partial x \mid F_Y = f_y\right)\,,$$

when σ^2 is known. (See appendix).

9.6 Sparse Bayes factor and F-ratios revisited

In Section 9.2.2, assuming σ^2 was known, we approximated the conditional probability of $\|\beta\|_A \geq \epsilon$ with

$$\frac{\rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|)}{1 + \rho\zeta_d(\|P_{\mathcal{X}} y/\sigma\|)}\,.$$

Now that we are estimating σ^2 with the residual mean square s^2, instead of $\{\|\beta\|_A \geq \epsilon\} = \{\|X\beta\|/\sigma \geq \epsilon\}$, we can consider $\{\|X\beta\|/s \geq \epsilon\} = \{\|X\beta\|^2/s^2 \geq \epsilon^2\} = \{dF_\beta \geq \epsilon^2\}$, and approximate the conditional probability of this event with

$$\frac{\rho\zeta_{d,k}^F(f_y)}{1 + \rho\,\zeta_{d,k}^F(f_y)}\,.$$

So when the unconditional odds for $\{dF_\beta \geq \epsilon^2\}$ are $\rho : 1$, then the Bayes factor for this event reduces to

$$\mathrm{BF}_{\epsilon^+}(y) = \frac{\mathrm{odds}(\|X\beta\|/s \geq \epsilon \mid f_y)}{\mathrm{odds}(\|X\beta\|/s \geq \epsilon)} = \zeta_{d,k}^F(f_y)\,. \tag{9.15}$$

In Figure 9.7, we show this Bayes factor for $d = 3, 5, 10, 15$ and $k = 3, 5, 10, 15, 30$. For all values of d, there is a transition observed approximately around $f_y = 1.2$, even though, in no case, the curves do all intersect at the same point (the appearance from the plots is misleading). For values of f_y smaller than roughly 1.2, larger k leads to smaller $\mathrm{BF}_{\epsilon^+}(f_y)$,

whereas for values of f_y greater than 1.2, larger k leads to larger $\mathrm{BF}_{\epsilon^+}(f_y)$. So larger values of k act as a deflating / inflating factor for $\mathrm{BF}_{\epsilon^+}(f_y)$, depending on the size of the observed F-ratio. On the other hand, larger degrees of freedom for $\|P_{\mathcal{X}}y\|^2$ has the effect of exaggerating this phenomenon, leading to larger Bayes factors when either both k and f_y are large, or both k and f_y are small. For all d, the value of f_y giving a Bayes factor of one, needs to be larger as k decreases.

In Figure 9.8, instead, we compare $\zeta^F_{d,k}(f_{d,k})$ with $\zeta_d(\sqrt{df_{d,k}})$, where we plot these two Bayes factors as functions of the dimension d, for different F-quantiles,

$$f_{d,k} = F^{-1}_{d,k}(p\text{-value})\,,$$

corresponding to p-values ranging from 0.5% to 10%. In Section 9.3, we investigated the sparse Bayes factor for the norm exceedance event $\{\|\beta\|_A > \epsilon\}$

$$\zeta_d\left(\|P_{\mathcal{X}}y\|/\sigma\right)\,,$$

and estimated it with

$$\zeta_d\left(\|P_{\mathcal{X}}y\|/s\right) = \zeta_d\left(\sqrt{df_y}\right)\,,$$

treating s^2 as if it was σ^2, i.e., as if s^2 was estimated on $k = \infty$ degrees of freedom. Yet, when k is in fact finite, treating σ^2 as known leads to some kind of divergence when $d \to \infty$. This divergence of the Bayes factor, $\zeta_d\left(\sqrt{df_{d,k}}\right)$, is shown in the top left panel of Figure 9.8, where $k = 25$ and d is growing large. By contrast, from the top right panel, we can see that, when we estimate the Bayes factor for the F-ratio with $\zeta^F_{d,k}(f_{d,k})$, as $d \to \infty$, it converges to one, even if $k = 25$. On the other hand, if we assume $k = \infty$, and consider

$$f_{d,\infty} = \chi^{2\,-1}_d(p\text{-value})/d\,,$$

then the two Bayes factors, $\zeta_d\left(\sqrt{df_{d,\infty}}\right)$ and $\zeta^F_{d,5000}(f_{d,\infty})$, practically coincide and slowly converge to one. This is because $\zeta^F_{d,k}(f_y) \to \zeta_d\left(\sqrt{df_y}\right)$ when $k \to \infty$.

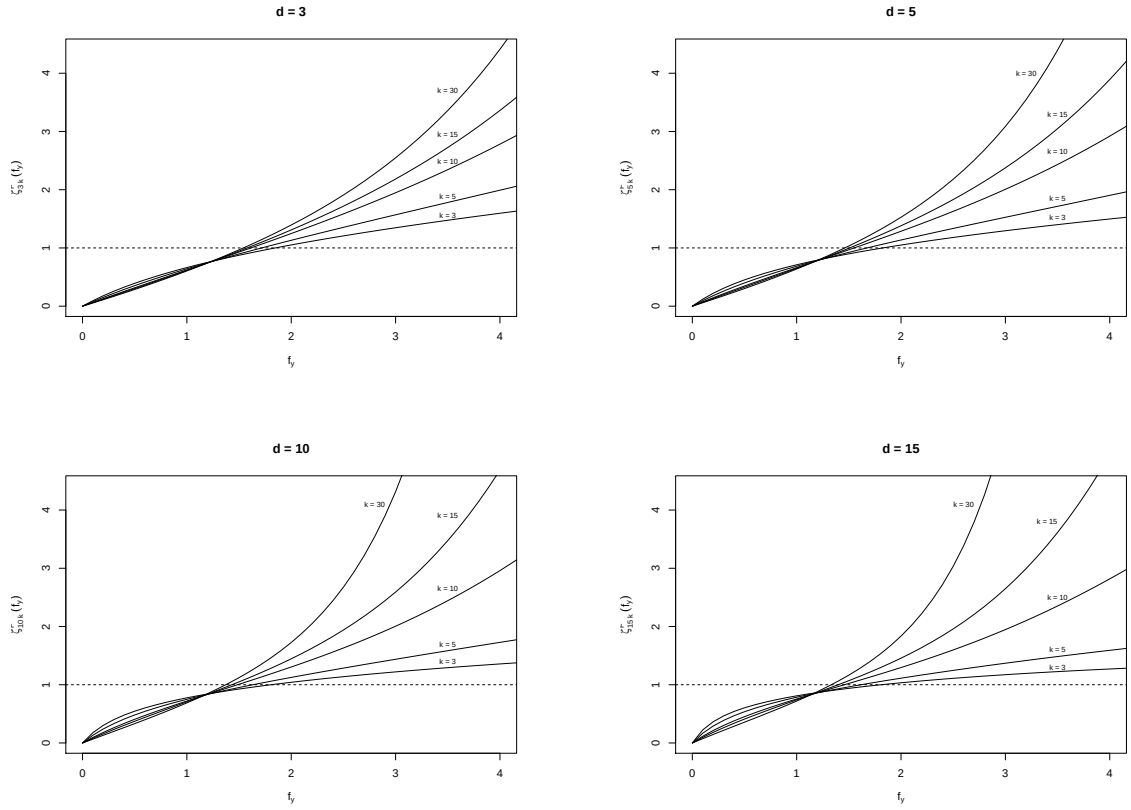


Figure 9.7: $\zeta^F_{d,k}(f_y)$ for $d = 3, 5, 10, 15$, and $k = 3, 5, 10, 15, 30$.

For a last comparison, we look at the asymptotic behavior of the zeta function for F-ratios for fixed values of the argument f. In Figure 9.9, we fix k to be small and plot $\zeta^F_{d,k}(f)$ as a function of $\log(d)$, for values of f ranging from 1 to 10. We can see that in the limit, even for large f, as $d \to \infty$, $\zeta^F_{d,k}(f) \to 1$. In Figure 9.10, instead, we let $k = d$, and look at $\zeta^F_{d,d}(f)$ when d gets large. Under this scenario, the limiting behavior depends on the argument: if $f \leq 1$, then $\zeta^F_{d,d}(f) \to 1 - ((1-f)/(1+f))^{\alpha/2}$, whereas if $f > 1$, $\zeta^F_{d,d}(f)$ diverges to infinity as $d \to \infty$.

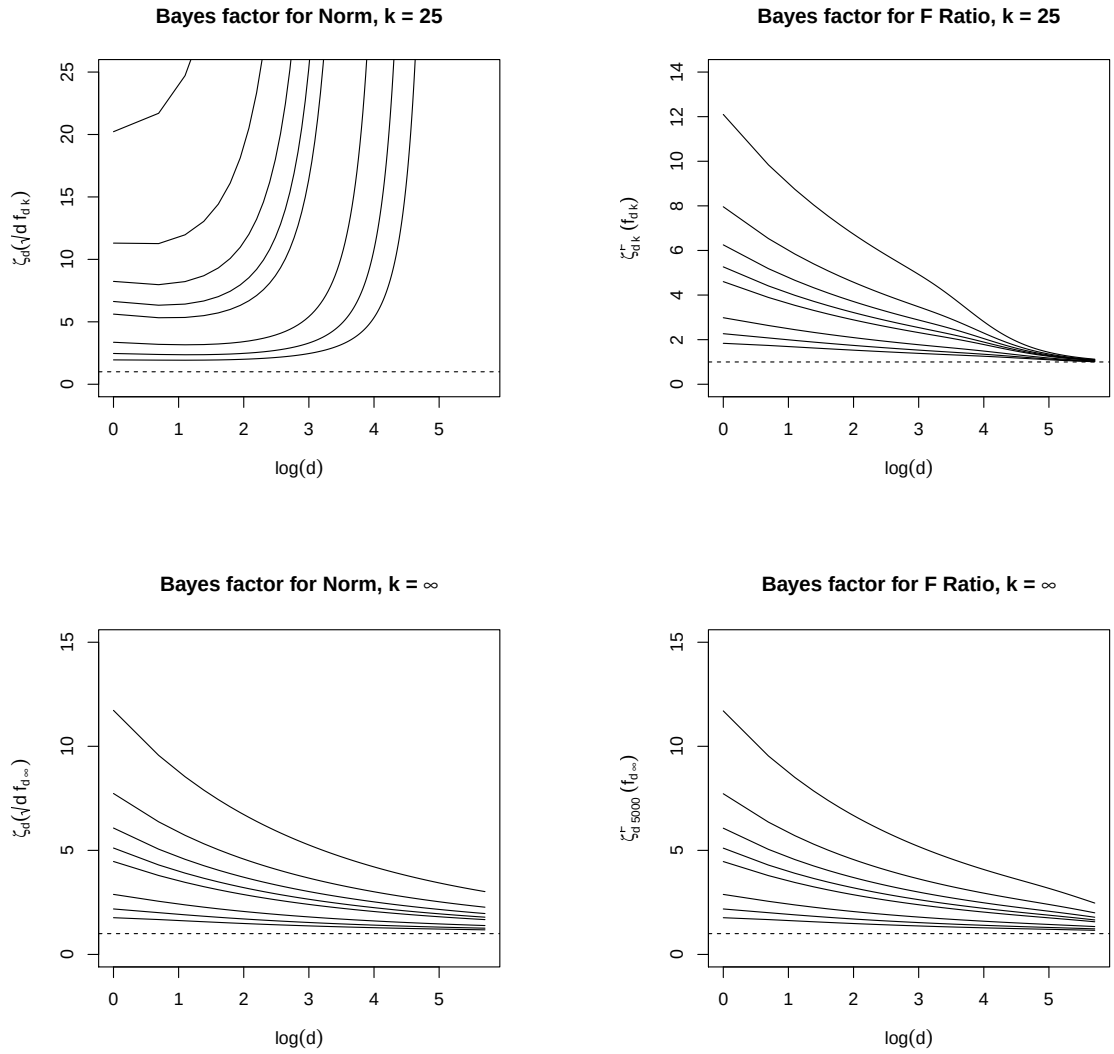


Figure 9.8: Comparison of $\zeta_d(\sqrt{df_{d,k}})$ (left panels) with $\zeta^F_{d,k}(f_{d,k})$ (right panels) as functions of $\log(d)$. Top panels have $k = 25$, while bottom panels have $k = \infty$. In each panel, different curves correspond to different $f_{d,k} = F^{-1}_{d,k}(p\text{-value})$, where p-value$\times 10^2 = 0.5, 1, 1.5, 2, 2.5, 5, 7.5, 10$.

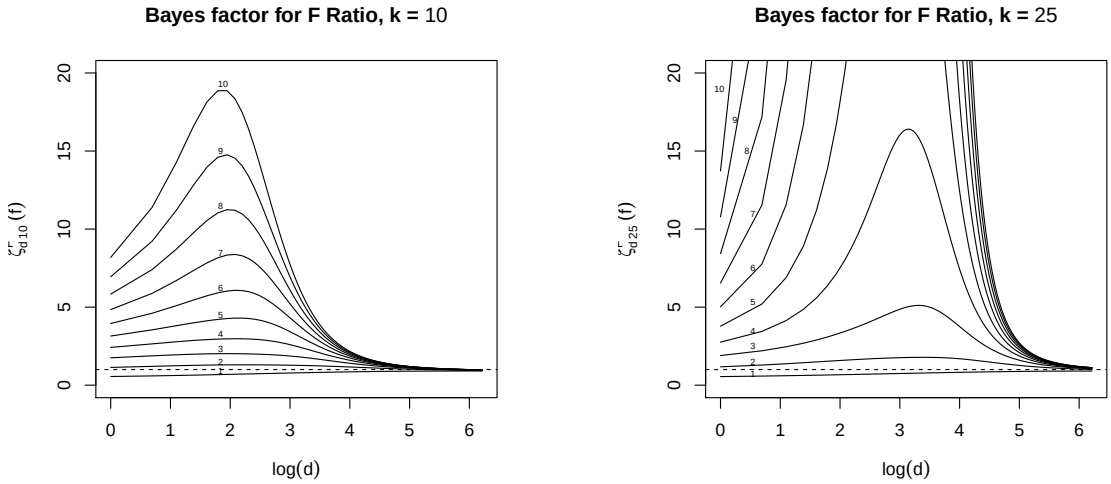


Figure 9.9: Behavior of the Bayes factor $\zeta_{d,k}^F(f)$ as a function of $\log(d)$, as $d \to \infty$, while k is fixed: $k = 10$ in left panel, $k = 25$ in right panel. Different curves correspond to different argument values f, ranging from 1 to 10.

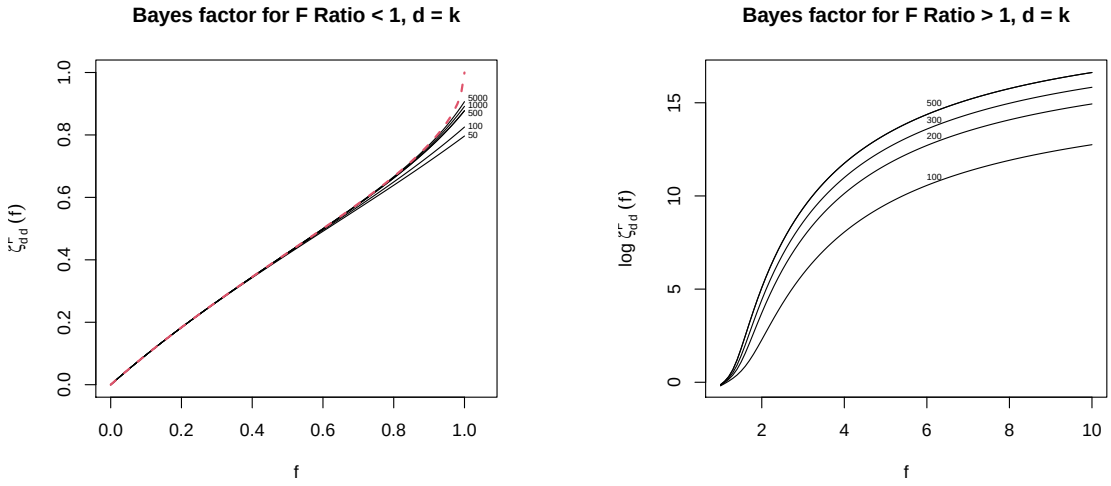


Figure 9.10: Behavior of the Bayes factor $\zeta_{d,k}^F(f)$ as a function of f, when $d = k$ are growing large. Left panel: when $f \leq 1$, $\zeta_{d,d}^F(f)$ converges to $1 - ((1-f)/(1+f))^{\alpha/2}$ (red dashed curve) as $d \to \infty$. Right panel: when $f > 1$, $\zeta_{d,d}^F(f)$ diverges as $d \to \infty$.

9.7 Illustrative example 2

To illustrate how it works in practice, we apply our extended version of the vector-sparse theory for linear regression to a genetic dataset. This dataset came from the department of Neurobiology of the University of Chicago, to the Statistics department, through the

consulting program. It contains the genetic expression level corresponding to 770 genes, recorded during a laboratory experiment on 21 five-week-old male transgenic mice. Each mouse belonged to one of three possible genotype groups, named huPS1WT, huPS1ΔE9 and huPS1M146L. The experiment was designed in such a way that, within each genotype group, approximately half of the mice were randomly assigned to a treatment, while the other half was not to receive the treatment and served as a control group. The treated mice were housed for one month in a so called 'enriched environment', consisting of large cages containing running wheels, tunnels, toys, and chewable materials. Control groups of animals were instead maintained in standard laboratory housing conditions.

Given the structure of the data, we model the expression level for each gene separately and consider the genotype as a covariate whose impact on the genes expression level is not necessarily expected to be negligible. By contrast, we look at the treatment effect and its interaction with the genotype as potentially not relevant.

Thus, for $g = 1, \dots, 770$, we assume

$$Y_g = \mu_{0,g} + X\beta_g + \eta_g \,.$$

Here $Y_g \in \mathbb{R}^{21}$ is the vector of the expression levels of gene g as recorded on the 21 mice; $\mu_{0,g} = X_0\beta_{0,g}$ is the additive effect of the matrix $X_0 = [\mathbf{1} \;\; G_1 \;\; G_2] \in \mathbb{R}^{21\times 3}$ spanning $\mathcal{X}_0$, the space of the genotype classification factor; $X_1 = [T \;\; G_1 * T \;\; G_2 * T] \in \mathbb{R}^{21\times 3}$ is the matrix containing the treatment factor and its interactions with the genotype factor. Yet, instead of X_1, we consider $X = LX_1$ which is the matrix X_1 after being projected onto the kernel of X_0. Really, $L \in \mathbb{R}^{21\times 3}$ is any full column rank matrix whose column span is a subspace of $\mathrm{Ker}(X_0) = \mathcal{X}_0^{\perp}$. This implies that X has a column span $\mathcal{X} \subset \mathcal{X}_0^{\perp}$, so that the effects of X refers solely to the treatment and interaction factors, as these are depurated of their

covariation with the genotype factor.

Concerning distributional assumptions, the errors $\eta_{g,i}$, $i = 1, \ldots, 21$, are independent Gaussian with unknown gene-specific variance σ_g^2,

$$\eta_g \sim N(0, \sigma_g^2 I_{21}) \, .$$

On the other hand, to reflect the expectation that the treatment and its interaction with genotype, might not have a relevant effect on the gene expression level, we assume that

$$\beta_g \sim P_{\nu,3} \, ,$$

where $P_{\nu,3}$ is a three-dimensional vector-sparse distribution, rotationally invariant with respect to the inner product $\langle u, v \rangle_A = u' X' X v / \sigma_g^2$. Letting ρ be the sparsity rate, we further assume that the radial exceedance measure for $\|\beta_g\|_A$ is the inverse-square measure, so that the three-dimensional exceedance measure for $P_{\nu,3}$ can be written as

$$H_{3,A}(dx) = \frac{1}{\text{Area}(\mathcal{S}_A^3)} \frac{\sqrt{2}}{\sqrt{\pi}} \|x\|_A^{-1-3} \, dx \, ,$$

where $\mathcal{S}_A^3 = \{z \in \mathbb{R}^3 : \|z\|_A^2 = z' X' X z / \sigma_g^2 = 1\}$ is the unit sphere with respect to $\langle \cdot, \cdot \rangle_A$.

Notice that

$$H_{3,A}(dx) = \frac{1}{\text{Area}(\mathcal{S}^3)} \frac{\sqrt{2}}{\sqrt{\pi}} \|A'x\|^{-1-3} \det(A) dx = H_3(d(A'x))$$

where H_3 is the inverse-square measure which is rotationally invariant with respect to $\langle \cdot, \cdot \rangle_2$. For this choice of radial measure, H_3 is proportional to the Lévy measure of the symmetric α-stable ($S\alpha S$) process generated by the three-dimensional Cauchy distribution. Therefore, we

can consider the scaled version of this distribution as a possibility for the sparse distribution $P_{\nu,3}$,

$$\begin{aligned}\text{Cauchy}_{\nu,3}(dx) &= \frac{\Gamma(\frac{3+1}{2})}{\sqrt{\pi}\,\pi^{\frac{3}{2}}} \frac{\nu\, dx}{(\|x\|_A^2 + \nu^2)^{\frac{3+1}{2}}} \\ &= \frac{d\tilde{x}}{\text{Area}(\mathcal{S}_A^3)} \frac{2\Gamma(\frac{3+1}{2})}{\sqrt{\pi}\,\Gamma(\frac{3}{2})} \frac{\nu\ \|x\|_A^{d-1}}{(\|x\|_A^2 + \nu^2)^{\frac{3+1}{2}}} d\,\|x\|_A\ ,\end{aligned} \tag{9.16}$$

where the last expression shows how the density factorizes into the spectral and radial components. For the three-dimensional scaled Cauchy, the sparsity rate is

$$\rho = \frac{\sqrt{2}}{\sqrt{\pi}}\nu \cdot \frac{\sqrt{\pi}\Gamma(3/2+1/2)}{\Gamma(3/2)}\,,$$

so that the one-dimensional rate $\frac{\sqrt{2}}{\sqrt{\pi}}\nu$ gets scaled by $\frac{\sqrt{\pi}\Gamma(3/2+1/2)}{\Gamma(3/2)} = 2$.

The aim of the analysis is to obtain, for each gene g, a sparse approximation to the exceedance conditional probability of the parameter F-ratio F_{β_g}, given the observed value of the observation F-ratio F_{Y_g}

$$\mathbb{P}(F_{\beta_g} > \epsilon \mid F_{Y_g} = f_{y_g})\,.$$

To this end, we start by estimating, for each gene, the noise variance σ_g^2 with the residual sum of squares

$$s_g^2 = \frac{\|Q_{\mathcal{X}\cup\mathcal{X}_0} Y_g\|^2}{k}\,,$$

where $k = 21 - 6 = 15$. Then we estimate by maximum likelihood, the sparsity rate ρ, which is assumed to be common to all genes. This is done by considering the sparse approximation to the marginal density of each F-ratio F_{Y_g},

$$m_{3,21}(f_y) = \mathcal{F}_{3,21}(f_y)\ \Big(1 - \rho + \rho\,\zeta_{3,21}\big(\sqrt{3f_y}\,\big)\Big) + o(\rho)\,,$$

and maximizing the log likelihood

$$\max_{\rho} \sum_{g} \log\left(1 - \rho + \rho\, \zeta_{3,21}\big(\sqrt{3 f_{y_g}}\big)\right) .$$

The result of this maximization gives the estimate $\hat{\rho} = 0.348$, which corresponds to $\hat{\nu} = 0.218$, and a one-dimensional rate of approximately 0.17.

We check this estimation by assuming that the sparse distribution for β_g is (9.16), and compute (numerically) the exact distribution for F_{Y_g}. The estimated ν, found by maximumum likelihood, is $\hat{\nu} = 0.218$, which leads to $\hat{\rho} = 0.347$. These estimates are very close to those obtained using the sparse approximation.

Because the estimated sparsity rate is not that small, instead of invoking the double limit regime, we numerically integrate (9.14), to compute the sparse approximation to

$$\mathbb{P}(F_{\beta_g} > \epsilon \mid F_{Y_g} = f_y) .$$

We choose $\epsilon = 0.945$ so that the unconditional probability is equal to the estimated sparsity rate

$$\mathbb{P}(F_{\beta_g} > \epsilon) = \hat{\rho} .$$

Figure 9.11 shows $\mathbb{P}(F_{\beta_g} > 0.945 \mid F_{Y_g} = f_y)$ for both the sparse approximated model (black curve) and the exact scale Cauchy model (green curve). Even though the first one is only based on the exceedance measure and the estimated sparsity rate, while the second one is derived from a fully specified model, the two conditional probabilities are quite close to each other. Interpreting the sparsity rate $\hat{\rho} = 0.347$, shown by the dashed horizontal line, as the probability of the event $F_{\beta_g} > 0.945$ before seeing the data, then this unconditional probability is overcome by the corresponding conditional probability when the observed value of F_{Y_g} is greater than four.

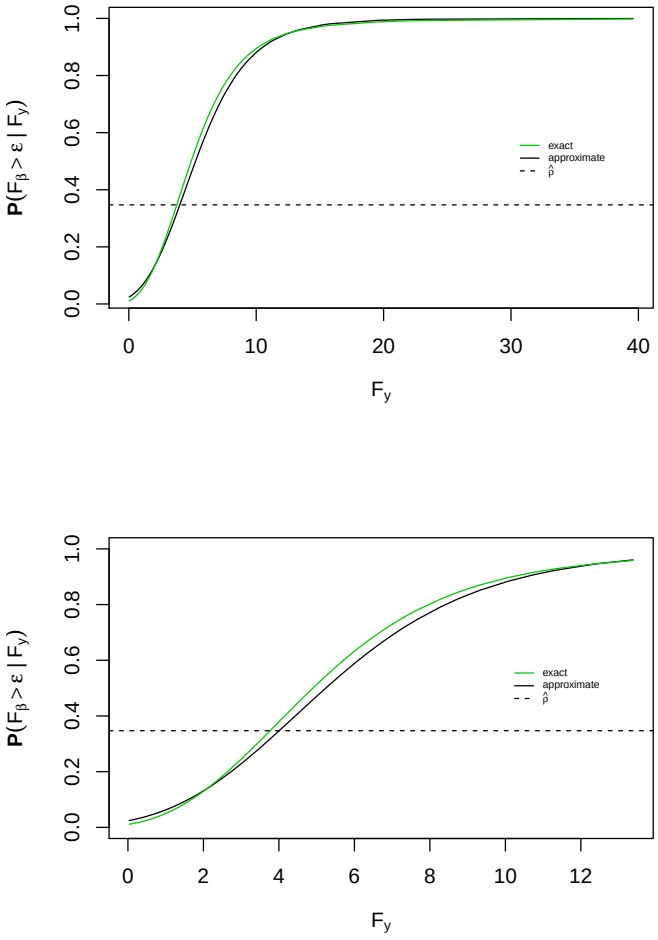


Figure 9.11: Conditional probability of $F_{\beta_g} > 0.945$ as a function of the observed value of F_{Y_g}. The green curves correspond to the exceedance probability obtained when P_ν is the scaled three-dimensional Cauchy. The black curves, instead, correspond to to the exceedance probability obtained under the sparse approximation. The dashed line shows $\hat{\rho}$ which is set to match the unconditional probability that $F_{\beta_g} > 0.945$.

9.8 Appendix

1. We here show that $\psi_{d,X}(y) = \phi_{n,\sigma}(y)\zeta_d((X'X)^{-1}X'y)$ is a probability density. Indeed, it is non negative for any y and it integrates to one. Indeed,

$$\int_{\mathbb{R}^d} \phi_{n,\sigma}(y)\zeta_d((X'X)^{-1}X'y)dy =$$
$$\int_{\mathbb{R}^d} \phi_{n,\sigma}(y) \int_0^\infty (\cosh_d(r(X'X)^{-1}X'y) - 1)e^{-r^2/2}\ H^R(dr)dy =$$
$$\int_0^\infty \int_{\mathbb{R}^d} \phi_{n,\sigma}(y) \int_{\mathcal{S}_A^d} (e^{r\langle (X'X)^{-1}X'y, \tilde{\beta}\rangle_A} - 1)\frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ dy\ e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty \int_{\mathcal{S}_A^d} \int_{\mathbb{R}^d} (\phi_{n,\sigma}(y)e^{ry'X(X'X)^{-1}AA'\tilde{\beta}} - \phi_{n,\sigma}(y))\, dy\ \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty \int_{\mathcal{S}_A^d} \int_{\mathbb{R}^d} (\phi_{n,\sigma}(y)e^{ry'X\tilde{\beta}/\sigma^2} - \phi_{n,\sigma}(y))\, dy\ \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty \int_{\mathcal{S}_A^d} (e^{r^2\tilde{\beta}'X'X\tilde{\beta}/2\sigma^2} - 1)\ \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty \int_{\mathcal{S}_A^d} (e^{r^2\langle\tilde{\beta},\tilde{\beta}\rangle_A/2} - 1)\ \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty \left(\int_{\mathcal{S}_A^d} e^{r^2/2}\ \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}_A^d)}\ - 1 \right) e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty \left(e^{r^2/2}\ - 1 \right) e^{-r^2/2}\ H^R(dr) =$$
$$\int_0^\infty (1 - e^{-r^2/2})\, H^R(dr) = 1$$

2. Here we derive the sparse approximation to the density function of $\|P_{\mathcal{X}}Y\|^2$, Eq. (9.11). Since $\sigma^2 = 1$, we can write the sparse approximation to the density of Y as

$$\phi_n(y)\,(1 - \rho + \rho\zeta_d(\|P_{\mathcal{X}}y\|)) + o(\rho)\,.$$

On the other hand, given β

$$Y \mid \beta \sim N(X\beta, I_n)\,,$$

so one has that

$$
\begin{aligned}
&\mathbb{P}\left(Y \in dy \mid \|\beta\|_A \in dr\right) = \\
&\int_{\mathcal{S}^d} \mathbb{P}\left(Y \in dy, \tilde{\beta} \in d\tilde{\beta} \mid \|\beta\|_A \in dr\right) = \\
&\int_{\mathcal{S}^d} \mathbb{P}\left(Y \in dy \mid \beta \in d(r\tilde{\beta})\right) \mathbb{P}\left(\tilde{\beta} \in d\tilde{\beta} \mid \|\beta\|_A \in dr\right) = \\
&\int_{\mathcal{S}^d} \phi_n(y - Xr\tilde{\beta}/\sigma) \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}^d)} = \\
&\int_{\mathcal{S}^d} \frac{1}{\sqrt{2\pi}^n} e^{-\|y\|^2/2} e^{y'Xr\tilde{\beta}} e^{-r^2\left\|X\tilde{\beta}\right\|^2/2} \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}^d)} = \\
&\frac{1}{\sqrt{2\pi}^n} e^{-\|y\|^2/2} \int_{\mathcal{S}^d} e^{y'Xr\tilde{\beta}} e^{-r^2\left\|\tilde{\beta}\right\|_A^2/2} \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}^d)} = \\
&\frac{1}{\sqrt{2\pi}^n} e^{-\|y\|^2/2} e^{-r^2/2} \int_{\mathcal{S}^d} e^{y'Xr\tilde{\beta}} \frac{d\tilde{\beta}}{\text{Area}(\mathcal{S}^d)} = \\
&\frac{1}{\sqrt{2\pi}^n} e^{-\|y\|^2/2} e^{-r^2/2} \cosh_d(r \left\|X'y\right\|_A) = \\
&\frac{1}{\sqrt{2\pi}^n} e^{-\|y\|^2/2} e^{-r^2/2} \cosh_d(r \left\|P_{\mathcal{X}} y\right\|) = \\
&\frac{1}{\sqrt{2\pi}^d} e^{-\|P_{\mathcal{X}} y\|^2/2} e^{-r^2/2} \cosh_d(r \left\|P_{\mathcal{X}} y\right\|) \frac{1}{\sqrt{2\pi}^{n-d}} e^{-\|Q_{\mathcal{X}} y\|^2/2} = \\
&\mathbb{P}\left(P_{\mathcal{X}} Y \in dP_{\mathcal{X}} y \mid \|\beta\|_A \in dr\right) \cdot \mathbb{P}\left(Q_{\mathcal{X}} Y \in dQ_{\mathcal{X}} y\right) .
\end{aligned}
$$

From this last expression, we can integrate over the $d-1$ spherical coordinates of $P_{\mathcal{X}} Y$, $\phi_1, \ldots, \phi_{d-1}$, to obtain that the density of $\|P_{\mathcal{X}} Y\|^2$ at p^2, given $\|\beta\|_A$, is

$$
\begin{aligned}
&\mathbb{P}\left(\|P_{\mathcal{X}} Y\|^2 \in dp^2 \mid \|\beta\|_A \in dr\right) = \\
&\int_{\mathcal{S}^d} \mathbb{P}\left(\|P_{\mathcal{X}} Y\|^2 \in dp^2, \phi_1 \in d\phi_1, \cdots, \phi_{d-1} \in d\phi_{d-1} \mid \|\beta\|_A \in dr\right) = \\
&\frac{1}{\sqrt{2\pi}^d} e^{-(p^2+r^2)/2} \cosh_d(rp)\, p^{d-1} \int_{\mathcal{S}^d} \sin^{d-2}\phi_1 \ldots \sin\phi_{d-2}\, d\phi_1 \ldots d\phi_{d-1} = \\
&\frac{1}{\sqrt{2\pi}^d}\, e^{-(p^2+r^2)/2} \cosh_d(rp)\ (p^2)^{\frac{d}{2}-1} \frac{\text{Area}(\mathcal{S}^d)}{2} = \\
&\chi^2_d(p^2) \cosh_d(rp)\ e^{-r^2/2}.
\end{aligned}
$$

So now to get the unconditional density, it suffices to integrate over the sparse radial measure,

$$\begin{aligned}
\mathbb{P}\left(\|P_{\mathcal{X}}Y\|^2 \in dp^2\right) &= \int \mathbb{P}\left(\|P_{\mathcal{X}}Y\|^2 \in dp^2 \mid \|\beta\|_A \in dr\right) \mathbb{P}\left(\|\beta\|_A \in dr\right) \\
&= \int \chi_d^2(p^2) \cosh_d(rp)\ e^{-r^2/2}\, P_\nu^R(dr) \\
&= \chi_d^2(p^2) \left(\int (\cosh_d(rp) - 1)\, e^{-r^2/2}\, P_\nu^R(dr) + 1 - \int (1 - e^{-r^2/2})\, P_\nu^R(dr) \right) \\
&= \chi_d^2(p^2) \left(\rho \int (\cosh_d(rp) - 1)\, e^{-r^2/2}\, H^R(dr) + 1 - \rho \right) \\
&= \chi_d^2(p^2) \left(\rho \zeta_d(p) + 1 - \rho \right) .
\end{aligned}$$

This last expression is the sparse approximation of order ρ to the density of $\|P_{\mathcal{X}}Y\|^2$ as shown in (9.11).

3. Here we derive the sparse approximation to the density function of F_Y, Eq. (9.12). Let $P = \|P_{\mathcal{X}}Y\|^2 / d$ and $Q = \|Q_{\mathcal{X}}Y\|^2 / k$, where dP has distribution given in (9.11) and is independent of

$$kQ \sim \chi_k^2 .$$

Then the marginal distribution of $F_Y = P/Q$ can be found by

$$\begin{aligned}
\mathbb{P}(P/Q \in \partial f_y) &= \int \mathbb{P}\left(P/(kQ) \in \partial(f_y/k),\, kQ \in \partial w\right) 1/k\, \partial w \\
&= \int \mathbb{P}\left((dP)/(kQ) \in \partial(d\, f_y/k),\, kQ \in \partial w\right) d/k\, \partial w \\
&= \int \mathbb{P}\left((dP) \in \partial(f_y\, w\, d/k),\, \right) \mathbb{P}\left(kQ \in \partial w\right) w\, d/k\, \partial w \\
&= \int \mathbb{P}\left((dP) \in \partial(f_y\, u),\, \right) \mathbb{P}\left(kQ \in \partial(u\, k/d)\right) u\, k/d\, \partial u \\
&= \int \chi_d^2(uf_y)(1 - \rho + \rho\zeta_d(\sqrt{uf_y}))\, \chi_k^2\left(\frac{k}{d}u\right) \frac{k}{d}\, u\, \partial u ,
\end{aligned}$$

where we used the notation ∂x instead of dx to denote the differential form, to avoid notational confusion with the d indicating the dimension of the vector.

Substituting the analytic form for the chi-square density function and the Taylor series for the $\cosh_d$ function, the last integral can be developed as follows

$$\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}(f_y)^{\frac{d}{2}-1}\left(\frac{k}{d}\right)^{\frac{k}{2}-1}\cdot$$
$$\cdot\int u^{\frac{d}{2}-1+\frac{k}{2}-1}e^{-u\frac{df_y+k}{2d}}\left(1-\rho+\rho\sum_{r=1}^{\infty}\frac{(uf_y)^r}{(d/2)^{\uparrow r}2^r r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\right)\frac{k}{d}\,u\,\partial u=$$

$$\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}(f_y)^{\frac{d}{2}-1}\left(\frac{k}{d}\right)^{\frac{k}{2}}\cdot$$
$$\cdot\int u^{\frac{d}{2}+\frac{k}{2}-1}e^{-u\frac{df_y+k}{2d}}\left(1-\rho+\rho\sum_{r=1}^{\infty}\frac{(uf_y)^r}{(d/2)^{\uparrow r}2^r r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\right)\partial u=$$

$$\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}(f_y)^{\frac{d}{2}-1}\left(\frac{k}{d}\right)^{\frac{k}{2}}\frac{\Gamma(\frac{d}{2}+\frac{k}{2})(2d)^{\frac{d}{2}+\frac{k}{2}}}{(df_y+k)^{\frac{d}{2}+\frac{k}{2}}}\cdot$$
$$\cdot\left(1-\rho+\rho\sum_{r=1}^{\infty}\frac{(df_y)^r}{(d/2)^{\uparrow r}2^r r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+r)2^r}{\Gamma(\frac{d}{2}+\frac{k}{2})(df_y+k)^r}\right)=$$

$$\frac{1}{Be(\frac{d}{2},\frac{k}{2})}\frac{(f_y)^{\frac{d}{2}-1}d^{\frac{d}{2}}k^{\frac{k}{2}}}{(df_y+k)^{\frac{d}{2}+\frac{k}{2}}}\cdot\left(1-\rho+\rho\sum_{r=1}^{\infty}\frac{(df_y)^r}{(df_y+k)^r r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+r)}{\Gamma(\frac{d}{2}+\frac{k}{2})(d/2)^{\uparrow r}}\right)=$$

$$\mathcal{F}_{d,k}(f_y)\left(1-\rho+\rho\sum_{r=1}^{\infty}\frac{(df_y)^r}{(df_y+k)^r}\frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow r}}{(d/2)^{\uparrow r}r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\right)=$$

$$\mathcal{F}_{d,k}(f_y)\left(1-\rho+\rho\,\zeta^F_{d,k}(f_y)\right).$$

This last expression is the sparse approximation of order ρ to the density of F_Y as shown in (9.12).

4. Here we show that as $k \to \infty$, (9.12) converges to

$$\chi^2_d(df_y)\,(1-\rho+\rho\,\zeta_d(f_y))$$

Now, as $k \to \infty$, Fisher's $F_{d,k}$ density converges to a scaled chi-square χ^2_d,

$$\mathcal{F}_{d,k}(f_y) \to \chi^2_d(df_y)\,,$$

so we really just need to show that

$$\zeta^F_{d,k}(f_y) \to \zeta_d\big(\sqrt{df_y}\,\big)\,.$$

To see this, notice that, for each r,

$$\frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow r}}{(df_y+k)^r} = \frac{\Gamma(\frac{d}{2}+\frac{k}{2}+r)2^{-r}}{\Gamma(\frac{d}{2}+\frac{k}{2})\left(\frac{df_y+k}{2}\right)^r} \to 2^{-r}$$

so then we have that, for each r,

$$\frac{(df_y)^r}{(df_y+k)^r}\frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow r}}{(d/2)^{\uparrow r}r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)} \to \frac{(df_y)^r}{(d/2)^{\uparrow r}2^r r!}\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}$$

Now because the convergence is monotone, we can pass the limit inside the infinite sum, and conclude that

$$\lim_{k\to\infty}\zeta^F_{d,k}(f_y) = \zeta_d\big(\sqrt{df_y}\,\big)\,.$$

5. Here we derive the sparse approximation to the joint density function of (F_β, F_Y). Again, let $P = \|P_{\mathcal{X}} Y\|^2 / d$ and $Q = \|Q_{\mathcal{X}} Y\|^2 / k$, and let $b = \|\beta\|_A^2 / d$. Also, denote by $p_\nu^{R^2}$ and h^{R^2}, the sparse density and corresponding exceedance density for the squared norm $\|\beta\|_A^2$. Then

$$\mathbb{P}\left(F_Y \in \partial f_y,\, F_\beta \in \partial f_\beta, \right) =$$

$$\int \mathbb{P}\left(\frac{P}{Q} \in \partial f_y,\, \frac{b}{Q} \in \partial f_\beta,\, Q \in \partial u\right) \partial u =$$

$$\int \mathbb{P}\left(P \in \partial(u f_y), \mid b \in \partial(u f_\beta)\right)\, \mathbb{P}(Q \in \partial u)\, \mathbb{P}(b \in \partial(u f_\beta))\; u^2 \partial u =$$

$$\int \chi_d^2(u f_y) \cosh_d\left(\sqrt{u f_y}\sqrt{u f_\beta}\right) e^{-u f_\beta}\, \chi_k^2\left(\frac{k}{d}u\right) \frac{k}{d}\, u^2\, p_\nu^{R^2}(u\, f_\beta)\, \partial u \;=$$

$$\rho \int \chi_d^2(u f_y) \cosh_d\left(\sqrt{u f_y}\sqrt{u f_\beta}\right) e^{-u f_\beta}\, \chi_k^2\left(\frac{k}{d}u\right) \frac{k}{d}\, u^2\, h^{R^2}(u\, f_\beta)\, \partial u \;=$$

$$\rho \int \chi_d^2(u f_y) \cosh_d\left(\sqrt{u f_y}\sqrt{u f_\beta}\right) e^{-u f_\beta}\, \chi_k^2\left(\frac{k}{d}u\right) \frac{k}{d}\, u^2\, \frac{\partial u}{(u f_\beta)^{\alpha/2+1}} K_\alpha =$$

$$\rho \int \chi_d^2(u f_y) \cosh_d\left(\sqrt{u f_y}\sqrt{u f_\beta}\right) e^{-u f_\beta}\, \chi_k^2\left(\frac{k}{d}u\right) \frac{k}{d}\, u^{2-\alpha/2-1} \partial u \cdot h^{R^2}(f_\beta)\,.$$

Again, by substituting the analytic form for the chi-square density function and the Taylor

series for the $\cosh_d$ function, the last integral can be developed as follows

$$\rho h^{R^2}(f_\beta)\,\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}\frac{f_y^{\frac{d}{2}-1}k^{\frac{k}{2}-1}}{d^{\frac{k}{2}-1}}\int u^{\frac{d}{2}-1+\frac{k}{2}-1+2-\frac{\alpha}{2}-1}e^{-u(f_y+f_\beta+\frac{k}{d})/2}\sum_{r=0}^{\infty}\frac{(uf_y)^r(uf_\beta)^r}{(d/2)^{\uparrow r}2^{2r}r!}\frac{k}{d}\,\partial u =$$

$$\rho h^{R^2}(f_\beta)\,\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}\frac{f_y^{\frac{d}{2}-1}k^{\frac{k}{2}-1}}{d^{\frac{k}{2}-1}}\int u^{\frac{d}{2}+\frac{k}{2}-\frac{\alpha}{2}-1}e^{-u\frac{df_y+df_\beta+k}{2d}}\sum_{r=0}^{\infty}\frac{(uf_y)^r(uf_\beta)^r}{(d/2)^{\uparrow r}2^{2r}r!}\,\partial u =$$

$$\rho h^{R^2}(f_\beta)\,\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}f_y^{\frac{d}{2}-1}\left(\frac{k}{d}\right)^{\frac{k}{2}}\sum_{r=0}^{\infty}\frac{f_y^r f_\beta^r}{(d/2)^{\uparrow r}2^{2r}r!}\int u^{2r+\frac{d}{2}+\frac{k}{2}-\frac{\alpha}{2}-1}e^{-u\frac{df_y+df_\beta+k}{2d}}\partial u =$$

$$\rho h^{R^2}(f_\beta)\,\frac{(1/2)^{\frac{d}{2}+\frac{k}{2}}}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}\frac{f_y^{\frac{d}{2}-1}k^{\frac{k}{2}-1}}{d^{\frac{k}{2}-1}}\sum_{r=0}^{\infty}\frac{f_y^r f_\beta^r}{(d/2)^{\uparrow r}2^{2r}r!}\frac{\Gamma(2r+\frac{d}{2}+\frac{k}{2}-\frac{\alpha}{2})(2d)^{2r+\frac{d}{2}+\frac{k}{2}-\frac{\alpha}{2}}}{(df_y+df_\beta+k)^{2r+\frac{d}{2}+\frac{k}{2}-\frac{\alpha}{2}}} =$$

$$\rho h^{R^2}(f_\beta)\,\frac{\Gamma(\frac{d}{2}+\frac{k}{2})}{\Gamma(\frac{d}{2})\Gamma(\frac{k}{2})}\frac{f_y^{\frac{d}{2}-1}k^{\frac{k}{2}-1}d^{\frac{d}{2}+1}}{(df_y+df_\beta+k)^{\frac{d}{2}+\frac{k}{2}}}\sum_{r=0}^{\infty}\frac{f_y^r f_\beta^r}{(d/2)^{\uparrow r}2^{2r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})(2d)^{2r-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y+df_\beta+k)^{2r-\frac{\alpha}{2}}} =$$

$$\rho h^{R^2}(f_\beta)\,\frac{1}{Be(\frac{d}{2},\frac{k}{2})}\frac{f_y^{\frac{d}{2}-1}d^{\frac{d}{2}}k^{\frac{k}{2}}}{(df_y+df_\beta+k)^{\frac{d}{2}+\frac{k}{2}}}d^{-\frac{\alpha}{2}}\sum_{r=0}^{\infty}\frac{(df_y)^r(df_\beta)^r}{(d/2)^{\uparrow r}2^{2r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{2r-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y+df_\beta+k)^{2r-\frac{\alpha}{2}}} =$$

$$\rho H^{R^2}(\partial(df_\beta))\,\mathcal{F}_{d,k}(f_y)\left(\frac{df_y+k}{df_y+k+df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}}\sum_{r=0}^{\infty}\frac{(df_y)^r(df_\beta)^r}{(d/2)^{\uparrow r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y+df_\beta+k)^{2r-\frac{\alpha}{2}}}\,.$$

Recalling that $P_\nu^{R^2}(\partial u) = \rho H^{R^2}(\partial u) + o(\rho)$, in the sense of integrals against functions in $\mathcal{W}^{\#}$, we can split the summation in two addends:

$$\mathcal{F}_{d,k}(f_y)P_\nu^{R^2}\left(\partial(df_\beta)\right)\left(\frac{df_y+k}{df_y+k+df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}}\frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow-\frac{\alpha}{2}}2^{-\frac{\alpha}{2}}}{(df_y+df_\beta+k)^{-\frac{\alpha}{2}}}\,,$$

and

$$\mathcal{F}_{d,k}(f_y)\rho H^{R^2}(\partial(df_\beta))\,\left(\frac{df_y+k}{df_y+k+df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}}\sum_{r=1}^{\infty}\frac{(df_y)^r(df_\beta)^r}{(d/2)^{\uparrow r}r!}\,\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y+df_\beta+k)^{2r-\frac{\alpha}{2}}}\,.$$

The sum of these two last expressions is the sparse approximation of order ρ to the joint density of (F_β, F_Y).

We can also write it in a more compact form,

$$\mathcal{F}_{d,k}(f_y)\left(P_\nu^{R^2}\left(\partial(df_\beta)\right)w_{d,k}(f_y,f_\beta)\;+\rho\zeta_{d,k}^F(\partial(df_\beta);df_y)\right) \tag{9.17}$$

where

$$w_{d,k}(f_y,f_\beta)=\left(\frac{df_y+k}{df_y+k+df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}}\frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow-\frac{\alpha}{2}}2^{-\frac{\alpha}{2}}}{(df_y+df_\beta+k)^{-\frac{\alpha}{2}}}$$

while

$$\zeta_{d,k}^F(\partial(df_\beta);df_y)=H^{R^2}(\partial(df_\beta))\,\left(\frac{df_y+k}{df_y+k+df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}}\cdot$$
$$\cdot\sum_{r=1}^{\infty}\frac{(df_y)^r(df_\beta)^r}{(d/2)^{\uparrow r}r!}\,\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y+df_\beta+k)^{2r-\frac{\alpha}{2}}}$$

is the zeta measure for F-ratios on d and k degrees of freedom.

6. Here we check that

$$\int\zeta_{d,k}^F(\partial(df_\beta);df_y)\partial(df_\beta)=\zeta_{d,k}^F(f_y)\,.$$

In fact, the RHS can be computed as

$$\int H^{R^2}(\partial(df_\beta))\left(\frac{df_y+k}{df_y+k+df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}}\sum_{r=1}^{\infty}\frac{(df_y)^r(df_\beta)^r}{(d/2)^{\uparrow r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})\,(df_y+df_\beta+k)^{2r-\frac{\alpha}{2}}}=$$
$$\sum_{r=1}^{\infty}\frac{(df_y)^r(df_y+k)^{\frac{d}{2}+\frac{k}{2}}}{(d/2)^{\uparrow r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})}\int\frac{(df_\beta)^r}{(df_y+k+df_\beta)^{\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2}}}H^{R^2}(\partial(df_\beta))=$$
$$\sum_{r=1}^{\infty}\frac{(df_y)^r(df_y+k)^{\frac{d}{2}+\frac{k}{2}}}{(d/2)^{\uparrow r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})}K_\alpha\int\frac{(df_\beta)^{r-\frac{\alpha}{2}-1}}{(df_y+k+df_\beta)^{\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2}}}\partial(df_\beta)\,.$$

We write the integral as

$$\int\frac{z^{r-\frac{\alpha}{2}-1}}{(v+z)^{\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2}}}\partial z=\int\left(\frac{z}{v+z}\right)^{r-\frac{\alpha}{2}-1}\left(\frac{1}{v+z}\right)^{\frac{d}{2}+\frac{k}{2}+r+1}\partial z\,.$$

Now make the change of variable $t=z/(v+z)$,

$$\int t^{r-\frac{\alpha}{2}-1}\left(\frac{1-t}{v}\right)^{\frac{d}{2}+\frac{k}{2}+r+1}\frac{v}{(1-t)^2}\,\partial t\;=$$
$$v^{-\frac{d}{2}-\frac{k}{2}-r}\int t^{r-\frac{\alpha}{2}-1}\,(1-t)^{\frac{d}{2}+\frac{k}{2}+r-1}\,\partial t\;=$$
$$v^{-\frac{d}{2}-\frac{k}{2}-r}\frac{\Gamma(r-\frac{\alpha}{2})\Gamma(\frac{d}{2}+\frac{k}{2}+r)}{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})}\,.$$

So plugging this back, we obtain,

$$\sum_{r=1}^{\infty}\frac{(df_y)^r(df_y+k)^{\frac{d}{2}+\frac{k}{2}}}{(d/2)^{\uparrow r}r!}\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})}K_\alpha\frac{\Gamma(r-\frac{\alpha}{2})(df_y+k)^{-\frac{d}{2}-\frac{k}{2}-r}\Gamma(\frac{d}{2}+\frac{k}{2}+r)}{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})}=$$
$$\sum_{r=1}^{\infty}\frac{(df_y)^r}{(df_y+k)^r(d/2)^{\uparrow r}r!}2^{-\frac{\alpha}{2}}\,\frac{\alpha 2^{\frac{\alpha}{2}-1}}{\Gamma(1-\alpha/2)}\,\frac{\Gamma(r-\frac{\alpha}{2})\Gamma(\frac{d}{2}+\frac{k}{2}+r)}{\Gamma(\frac{d}{2}+\frac{k}{2})}=$$
$$\sum_{r=1}^{\infty}\frac{(df_y)^r}{(df_y+k)^r}\,\frac{\alpha\Gamma(r-\alpha/2)}{2\Gamma(1-\alpha/2)}\,\frac{(\frac{d}{2}+\frac{k}{2})^{\uparrow r}}{(d/2)^{\uparrow r}r!}\,,$$

which indeed is the same expression we have for $\zeta^F_{d,k}(f_y)$.

7. Here we show the convergence of (9.17), as $k \to \infty$. Start by noticing that, as $k \to \infty$, Fisher's $F_{d,k}$ density converges to a scaled chi-square χ^2_d,

$$\mathcal{F}_{d,k}(f_y) \to \chi^2_d(df_y)\,,$$

while

$$\left(\frac{df_y + k}{df_y + k + df_\beta}\right)^{\frac{d}{2}+\frac{k}{2}} \to e^{-df_\beta/2}\,.$$

On the other hand, as $k \to \infty$, both

$$\frac{\Gamma(\frac{d}{2}+\frac{k}{2}-\frac{\alpha}{2})\,2^{-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})(d(f_y+f_\beta)+k)^{-\frac{\alpha}{2}}} \to 1\,,$$

and

$$\frac{\Gamma(\frac{d}{2}+\frac{k}{2}+2r-\frac{\alpha}{2})\,2^{2r-\frac{\alpha}{2}}}{\Gamma(\frac{d}{2}+\frac{k}{2})(d(f_y+f_\beta)+k)^{2r-\frac{\alpha}{2}}} \to 1\,.$$

So as $k \to \infty$,

$$w_{d,k}(f_y, f_\beta) \to e^{-df_\beta/2}$$

while

$$\zeta^F_{d,k}(\partial(df_\beta); df_y) \to H^{R^2}(\partial(df_\beta))\, e^{-df_\beta/2} \sum_{r=1}^{\infty} \frac{(df_y)^r (df_\beta)^r}{(d/2)^{\uparrow r} 2^{2r} r!}\,.$$

Thus, the sparse approximation of the joint density of (F_β, F_y) converges to

$$\begin{aligned}
&\chi^2_d(df_y) e^{-df_\beta/2}\left(P^{R^2}_\nu(\partial(df_\beta)) + \rho H^{R^2}(\partial(df_\beta)) \sum_{r=1}^{\infty} \frac{(df_y\, df_\beta)^r}{(d/2)^{\uparrow r} 2^{2r} r!}\right) = \\
&\chi^2_d(df_y)\left(e^{-df_\beta/2} P^{R^2}_\nu(\partial(df_\beta)) + \rho(\cosh_d(\sqrt{df_y}\sqrt{df_\beta}) - 1) e^{-df_\beta/2} H^{R^2}(\partial(df_\beta))\right) = \\
&\chi^2_d(df_y)\left(e^{-df_\beta/2} P^{R^2}_\nu(\partial(df_\beta)) + \rho\zeta_d(\partial(df_\beta); df_y)\right)\,.
\end{aligned}$$

From this convergence and the one proved for the marginal density of F_Y, one can easily deduce the analog for the conditional density of F_β given F_Y.

Bibliography

[1] F. Abramovich, T. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):725–749, 1998.

[2] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. Banday, R. Barreiro, N. Bartolo, S. Basak, et al. Planck 2018 results-VI. Cosmological parameters. *Astronomy & Astrophysics*, 641:A6, 2020.

[3] A. Atay-Kayis and H. Massam. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335, 2005.

[4] M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018.

[5] S. Banerjee and S. Ghosal. Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162, 2015.

[6] R. F. Barber, M. Drton, and K. M. Tan. Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data*, pages 15–36. Abel Symposia, vol 11. Springer, 2016.

[7] O. E. Barndorff-Nielsen and F. Hubalek. Probability measures, lévy measures and analyticity in time. *Bernoulli*, 14(3):764–790, 2008.

[8] H. Battey. On sparsity scales and covariance matrix transformations. *Biometrika*, 106 (3):605–617, 2019.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[10] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[11] J. Berger et al. A robust generalized bayes estimator and confidence region for a multivariate normal mean. *Annals of Statistics*, 8(4):716–761, 1980.

[12] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[13] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[14] P. J. Brown and J. E. Griffin. Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis*, 5(1):171–188, 2010.

[15] B. Burwinkel, J. W. Scott, C. Bührer, F. K. Van Landeghem, G. F. Cox, C. J. Wilson, D. G. Hardie, and M. W. Kilimann. Fatal congenital heart glycogenosis caused by a recurrent activating r531q mutation in the γ2-subunit of amp-activated protein kinase (prkag2), not by phosphorylase kinase deficiency. *The American Journal of Human Genetics*, 76(6):1034–1049, 2005.

[16] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[17] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch. Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92(440):1413–1421, 1997.

[18] M. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2):391–401, 1998.

[19] R. R. Coifman and D. L. Donoho. Translation-invariant de-noising. In *Wavelets and statistics*, pages 125–150. Springer, 1995.

[20] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

[21] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21(3):1272–1317, 1993.

[22] D. L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and computational harmonic analysis*, 1(1):100–115, 1993.

[23] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[24] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67, 1992.

[25] M. L. Eaton. Group invariance applications in statistics. In *Regional conference series in Probability and Statistics*, pages i–133. JSTOR, 1989.

[26] B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[27] B. Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, 2007.

[28] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012.

[29] B. Efron and R. Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

[30] B. Efron, J. D. Storey, and R. Tibshirani. *Microarrays empirical Bayes methods, and false discovery rates*. Department of Statistics, Stanford University, 2001.

[31] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[32] N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36(6):2757–2790, 2008.

[33] K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. Chapman and Hall, 1990.

[34] S. M. Farabaugh, D. N. Boone, and A. V. Lee. Role of igf1r in breast cancer subtypes, stemness, and lineage differentiation. *Frontiers in endocrinology*, 6:59, 2015.

[35] W. Feller. *An Introduction to Probability Theory and Its Application: Vol. 1-2*. 1966.

[36] R. A. Fisher. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 121(788):654–673, 1928.

[37] W. L. Freedman, B. F. Madore, D. Hatt, T. J. Hoyt, I. S. Jang, R. L. Beaton, C. R. Burns, M. G. Lee, A. J. Monson, and J. R. Neeley. The Carnegie-Chicago Hubble Program VIII. An independent determination of the Hubble constant based on the tip of the red giant branch. *The Astrophysical Journal*, 882(1):34, 2019.

[38] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[39] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061, 2004.

[40] E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[41] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286 (5439):531–537, 1999.

[42] N. Hinz and M. Jücker. Distinct functions of akt isoforms in breast cancer: a comprehensive review. *Cell Communication and Signaling*, 17(1):1–29, 2019.

[43] D. P. Hollern, M. R. Swiatnicki, J. P. Rennhack, S. A. Misek, B. C. Matson, A. McAuliff, K. A. Gallo, K. M. Caron, and E. R. Andrechek. E2f1 drives breast cancer metastasis by regulating the target gene fgf13 and altering cell migration. *Scientific reports*, 9(1): 1–13, 2019.

[44] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the royal statistical society: series B (statistical methodology)*, 59(2):319–351, 1997.

[45] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):1594–1649, 2004.

[46] I. M. Johnstone and B. W. Silverman. EbayesThresh: R and S-Plus programs for Empirical Bayes thresholding. *Journal of Statistical Software*, 12:1–38, 2005.

[47] I. M. Johnstone and B. W. Silverman. Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, 33(4):1700–1752, 2005.

[48] A. Lenkoski and A. Dobra. Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157, 2011.

[49] C. Liu and R. Martin. An empirical G-Wishart prior for sparse high-dimensional Gaussian graphical models. *arXiv preprint arXiv:1912.03807*, 2019.

[50] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[51] P. McCullagh and N. G. Polson. Statistical sparsity. *Biometrika*, 105(4):797–814, 2018.

[52] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[53] K. R. Monson, M. Goldberg, H.-C. Wu, R. M. Santella, W. K. Chung, and M. B. Terry. Circulating growth factor concentrations and breast cancer risk: a nested case-control study of igf-1, igfbp-3, and breast cancer in a family-based cohort. *Breast Cancer Research*, 22(1):1–5, 2020.

[54] H. Nalwoga, J. B. Arnes, H. Wabinga, and L. A. Akslen. Expression of egfr and c-kit is associated with the basal-like phenotype in breast carcinomas of african women. *Apmis*, 116(6):515–525, 2008.

[55] G. Nason. *Wavelet methods in statistics with R*. Springer Science & Business Media, 2008.

[56] G. Nason. **wavethresh**: Wavelets statistics and transforms, 2016. URL `https://CRAN.R-project.org/package=wavethresh`.

[57] G. Pérez-Tenorio, E. Karlsson, M. A. Waltersson, B. Olsson, B. Holmlund, B. Nordenskjöld, T. Fornander, L. Skoog, and O. Stål. Clinical potential of the mtor targets s6k1 and s6k2 in breast cancer. *Breast cancer research and treatment*, 128(3):713–723, 2011.

[58] A. G. Riess. The expansion of the universe is faster than expected. *Nature Reviews Physics*, 2(1):10–12, 2020.

[59] A. G. Riess, S. Casertano, W. Yuan, L. M. Macri, and D. Scolnic. Large Magellanic Cloud Cepheid standards provide a 1% foundation for the determination of the Hubble constant and stronger evidence for physics beyond ΛCDM. *The Astrophysical Journal*, 876(1):85, 2019.

[60] V. Ročková and E. I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[61] A. Roverato. Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, 87(1):99–112, 2000.

[62] G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes*. Chapman and Hall, 1994.

[63] K.-I. Sato. *Lévy processes and infinitely divisible distributions*. Cambridge University Press, 1999.

[64] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

[65] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[66] M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[67] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[68] J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035, 2003.

[69] W. E. Strawderman. Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):385–388, 1971.

[70] L. Sun and M. Stephens. Solving the empirical Bayes normal means problem with correlated noise. *arXiv preprint arXiv:1812.07488*, 2018.

[71] L. Verde, T. Treu, and A. G. Riess. Tensions between the early and late universe. *Nature Astronomy*, 3:891–895, 2019.

[72] S. M. Villa, J. C. Altuna, J. S. Ruff, A. B. Beach, L. I. Mulvey, E. J. Poole, H. E. Campbell, K. P. Johnson, M. D. Shapiro, S. E. Bush, et al. Rapid experimental evolution of reproductive isolation from a single natural population. *Proceedings of the National Academy of Sciences*, 116(27):13440–13445, 2019.

[73] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis*, 11(1):89–123, 2001.

[74] H. Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

[75] H. Wang and S. Z. Li. Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198, 2012.

[76] E. B. Wilson and M. M. Hilferty. The distribution of chi-square. *proceedings of the National Academy of Sciences of the United States of America*, 17(12):684, 1931.

[77] N. Wolchover. Cosmologists debate how fast the universe is expanding. *Quanta Magazine*, 2019.

[78] N. Wolchover. New wrinkle added to cosmology's Hubble crisis. *Quanta Magazine*, 2020.

[79] Z. Xing, P. Carbonetto, and M. Stephens. Flexible signal denoising via flexible empirical Bayes shrinkage. *arXiv preprint arXiv:1605.07787*, 2016.

[80] Y. Zhang, P. K.-S. Ng, M. Kucherlapati, F. Chen, Y. Liu, Y. H. Tsang, G. de Velasco, K. J. Jeong, R. Akbani, A. Hadjipanayis, et al. A pan-cancer proteogenomic atlas of pi3k/akt/mtor pathway alterations. *Cancer cell*, 31(6):820–832, 2017.