THE UNIVERSITY OF CHICAGO

FLEXIBLE BAYESIAN METHODS FOR HIGH DIMENSIONAL DATA

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

ENAKSHI SAHA

CHICAGO, ILLINOIS

JUNE, 2021

To my loving family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor Professor Veronika Rockova for introducing me to some very interesting topics in Bayesian Statistics. Her constant tenacity and drive for excellence will always be a great source of inspiration. I shall be eternally grateful to Professor Chao Gao for motivating me to be a better researcher and helping me make important career decisions, through his invaluable insight and meticulous feedback. I would also like to thank Professor Matthew Stephens whose kind and generous suggestions have not only improved my dissertation but also made lasting impact on my entire research perspective. Professor Stephens has been incredibly supportive and compassionate throughout the process of completion of my dissertation. In this regard I would also like to thank Professor Yali Amit, who as the Director of Graduate Studies has always been extremely kind and helpful. Without their strong support, this dissertation would never have come to fruition. I am extremely thankful to my amazing collaborator Professor Carlo Graziani who is not only a great mentor but also a trusted friend whose wisdom and constant encouragement have helped me move forward in my research, during the tough times of the pandemic. I would also like to thank all the faculty and staff members and my fellow graduate students in the Department of Statistics and my dear friends Koushiki, Kushal, Ritabrata, Sayar, Soudeep and Swarnali, with whom I have spent the wonderful years of my graduate life.

I extend my sincere gratitude to Professor Anil Kumar Ghosh at ISI, Kolkata who taught me how to take big problems of the world and formulate them in the form of a much more tractable research question. I would also like to thank my amazing teacher Mr. Basudeb Polley who taught me to look at tough mathematical problems through an explorer's vision and forever changed my life in the process.

I would like to thank my loving parents Dr. Biswa Kumar Saha and Mrs. Putul Saha, my amazing brother and role model Dipankar, my sweetest sister-in-law Subhra, my beloved soulmate Rahul and my two precious little friends Irtish and Aaditri, for their unconditional love, unwavering support and unfathomable faith in me.

ABSTRACT

We study flexible Bayesian methods that are amenable to a wide range of learning problems involving complex high dimensional data structures, with minimal tuning. We consider parametric and semiparametric Bayesian models, that are applicable to both static and dynamic data, arising from a multitude of areas such as economics, finance and marketing, to name a few. A special emphasis is given on deriving probabilistic guarantees of these models, that corroborate their strong empirical performance and can potentially provide insight into interesting avenues for future research.

Chapter 1 describes the broader theme of our research. We focus on two important domains of Bayesian Statistics: Bayesian ensemble learning and latent factor models. As part of the first topic, we explore the theoretical properties and empirical adaptability of Bayesian trees and their additive ensembles, along with their multiple incarnations. In the second part of our research we propose a sparse factor analysis model for dynamic data that is suitable for discovering latent structures in multivariate time series arising from a wide range of real life applications.

Bayesian additive regression trees (BART) is an ensemble learning technique that has been adapted to a wide range of high dimensional learning tasks. In Chapter 2 we demonstrate that the BART model has a near-optimal posterior concentration rate when the underlying regression function is ν-Hölder continuous with $0 < \nu \leq 1$. In Chapter 3 we demonstrate that this theoretical guarantee extends beyond the regression problem, to encompass response variables belonging to the exponential family, thereby including variants of BART that are adaptable to other important applications, such as classification and count regression. We also prove that these results can be replicated not only for Hölder continuous functions but also when the regression function is a step function or a monotone function. In Chapter 4 we demonstrate the scope of BART for discrete choice modeling. We demonstrate that BART exhibits superior predictive accuracy on several benchmark datasets compared to some popular discrete choice models.

In Chapter 5, we propose a Bayesian sparse factor analysis model for high dimensional dynamic data. We address some important challenges that often hinder the practical deployment of many existing dynamic factor analysis tools. Firstly our model infers the number of latent factors from the data, instead of fixing this number to a user-defined value. Moreover both the number of latent factors, as well as the factor loadings are allowed to vary over time. Second, we propose an EM implementation that requires minimal identification constraints and is considerably faster than the MCMC sampler, for high dimensional applications. To demonstrate the efficacy of our model, we study a large scale US macroeconomic data with a special focus on the 2008 financial crisis.

Finally Chapter 6 concludes with a discussion on possible implications of our work and some promising future research directions.

CHAPTER 1

INTRODUCTION

In this age of information revolution, researchers in every field are facing challenges posed by a massive influx of complex high-dimensional data structures. These data, where number of features often exceed the sample size, require novel data analysis tools that are fast, flexible and require minimal calibration by the practitioner. Our research focuses on Bayesian models, that are capable of handling such large complex datasets, arising out of important real life applications. We are particularly interested in deriving theoretical guarantees for semiparametric Bayesian tools such as Bayesian trees and forests, that have demonstrated significant empirical success but lack any theoretical justification supporting their impressive performance. Another strand of our research concerns developing a sparse factor analysis technique, capable of discovering interpretable latent structures in high dimensional dynamic datasets. The objective is to answer key questions about the dynamic system and efficiently employ these findings for forecasting purpose. While the common theme of our research is Bayesian models for high dimensional data, we address two distinct areas of Bayesian Statistics: ensemble learning and latent variable models. The first portion of this dissertation is primarily a theoretical pursuit of an existing ensemble learning model applicable to static data, the second portion concerns a novel methodology for dynamic data analysis, motivated by a large-scale macroeconomic application.

Ensemble learning is a statistical paradigm built on the premise that many weak learners can perform exceptionally well when deployed collectively. The BART method of Chipman et al. [2010] is a prominent example of *Bayesian* ensemble learning, where each learner is a tree. Due to its impressive performance, BART has received a lot of attention from practitioners. Despite its wide popularity, however, theoretical studies of BART have begun emerging only very recently. Laying the foundations for the theoretical analysis of Bayesian forests, Rockova et al. [2020] showed optimal posterior concentration under *conditionally uniform tree priors*. These priors deviate from the actual priors implemented in BART. In

Chapter 2 [Rockova and Saha, 2019], we study the exact BART prior and propose a simple modification so that it *also* enjoys similar optimality properties. To this end, we dive into branching process theory. We obtain tail bounds for the distribution of total progeny under heterogeneous Galton-Watson (GW) processes exploiting their connection to random walks. We conclude with a result stating the optimal rate of posterior convergence for BART.

Since its introduction in 2010, the scope of BART has extended beyond high dimensional regression and binary classification, to include a broad range of applications such as survival analysis, causal inference, variable selection, interaction detection and varying coefficient models, to name a few. Despite its wide adaptability, the existing theoretical results on BART concentrate primarily around the continuous regression problem. In Chapter 3, we try to remedy this situation by exploring the theoretical properties of some important variants of BART that extend beyond regression. We describe a Generalized BART (G-BART) model, analogous to generalized linear models that extend beyond the continuous regression and classification setup, to response variables falling under the broader exponential family distributions and hence encompasses the classical BART model [Chipman et al., 2010], along with several of its important variants (e.g. log-linear BART by Murray [2020]). In the G-BART framework described in this Chapter, another generalization emerges in terms of the distribution of 'step-heights', that are allowed to come from a more general class of distributions, including the Gaussian distribution typically considered in Bayesian trees and forests. We examine the theoretical properties of G-BART, when the underlying regression function is either a step function, a monotone function or a Hölder continuous function. These results supplement the recently emerging literature on theoretical aspects of BART, that have till date primarily focused on only Hölder continuous regression functions. We demonstrate that this theoretical optimality persists through conventional adaptations of BART, such as for classification [Chipman et al., 2010, Denison et al., 1998] and count regression [Murray, 2020].

In Chapter 4, we introduce a novel application of BART for discrete consumer choice

models. We demonstrate that BART has significantly superior predictive performance than the widely popular Generalized Linear Models with fixed and/or random effects for ten benchmark consumer choice datasets. We also discuss how the BART model for multi-class classification can be suitably adapted to incorporate monotonic dependence of customer preference on specific covariates such as price or income. Such monotonicity assumptions are typical of consumer demand theory [Deaton and Muellbauer, 1980].

In the second portion of our research, Chapter 5 focuses on a Bayesian latent factor model geared towards discovering interpretable unobserved structures in high dimensional time series data. Latent factor models are extremely important in the arsenal of multivariate data analysis. Besides being conceptually attractive due to their strong probabilistic foundation, these models are efficiently adaptable to a broad range of static and dynamic applications. Despite its popularity across many fields, there are outstanding methodological challenges that have hampered wider implementation in practice. One major challenge is the selection of the number of factors. This issue is exacerbated in dynamic factor models where factors can disappear, emerge, and/or reoccur over time. Existing models that assume a known fixed number of factors may provide a misguided data representation, especially when the factor dimension is grossly misspecified. Another challenge is interpretability which is often regarded as an unattainable objective due to the lack of identifiability. Motivated by a topical macroeconomic application, we develop a flexible Bayesian method for dynamic factor analysis (DFA) that can simultaneously accommodate a time-varying number of factors and enhance interpretability through sparse mode detection. To this end, we turn to dynamic sparsity by employing Dynamic Spike-and-Slab (DSS) priors within DFA. Besides MCMC, a scalable Bayesian EM estimation is proposed for fast posterior mode identification via rotations to sparsity, enabling Bayesian data analysis at larger scales. To highlight the efficacy and usefulness of our proposed method, we study a high-dimensional balanced panel of macroeconomic variables covering multiple facets of the US economy, with a focus on the Great Recession of 2008. We demonstrate that the interpretation of the latent structures

discovered by our model corroborate the consensus among existing economic literature that pin down the housing bubble deflation and the resulting devaluation of mortgage-backed securities as a possible harbinger of the 2008 financial crisis.

In Chapter 6 we revisit the implications of our work and describe several promising avenues for future research that include extensions of BART in novel empirical applications, as well as exploring semiparametric inference from a theoretical perspective. We also discuss several potential adaptations of our dynamic sparse factor analysis model to networks and datasets with multiple response categories.

CHAPTER 2

ON THEORY FOR BART

2.1 Bayesian Machine Learning

Bayesian Machine Learning and Bayesian Non-parametrics share the same objective: increasing flexibility necessary to address very complex problems using a Bayesian approach with minimal subjective input. While the two fields can be, to some extent, regarded as synonymous, their emphasis is quite different. Bayesian non-parametrics has evolved into a largely theoretical field, studying frequentist properties of posterior objects in inifinite-dimensional parameter spaces. Bayesian machine learning, on the other hand, has been primarily concerned with developing scalable tools for computing such posterior objects. In this work, we bridge these two fields by providing theoretical insights into one of the workhorses of Bayesian machine learning, the BART method.

Bayesian Additive Regression Trees (BART) are one of the more widely used Bayesian prediction tools and their popularity continues to grow. Compared to its competitors (e.g. Gaussian processes, random forests or neural networks) BART requires considerably less tuning, while maintaining robust and relatively scalable performance (`BART` R package of McCulloch (2017), `bartMachine` R package of Bleich et al. [2014], top down particle filtering of Lakshminarayanan et al. [2013], `parallel BART` of Pratola et al. [2014] and `XBART` of He et al. [2019]). BART has been successfully deployed in many prediction tasks, often outperforming its competitors (see predictive comparisons on 42 data sets in Chipman et al. [2010]). More recently, its flexibility and stellar prediction has been capitalized in multiple application areas such as in causal inference tasks for heterogeneous/average treatment effect estimation (Hill [2011], Hahn et al. [2017] and references therein), interaction detection [Du and Linero, 2019], survival analysis [Sparapani et al., 2016], time series analysis [Taddy et al.,

. This chapter is based on the paper "On theory for BART" by Veronika Rockova and Enakshi Saha. The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019.

2011, Deshpande et al., 2020] and variable Selection [Bleich et al., 2014, Linero, 2018, Liu et al., 2018, Liu and Rockova, 2020], to name a few. BART has also served as a springboard for various incarnations and extensions including: Monotone BART (Chipman et al. [2016]), Heteroscedastic BART (Pratola et al. [2017]), treed Gaussian processes (Gramacy and Lee [2008]), dynamic trees (Taddy et al. [2011]), varying coefficient BART [Deshpande et al., 2020] and BART with targeted smoothing [Starling et al., 2020] to list a few. Related non-parametric constructions based on recursive partitioning have proliferated in the Bayesian machine learning community for modeling relational data (Mondrian process of Roy et al. [2008], Mondian forests (Lakshminarayanan et al. [2014]). In short, BART continues to have a decided impact on the field of Bayesian non-parametrics/machine learning.

Despite its widespread popularity, however, the theory has not caught up with its applications. First theoretical results were obtained only very recently. As a precursor to these developments, Coram et al. [2006] obtained a consistency result for Bayesian histograms in binary regression with a single predictor. van der Pas and Rockova [2017] provided a posterior concentration result for Bayesian regression histograms in Gaussian non-parametric regression, also with one predictor. Rockova et al. [2020] (further referred to as RP20) then extended their study to trees and forests in a high-dimensional setup where $p > n$ and where variable selection uncertainty is present. They obtained the first theoretical results for Bayesian CART, showing optimal posterior concentration (up to a log factor) around a ν-Hölder continuous regression function (with a smoothness $0 < \nu \leq 1$). Going further, they also show optimal performance for Bayesian forests, both in additive and non-additive regression. Linero and Yang [2017] obtained similar results for Bayesian ensembles, but for *fractional* posteriors (raised to a power). The proof of RP20, on the other hand, relies on a careful construction of sieves and applies to regular posteriors. In addition, Linero and Yang [2017] do not study step functions (the essence of Bayesian CART and BART) but aggregated smooth kernels, allowing for $\nu > 1$. Building on RP20, Liu et al. [2018] obtained model selection consistency results (for variable and regularity selection) for Bayesian forests.

Albeit related, the tree priors studied in RP20 are *not* the actual priors deployed in BART. Here, we develop new tools for the analysis of the actual BART prior and obtain parallel results to those in RP20. To begin, we dive into branching process theory to characterize aspects of the distribution on total progeny under heterogeneous Galton-Watson processes. Revisiting several useful facts about Galton-Watson processes, including their connection to random walks, we derive a new prior tail bound for the tree size under the BART prior. With our proving strategy, the actual prior of Chipman et al. [2010] *does not* appear to penalize large trees aggressively enough. We suggest a very simple modification of the prior by altering the splitting probability. With this minor change, the prior is shown to induce the right amount of regularization and optimal speed of posterior convergence.

This Chapter is structured as follows. Section 2.2 revisits trees and forests in the context of non-parametric regression and discusses the BART prior. Section 2.3 reviews the notion of posterior concentration. Section 2.4 discusses Galton Watson processes and their connection to Bayesian CART. Section 2.5 is concerned with tail bounds on total progeny. Section 2.6 and 2.7 describe prior and concentration properties of BART. Section 2.7 wraps up with a discussion. Relevant proofs are given in Section 2.9.

2.2 The Appeal of Trees/Forests

The data setup under consideration consists of $Y_i \in \mathbb{R}$, a set of low dimensional outputs, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in [0,1]^p$, a set of high dimensional inputs for $1 \leq i \leq n$. Our statistical framework is non-parametric regression, which characterizes the input-output relationship through

$$Y_i = f_0(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1),$$

where $f_0 : [0,1]^p \to \mathbb{R}$ is an unknown regression function. A regression tree can be used to reconstruct f_0 via a mapping $f_{\mathcal{T},\boldsymbol{\beta}} : [0,1]^p \to \mathbb{R}$ so that $f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) \approx f_0(\boldsymbol{x})$ for $\boldsymbol{x} \notin \{\boldsymbol{x}_i\}_{i=1}^n$.

Each such mapping is essentially a step function

$$f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \beta_k \mathbb{I}(\boldsymbol{x} \in \Omega_k) \tag{2.1}$$

underpinned by a tree-shaped partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ and a vector of step heights $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. The vector $\boldsymbol{\beta}$ represents the quantitative guesses of the average outcome inside each cell. Each partition $\mathcal{T}$ consists of rectangles obtained by recursively applying a splitting rule (an axis-parallel bisection of the predictor space). We focus on *binary* tree partitions, where each internal node (box) is split into two children (formal definition below).

Definition 2.2.1. *(A Binary Tree Partition) A binary tree partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ consists of K rectangular cells Ω_k obtained with $K-1$ successive recursive binary splits of the form $\{\boldsymbol{x}_j \leq c\}$ vs $\{\boldsymbol{x}_j > c\}$ for some $j \in \{1, \ldots, p\}$, where the splitting value c is chosen from observed values $\{x_{ij}\}_{i=1}^n$.*

Partitioning is intended to increase within-node homogeneity of outcomes. In the traditional CART method (Brieman et al. [1984]), the tree is obtained by "greedy growing" (i.e. sequential optimization of some impurity criterion) until homogeneity cannot be substantially improved. The tree growing process is often followed by "optimal pruning" to increase generalizability. Prediction is then determined by terminal nodes of the pruned tree and takes the form either of a class level in classification problems, or the average of the response variable in least squares regression problems (Brieman et al. [1984]).

In tree *ensemble* learning, each constituent is designed to be a weak learner, addressing a slightly different aspect of the prediction problem. These trees are intended to be shallow and are woven into a forest mapping

$$f_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}) = \sum_{t=1}^{T} f_{\mathcal{T}_t,\boldsymbol{\beta}_t}(\boldsymbol{x}), \tag{2.2}$$

where each $f_{\mathcal{T}_t,\boldsymbol{\beta}_t}(\boldsymbol{x})$ is of the form (2.1), $\mathcal{E} = \{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$ is an ensemble of trees and

$\boldsymbol{B} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}'$ is a collection of jump sizes for the T trees. Random forests obtain each tree learner from a bootstrapped version of the data. Here, we consider a Bayesian variant, the BART method of Chipman et al. [2010], which relies on the posterior distribution over $f_{\mathcal{E},\boldsymbol{B}}$ to reconstruct the unknown regression function f_0.

Bayesian Trees and Forests

Bayesian CART was introduced as a Bayesian alternative to CART, where regularization/stabilization is obtained with a prior rather than with pruning (Chipman et al. [1998], Denison et al. [1998]). The prior distribution is assigned over a class of step functions

$$\mathcal{F} = \{f_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}) \quad \text{of the form (2.2) for some } \mathcal{E} \text{ and } \boldsymbol{B}\} \tag{2.3}$$

in a hierarchical manner.

The BART prior by Chipman et al. [2010] assumes that the number of trees T is fixed. The authors recommend a default choice $T = 200$ which was seen to provide good results. Next, the tree components $(\mathcal{T}_t, \boldsymbol{\beta}_t)$ are a-priori independent of each other in the sense that

$$\pi(\mathcal{E}, \boldsymbol{B}) = \prod_{t=1}^{T} \pi(\mathcal{T}_t)\pi(\boldsymbol{\beta}_t \,|\, \mathcal{T}_t), \tag{2.4}$$

where $\pi(\mathcal{T}_t)$ is the prior probability of a partition $\mathcal{T}_t$ and $\pi(\boldsymbol{\beta}_t \,|\, \mathcal{T}_t)$ is the prior distribution over the jump sizes.

Prior on Partitions $\pi(\mathcal{T})$

In BART and Bayesian CART of Chipman et al. [1998], the prior over trees is specified implicitly as a tree generating stochastic process, described as follows:

1. Start with a single leave (a root node) $[0, 1]^p$.

2. Split a terminal node, say Ω_t, with a probability

$$p_{split}(\Omega_t) = \frac{\alpha}{(1 + d(\Omega_t))^\gamma} \tag{2.5}$$

for some $\alpha \in (0, 1)$ and $\gamma \geq 0$, where $d(\Omega_t)$ is the depth of the node Ω_t in the tree architecture.

3. If the node Ω_t splits, assign a splitting rule and create left and right children nodes. The splitting rule consists of picking a split variable j uniformly from available directions $\{1, \ldots, p\}$ and picking a split point c uniformly from available data values $x_{1j}, \ldots, x_{nj}$. Non-uniform priors can also be used to favor splitting values that are thought to be more important. For example, splitting values can be given more weight towards the center and less weight towards the edges.

Prior on Step Heights $\pi(\boldsymbol{\beta} \,|\, \mathcal{T})$

Given a tree partition $\mathcal{T}_t$ with K_t steps, we consider iid Gaussian jumps

$$\pi(\boldsymbol{\beta}_t \,|\, \mathcal{T}_t) = \prod_{k=1}^{K_t} \phi(\beta_{tj}; 0, 1/T),$$

where $\phi(x; 0, \sigma^2)$ is a Gaussian density with mean 0 and variance σ^2. Chipman et al. [2010] recommend first shifting and rescaling Y_i's so that the observed transformed values range from -0.5 to 0.5. Then they assign a conjugate normal prior $\beta_{tj} \sim N(0, \sigma^2)$, where $\sigma = 0.5/k\sqrt{T}$ for some suitable value of k. This is to ensure that the prior assigns substantial probability to the range of the Y_i's.

The BART prior also involves an inverse chi-squared distribution on residual variance, with hyper-parameters chosen so that the q^{th} quantile of the prior is located at some sample based variance estimate. While the case of random variance can be incorporated in our analysis (van der Vaart et al. [2008]), we will for simplicity assume that the residual variance

is fixed.

Existing theoretical work for Bayesian forests (RP20) is available for a different prior on tree partitions $\mathcal{T}$. Their analysis assumes a hierarchical prior consisting of (a) a prior on the size of a tree K and (b) a uniform prior over trees of size K. This prior is equalitarian in the sense that trees with the same number of leaves are a-priori equally likely regardless of their topology. RP20 also imposed a diversification restriction in their prior, focusing on δ-valid ensembles (Definition 5.3) which consist of trees that do not overlap too much. The prior on the number of leaves K is a very important ingredient for regularization. We will study aspects of its distribution under the actual BART prior in later sections.

2.3 Bayesian Non-parametrics Lense

One way of assessing the quality of a Bayesian procedure is by studying the learning rate of its posterior, i.e. the speed at which the posterior distribution shrinks around the truth as $n \to \infty$. These statements are ultimately framed in a frequentist way, describing the typical behavior of the posterior under the true generative model $\mathbb{P}_{f_0}^{(n)}$. Posterior concentration rate results have been valuable for the proposal and calibration of priors. In infinite-dimensional parameter spaces, such as the one considered here, seemingly innocuous priors can lead to inconsistencies (Cox [1993], Diaconis and Freedman [1986]) and far more care has to be exercised to come up with well-behaved priors.

The Bayesian approach requires placing a prior measure $\Pi(\cdot)$ on $\mathcal{F}$, the set of qualitative guesses of f_0. Given observed data $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)'$, inference about f_0 is then carried out via the posterior distribution

$$\Pi(A \mid \boldsymbol{Y}^{(n)}) = \frac{\int_A \prod_{i=1}^n \Pi_f(Y_i \mid \boldsymbol{x}_i) \mathrm{d}\,\Pi(f)}{\int \prod_{i=1}^n \Pi_f(Y_i \mid \boldsymbol{x}_i) \mathrm{d}\,\Pi(f)} \quad \forall A \in \mathcal{B} \tag{2.6}$$

where $\mathcal{B}$ is a σ-field on $\mathcal{F}$ and where $\Pi_f(Y_i \,|\, \boldsymbol{x}_i)$ is the likelihood function for the output Y_i under f.

In Bayesian non-parametrics, one of the usual goals is determining *how fast the posterior probability measure concentrates around* f_0 as $n \to \infty$. This speed can be assessed by inspecting the size of the smallest $\|\cdot\|_n$-neighborhoods around f_0 that contain most of the posterior probability (Ghosal et al. [2007]), where $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)^2$ denotes the L_2 norm averaged over the available sample.

For a diameter $\varepsilon > 0$, we denote with

$$A_\varepsilon = \{f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F} : \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n \leq \varepsilon\} \tag{2.7}$$

the ε-neighborhood centered around f_0. We say that the posterior distribution concentrates at speed $\varepsilon_n \to 0$ such that $n\,\varepsilon_n^2 \to \infty$ when

$$\Pi(A^c_{\varepsilon_n, M_n} \mid \boldsymbol{Y}^{(n)}) \to 0 \quad \text{in } \mathbb{P}^{(n)}_{f_0}\text{-probability as } n \to \infty \tag{2.8}$$

for any $M_n \to \infty$. Posterior consistency statements are a bit weaker, where ε_n in (2.8) is replaced with a fixed neighborhood $\varepsilon > 0$. We will position our results using $\varepsilon_n = n^{-\nu/(2\nu+p)} \log^{1/2} n$, the near-minimax rate for estimating a p-dimensional ν-smooth function. We will also assume that f_0 is Hölder continuous, i.e. ν-Hölder smooth with $0 < \nu \leq 1$. The limitation $\nu \leq 1$ is an unavoidable consequence of using step functions to approximate smooth f_0 and can be avoided with smooth kernel methods (Linero and Yang [2017]).

The statement (2.8) can be proved by verifying the following three conditions (suitably adapted from Theorem 4 of Ghosal et al. [2007]):

$$\sup_{\varepsilon > \varepsilon_n} \log N\left(\tfrac{\varepsilon}{36}; A_{\varepsilon,1} \cap \mathcal{F}_n; \|.\|_n\right) \leq n\,\varepsilon_n^2 \tag{2.9}$$

$$\Pi(A_{\varepsilon_n,1}) \geq \mathrm{e}^{-d\,n\,\varepsilon_n^2} \tag{2.10}$$

$$\Pi(\mathcal{F} \backslash \mathcal{F}_n) = o(\mathrm{e}^{-(d+2)\,n\,\varepsilon_n^2}) \tag{2.11}$$

for some $d > 2$. In (2.9), $N(\varepsilon; \Omega; d)$ is the ε-covering number of a set Ω for a semimetric d, i.e. the minimal number of d-balls of radius ε needed to cover a set Ω.

A few remarks are in place. The condition (2.11) ensures that the prior zooms in on smaller, and thus more manageable, sets of models $\mathcal{F}_n$ by assigning only a small probability outside these sets. The condition (2.9) is known as "the entropy condition" and controls the combinatorial richness of the approximating sets $\mathcal{F}_n$. Finally, condition (2.10) requires that the prior charges an ε_n neighborhood of the true function. The results of type (**??**) quantify not only the typical distance between a point estimator (posterior mean/median) and the truth, but also the typical spread of the posterior around the truth. These results are typically the first step towards further uncertainty quantification statements.

2.4 The Galton-Watson Process Prior

The Galton-Watson (GW) process provides a mathematical representation of an evolving population of individuals who reproduce and die subject to laws of chance. Binary tree partitions $\mathcal{T}$ under the prior (2.5) can be thought of as realizations of such a branching process. Below, we review some terminology of branching processes and link them to Bayesian CART.

We denote with Z_t the population size at time t (i.e. the number of nodes in the t^{th} layer of the tree). The process starts at time $t = 0$ with a single individual, i.e. $Z_0 = 1$. At time t, each member is split *independently* of one another into a random number of offsprings. Let Y_{ti} denote the number of offsprings produced by the i^{th} individual of the t^{th} generation and let $g_t(s)$ be the associated probability generating function. A binary tree is obtained when each node has either *zero* or *two* offsprings, as characterized by

$$g_t(s) = s^0 \mathbb{P}(Y_{t1} = 0) + s^2 \mathbb{P}(Y_{t1} = 2), \quad 0 \leq s \leq 1. \tag{2.12}$$

Homogeneous GW process is obtained when all Y_{ti}'s are iid. A *heterogeneous* GW process is a generalization where the offspring distribution is allowed to vary according to the generations,

i.e. the variables Y_{ti} are independent but *not identically distributed*. The Bayesian CART prior of Chipman et al. [1998] can be framed as a heterogeneous GW process, where the probability of splitting a node (generating offsprings) depends on the depth t of the node in the tree. In particular, using (2.5) one obtains for $0 < \alpha < 1$ and $\gamma > 0$

$$\mathbb{P}(Y_{t1} = 2) = 1 - \mathbb{P}(Y_{t1} = 0) = \frac{\alpha}{(1+t)^{\gamma}}. \tag{2.13}$$

The population size at time t satisfies $Z_t = \sum_{i=1}^{Z_{t-1}} Y_{ti}$ and its expectation can be written as

$$\mathbb{E}Z_t = \mathbb{E}[\mathbb{E}(Z_t \mid Z_{t-1})] = (2\alpha)^t[(t+1)!]^{-\gamma}.$$

Since $\mathbb{E}Z_1 < 1$ under (2.13), the process is subcritical and thereby it dies out with probability one. This means that the random sequence $\{Z_t\}$ consists of zeros for all but a finite number of t's. The overall number of nodes in the tree (all ancestors in the family pedigree)

$$X = \sum_{t=0}^{\infty} Z_t \tag{2.14}$$

is thus finite with probability one. The number of *leaves* (bottom nodes) K can be related to X through

$$K = (X+1)/2 \tag{2.15}$$

and satisfies

$$T_{ex} + 1 \leq K \leq 2^{T_{ex}}, \tag{2.16}$$

where $T_{ex} = \min\{t : Z_t = 0\}$ is the time of extinction. In (2.16), we have used the fact that $T_{ex} - 1$ is the depth of the tree, where the lower bound is obtained with asymmetric trees with only one node split at each level and the upper bound is obtained with symmetric full binary trees (all nodes are split at each level).

Regularization is an essential remedy against overfitting and Bayesian procedures have

a natural way of doing so through a prior. In the context of trees, the key regularization element is the prior on the number of bottom leaves K, which is completely characterized by the distribution of total progeny X via (2.15). Using this connection, in the next section we study the tail bounds of the distribution $\pi(K)$ implied by the Galton-Watson process.

2.5 Bayesian Tree Regularization

If we knew ν, the optimal (rate-minimax) choice of the number of tree leaves would be $K \asymp K_\nu = n^{p/(2\nu+p)}$ (RP20). When ν is unknown, one can do almost as well (sacrificing only a log factor in the convergence rate) using a suitable prior $\pi(K)$. As noted by Coram et al. [2006], the tail behavior of $\pi(K)$ is critical for controlling the vulnerability/resilience to overfitting. The anticipation is that with smooth f_0, more rapid posterior concentration takes place when $\pi(K)$ has a heavier tail. However, too heavy tails make it easier to overfit when the true function is less smooth. To achieve an equilibrium, Denison et al. [1998] suggest the Poisson distribution (constrained to $\mathbb{N}\backslash\{0\}$), which satisfies

$$\mathbb{P}(K > k) \lesssim \mathrm{e}^{-C_K\, k \log k} \quad \text{for some } C_K > 0. \tag{2.17}$$

Under this prior, one can show that $\mathbb{P}(K > C\, K_\nu \,|\, \boldsymbol{Y}^{(n)}) \to 0$ in $\mathbb{P}_{f_0}^{(n)}$ probability (RP20). The posterior thus does not overshoot the oracle K_ν too much.

In the BART prior, the distribution $\pi(K)$ is implicitly defined through the GW process rather than directly through (2.17). In order to see whether BART induces a sufficient amount of regularization, we first need to obtain a tail bound of $\pi(K)$ under the GW process and show that it decays fast enough. One seemingly simple remedy would be to set $\gamma = 0$ (which coincides with the homogeneous GW case) and $\alpha = c/n$ with some $c > 0$. Standard branching process theory then implies $\Pi(K > k) \lesssim \mathrm{e}^{-C_K\, k \log n}$. This prior is more aggressive than (2.17). Moreover, letting the split probability $p_{split}(\Omega_k)$ decay with sample size is counterintuitive. By choosing $\alpha = c$, on the other hand, one obtains $\Pi(K > k) \lesssim \mathrm{e}^{-C_K\, k}$

which is not aggressive enough.

While the homogeneous GW processes have been studied quite extensively, the literature on tail bounds for *heterogeneous* GW processes (for when $\gamma \neq 0$) has been relatively deserted. We first review one interesting approach in the next section and then come up with a new bound in Section 2.5.2.

2.5.1 Tail Bounds à la Agresti

Agresti [1975] obtained both upper and lower bounds for the extinction time distribution of branching processes with independent non-identically distributed environmental random variables Y_{ti}. These bounds correspond to bounding the extinction time of the process by the extinction times of two varying environment GW processes. For our purposes, we will require only the upper bound, as given in the next theorem.

Theorem 2.5.1. *[Agresti, 1975] Consider the heterogeneous Galton-Watson branching process with offspring probability generating functions $\{g_j(s); j \geq 0\}$. For each j, let $g_j^{'}$ and $g_j^{''}$ denote the first and second order derivatives of the function g_j, respectively, satisfying $g_j^{''}(1) < \infty$ for $j \geq 0$. Denote $P_t = \prod_{j=0}^{t-1} g_j'(1)$. Then*

$$\mathbb{P}(T_{ex} > t) \leq \left[P_t^{-1} + \frac{1}{2} \sum_{j=0}^{t-1} (g_j^{''}(0)/g_j'(1) P_{j+1}) \right]^{-1}. \tag{2.18}$$

Using this result, we can obtain a tail bound on the extinction time under the Bayesian CART prior.

Corollary 2.5.1. *For the heterogeneous Galton-Watson branching process with offspring p.g.f.'s* (2.12) *with* (2.13) *we have*

$$\mathbb{P}(T_{ex} > t) < C_0 \left(\frac{t^\gamma}{2\alpha \mathrm{e}^\gamma} \right)^{-t} \tag{2.19}$$

for a positive constant C_0 that depends on α and γ.

Proof. We have $g_0(s) = s$ and for $j \geq 1$

$$g_j(s) = 1 - \alpha(1+j)^{-\gamma} + s^2\alpha(1+j)^{-\gamma},$$

Taking first and second order derivatives with respect to s, we get

$$g_j^{'}(s) = 2s\alpha(1+j)^{-\gamma},$$
$$g_j^{''}(s) = 2\alpha(1+j)^{-\gamma}.$$

Thus we have $g_0'(1) = 1$ and $g_j^{'}(1) = g_j^{''}(0) = 2\alpha(1+j)^{-\gamma}$ for $j \geq 1$. Then we can write

$$P_t^{-1} = \frac{\prod_{i=0}^{t-1}(1+i)^{\gamma}}{(2\alpha)^t} = \frac{(t!)^{\gamma}}{(2\alpha)^t} \tag{2.20}$$

and

$$\sum_{j=0}^{t-1}(g_j^{''}(0)/g_j'(1)P_{j+1}) = \sum_{j=0}^{t-1}\frac{1}{P_{j+1}} = \sum_{j=1}^{t}\frac{(j)!^{\gamma}}{(2\alpha)^j} > \frac{(t!)^{\gamma}}{(2\alpha)^t}.$$

Using (2.20) and the fact that $t! > (t/\mathrm{e})^t\mathrm{e}$, we can upper-bound the right hand side of (2.18) with $C_0[t^{\gamma}/(\mathrm{e}^{\gamma}2\alpha)]^{-t}$. □

Remark 2.5.1. *A simpler bound on the extinction time can be obtained using Markov's inequality as follows:* $\mathbb{P}(T_{ex} > t) = \mathbb{P}(Z_t \geq 1) \leq \mathbb{E}Z_t \leq (2\alpha)^t[(t+1)!]^{-\gamma}$.

Using the upper bound in (2.16) we immediately conclude that

$$\mathbb{P}(K > k) < \mathbb{P}(T_{ex} > \log_2 k) \leq C_0\left(\frac{\log_2^{\gamma} k}{2\alpha\mathrm{e}^{\gamma}}\right)^{-\log_2 k}.$$

This decay, however, is not fast enough as we would ideally like to show (2.17). We try a bit different approach in the next section.

2.5.2 Trees as Random Walks

There is a curious connection between branching processes and random walks (see e.g. Dwass [1969]). Suppose that a binary tree $\mathcal{T}$ is revealed in the following node-by-node exploration process: one exhausts all nodes in generation d before revealing nodes in generation $d+1$. Namely, nodes are implicitly numbered (and explored) according to their priority and this is done in a top/down manner according to their layer and a left-to-right manner within each layer (i.e. Ω_0 is the root node and, if split, Ω_1 and Ω_2 are the two children (left and right) etc.)

Nodes that are waiting to be explored can be organized in a queue Q. We say that a node is *active* at time t if it resides in a queue. Starting with one active node at $t = 0$ (the root node), at each time t we deactivate (remove from Q) the node with the highest priority (lowest index) and add its children to Q. Letting S_t be the number of active nodes at time t, one finds that $\{S_t\}$ satisfies

$$S_t = S_{t-1} - 1 + Y_t, \quad t \geq 1,$$

and $S_0 = 1$, where Y_t are sampled from the offspring distribution. For the *homogeneous* GW process, S_t is an actual random walk where Y_t are iid with a probability generating function (2.12). For the *heterogeneous* GW process, S_t is not strictly a random walk in the sense that $Y_t's$ are not iid. Nevertheless, using this construction one can see that the total population X equals the first time the queue is empty:

$$X = \min\{t \geq 0 : S_t = 0\}.$$

Linking Galton-Watson trees to random walk excursions in this way, one can obtain a useful tail bound of the distribution of the population size X. While perhaps not surprising, we believe that this bound is new, as we could not find any equivalent in the literature.

Lemma 2.5.1. *Denote by X the total population size* (2.14) *arising from the heterogeneous Galton-Watson process. Then we have for any $c > 0$*

$$\mathbb{P}(X > k) \leq \mathrm{e}^{-k\,c+(\mathrm{e}^{2c}-1)\mu}, \tag{2.21}$$

where $\mu = \sum_{i=1}^{k} p_i$ and $p_i = p_{split}(\Omega_i)$, where nodes Ω_i are ordered in a top-down left-to-right fashion.

Proof. For $k > 0$, we can write

$$\mathbb{P}(X > k) \leq \mathbb{P}(S_k > 0) = \mathbb{P}\left(\sum_{i=1}^{k} Y_i > k-1\right),$$

where X is the number of all nodes (internal and external) in the tree and Y_i has a two-point distribution characterized by $\mathbb{P}(Y_i = 2) = 1 - \mathbb{P}(Y_i = 0) = p_i$. Using the Chernoff bound, one deduces that for any $c > 0$

$$\mathbb{P}\left(\sum_{i=1}^{k} Y_i > k-1\right) \leq \mathrm{e}^{-k\,c}\,\mathbb{E}\mathrm{e}^{c\sum_{i=1}^{k} Y_i} = \mathrm{e}^{-k\,c}\prod_{i=1}^{k}[p_i\mathrm{e}^{2c} + 1 - p_i] \leq \mathrm{e}^{-k\,c+(\mathrm{e}^{2c}-1)\mu}$$

where $\mu = \sum_{i=1}^{k} p_i$. □

The goal throughout this section has been to understand whether the Bayesian CART prior of Chipman et al. [1998] yields (2.17) for some $C_K > 0$. The prior assumes $p_i = \alpha/(1+d(\Omega_i))^{\gamma}$. Choosing $c = (\log k)/2$ in (2.21), the right hand side will be smaller than $\mathrm{e}^{-a\,k\log k}$, for some suitable $0 < a < 1/2$, as long as $\mu \leq (1/2 - a)\log k$. We note that

$$\mu = \sum_{i=1}^{k} p_i < \sum_{d=1}^{\lceil \log_2 k \rceil} \frac{\alpha}{(1+d)^{\gamma}} 2^d.$$

Because the split probability p_i decreases only polynomially in depth of Ω_i, this *is not enough* to ensure $\mu < (1/2 - a)\log(k)$. The optimal decay, however, will be guaranteed if

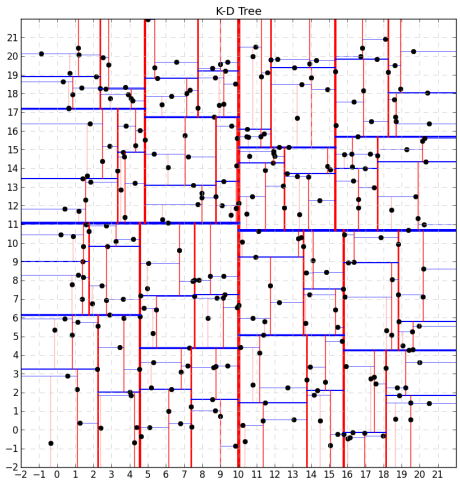


Figure 2.1: The *k-d* trees in two dimensions at various resolution levels.

we instead choose

$$p_{split}(\Omega) \propto \alpha^{d(\Omega)} \quad \text{for some } 0 < \alpha < 1/2. \tag{2.22}$$

To conclude, from our considerations it is not clear that the Bayesian CART prior of Chipman et al. [1998] has the optimal tail-bound decay. The following Corollary certifies that the optimal tail behavior can be obtained with a suitable modification of $p_{split}(\Omega)$.

Corollary 2.5.2. *Under the Bayesian CART prior of Chipman et al. [1998] with* (2.22)*, we obtain* (2.17)*.*

Proof. Follows from the considerations above and from (2.15).

2.6 Prior Concentration for BART

One of the prerequisites for optimal posterior concentration (2.8) is optimal prior concentration (Condition (2.10)). This condition ensures that there is enough prior support around the truth. It can be verified by constructing one approximating tree and by showing that it has enough prior mass. RP20 use the *k-d* approximating tree (Remark 3.1), which is a balanced full binary tree which partitions $[0,1]^p$ into nearly identical rectangles (in sufficiently regular designs). This tree can be regarded as the most regular partition that can be

obtained by splitting at observed values. A formal definition of the *k-d* tree is below and a few two-dimensional examples.

Definition 2.6.1. *(k-d tree partition) The k-d tree partition is constructed by cycling over coordinate directions* $\{1, \ldots, p\}$*, where all nodes at the same level are split along the same axis. For a given direction* $j \in \{1, \ldots, p\}$*, each internal node, say* Ω_k*, will be split at a median of the point set (along the* j^{th} *axis). Each split thus roughly halves the number of points inside the cell.*

After s rounds of splits on each variable, all K terminal nodes have at least $\lfloor n/K \rfloor$ observations, where $K = 2^{s\,p}$. The *k-d* tree partitions are thus balanced in light of Definition 2.4 of Rockova et al. [2020] (i.e. have roughly the same number of observations inside).

The *k-d* tree construction is instrumental in establishing optimal prior/posterior concentration. Lemma 3.2 of RP20 shows that there *exists* a step function supported by a *k-d* partition that safely approximates f_0 with an error smaller than a constant multiple of the minimax rate. The approximating *k-d* tree partition, denoted with $\widehat{\mathcal{T}}$, has $\widehat{K}$ steps where $\widehat{K} \asymp n\varepsilon_n^2 / \log n$ when $p \lesssim \log^{1/2} n$ (as shown in Section 8.3 of RP20 and detailed in the proof of Theorem 2.7.1).

In order to complete the proof of posterior concentration for the Bayesian CART under the Galton-Watson process prior, we need to show that $\pi(\widehat{\mathcal{T}}) \geq \mathrm{e}^{-c_1 n \varepsilon_n^2}$ for some $c_1 > 0$. This is verified in the next lemma.

Lemma 2.6.1. *Denote with* $\widehat{\mathcal{T}}$ *the k-d tree partition described above. Assume the heterogeneous Galton-Watson process tree prior with* $p_{split}(\Omega_k) \propto \alpha^{d(\Omega_k)}$ *for some suitable* $1/n \leq \alpha < 1/2$*. Assume* $p \lesssim \log^{1/2} n$*. Then we have for some suitable* $c_1 > 0$

$$\pi(\widehat{\mathcal{T}}) \geq \mathrm{e}^{-c_1\, n\, \varepsilon_n^2}.$$

Proof. By construction, the *k-d* tree $\widehat{\mathcal{T}}$ has $\widehat{K} = 2^{p \times s}$ leaves and $p \times s$ layers for some $s \in \mathbb{N}$ where p is the number of predictors. In addition, the *k-d* tree is complete and balanced (i.e.

every layer d, including the last one, has the maximal number 2^d of nodes). Since there are $\widehat{K}-1$ internal nodes and at least $1/(p\,n)$ splitting rules for each internal node, we have

$$\begin{aligned}\pi(\widehat{\mathcal{T}}) &\geq \frac{(1-\alpha^{s\,p})^{\widehat{K}}}{(p\,n)^{\widehat{K}-1}} \prod_{d=0}^{\log_2 \widehat{K}-1} \alpha^{2^d} \geq \frac{(1-\alpha^{s\,p})^{\widehat{K}}}{(p\,n)^{\widehat{K}-1}} \alpha^{\widehat{K}-1} \\ &\geq [\alpha(1-\alpha)]^{\widehat{K}} \left(\frac{1}{p\,n}\right)^{\widehat{K}-1} > \mathrm{e}^{-\widehat{K}\log(2n)-(\widehat{K}-1)\log(p\,n)}.\end{aligned}$$

Since $p \lesssim \log^{1/2} n$ and $\widehat{K} \asymp n\,\varepsilon_n^2/\log n$ we can lower-bound the above with $\mathrm{e}^{-c_1\,n\varepsilon_n^2}$ for some $c_1 > 0$. □

For the actual BART prior [Chipman et al., 2010] (similarly as in Theorem 5.1 of RP20), one needs to find an approximating *tree ensemble* and show that it has enough prior support. The approximating ensemble can be found in Lemma 10.1 of RP20 and consists of $\widehat{\mathcal{E}} = \{\widehat{\mathcal{T}}_1, \ldots, \widehat{\mathcal{T}}_T\}$ tree partitions obtained by chopping off branches of $\widehat{\mathcal{T}}$. The number of trees T is fixed and the trees $\mathcal{T}_t$ will not overlap much when $1 \leq T \leq \widehat{K}/2$. The default BART choice $T = 200$ safely satisfies this as long as $p > 9$, thus ensuring a diversified ensemble of shallow trees $\{\widehat{\mathcal{T}}_1, \ldots, \widehat{\mathcal{T}}_T\}$. The component little trees $\widehat{\mathcal{T}}_t$ have $\widehat{K}^t$ leaves and satisfy $\log_2 \widehat{K} + 1 \leq \widehat{K}^t \leq \widehat{K}$, depending upon the choice of T. Since T is fixed, using Lemma 2.6.1 and the fact that the trees are independent a-priori (from (2.4)), we then obtain a lower bound for the prior probability on the forest mapping:

$$\begin{aligned}\pi(\widehat{\mathcal{E}}) &\geq \mathrm{e}^{-\sum_{t=1}^{T}[\widehat{K}^t \log 2n + (\widehat{K}^t-1)\log(p\,n)]} \\ &> \mathrm{e}^{-T\widehat{K}\log 2n - T(\widehat{K}-1)\log(p\,n)} > \mathrm{e}^{-c_2\,n\varepsilon_n^2}\end{aligned}$$

for some constant $c_2 > 0$.

The BART prior thus concentrates enough mass around the truth. Condition (2.10) also requires verification that the prior on jump sizes concentrates around the forest sitting on $\widehat{\mathcal{E}}$. This follows directly from Section 9.2 of RP20. We detail the steps in the proof of Theorem

2.7.1, in Section 2.9.

2.7 Posterior Concentration for BART

We now have all the ingredients needed to state the posterior concentration result for BART. The result is *different* from Theorem 5.1 of RP20 because here we (a) assume that T is fixed, (b) assume the branching process prior on $\mathcal{T}$ and (c) we do not have subset selection uncertainty. We will treat the design as fixed and *regular* according to Definition 3.3 of RP20. Moreover, the BART prior support will be restricted to δ-valid ensembles with $\delta \geq 1$.

Theorem 2.7.1. *(Posterior Concentration for BART) Assume that f_0 is ν-Hölder continuous with $0 < \nu \leq 1$ where $\|f_0\|_\infty \lesssim \log^{1/2} n$. Assume a regular design $\{\boldsymbol{x}_i\}_{i=1}^n$ where $p \lesssim \log^{1/2} n$. Assume the BART prior with T fixed and with $p_{split}(\Omega_t) = \alpha^{d(\Omega_t)}$ for $1/n \leq \alpha < 1/2$. With $\varepsilon_n = n^{-\nu/(2\nu+p)} \log^{1/2} n$ we have*

$$\Pi\left(f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F} : \|f_0 - f_{\mathcal{E},\boldsymbol{B}}\|_n > \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

in $\mathbb{P}_{f_0}^{(n)}$-probability, as $n, p \to \infty$.

Proof. Section 2.9. □

The logarithmic term in the expression $\varepsilon_n = n^{-\nu/(2\nu+p)} \log^{1/2} n$, is an unavoidable consequence of the limitation of trees (step functions) in recovering smoother functions. As will be seen in Section 2.9, in Theorem 2.7.1 we derive an approximating k-d tree partition with k_n leaves, that estimates f_0 with a desired level of accuracy. In particular the prior concentration condition (2.10) involves lower-bounding the left hand side of the inequality by the prior probability on estimating partitions supported on $k_n \approx n^{p/(2\nu+p)} = n\varepsilon_n^2/\log n$ leaves, which is "too coarse" to estimate the true function with high enough accuracy. For example the tree construction procedure (2.27) suggests that while estimating the function $f_0(x) = x$, a Bayesian tree will require approximately $n^{1/3}$ leaves (up to a multiplicative

constant), yielding an approximation error of the order $n^{-1/3}$, which is "too large" to attain the minimax rate of $n^{-5/2}$, over the class of all continuously differentiable functions.

Theorem 2.7.1 has very important implications. It provides a frequentist theoretical justification for BART claiming that the posterior is wrapped around the truth and its learning rate is near-optimal. As a by-product, one also obtains a statement which supports the empirical observation that BART is resilient to overfitting.

Corollary 2.7.1. *Under the assumptions of Theorem 2.7.1 we have*

$$\Pi\left(\bigcup_{t=1}^{T}\{K^t > C\, n^{p/(2\nu+p)}\} \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

in $\mathbb{P}_{f_0}^{(n)}$-probability, as $n, p \to \infty$, for a suitable constant $C > 0$.

Proof. The proof follows from the proof of Theorem 2.7.1 and Lemma 1 of Ghosal et al. [2007]. □

In other words, the posterior distribution rewards ensembles that consist of small trees whose size does not overshoot the optimal number of steps $K_\nu = n^{p/(2\nu+p)}$ by much. In this way, the posterior is fully adaptive to unknown smoothness, not overfitting in the sense of split overuse.

Remark 2.7.1. *Interestingly, the posterior concentration rate derived in Theorem 2.7.1, does not depend on the number of trees T in the BART ensemble. In other words the concentration rate is equally valid for a single tree (i.e. $T = 1$), as well as for tree ensembles (i.e. $T > 1$), when the true regression function f_0 is ν-Hölder continuous with $0 < \nu \leq 1$. However as has been seen in multiple empirical applications [Chipman et al., 2010], Bayesian forests consisting of multiple trees provide superior predictive performance, compared to a single tree, the reason being that multiple weak tree learners, when woven together into a forest, can accommodate a wider class of partitions, as opposed to a single tree. This phenomenon can be reinforced by theoretical results, such as Theorem 6.1 of RP20. When the*

true function f_0 is of the form $f_0 = \sum_{t=1}^{T_0} f_0^t$, where f_0^t is a ν_t-Hölder continuous function, with $0 \leq \nu^t \leq 1$, a forest with multiple trees have a posterior concentration rate equal to $\varepsilon_n^2 = \sum_{t=1}^{T_0} n^{-2\nu_t/(2\nu_t+p)} \log n$, provided $T_0 \lesssim n$, whereas single regression trees fail to recognize the additive nature of the true function and attain a slower concentration rate[1]*. Although the BART prior considered by RP20 is fundamentally different from the classical BART prior [Chipman et al., 2010] considered here, their result on additive functions can be replicated in the present set up as well, provided we allow the number of trees T in the BART ensemble to be random. In particular, following in the footsteps of RP20, we can assume*

$$\pi(T) \propto e^{-C_T T}, \text{ for } T \in \mathbb{N} \setminus \{0\}, \text{ with } C_T > \log 2, \tag{2.23}$$

thus enabling the number of trees in the forest to adapt to unknown T_0, as $n, p \to \infty$.

2.8 Discussion

In this work, we have built on results in Rockova et al. [2020] to show optimal posterior convergence rate of the BART method in the $\|\cdot\|_n$ sense. We have proposed a minor modification of the prior that guarantees this optimal performance. Similar results have been obtained for other Bayesian non-parametric constructions such as Polya trees (Castillo et al. [2017]), Gaussian processes (van der Vaart et al. [2008], Castillo et al. [2008]) and deep ReLU neural networks [Polson and Rockova]. Up to now, the increasing popularity of BART has relied on its practical performance across a wide variety of problems. The goal of this and future theoretical developments is to establish BART as a rigorous statistical tool with solid theoretical guarantees. Similar guarantees have been obtained for variants of the traditional forests/trees by multiple authors including Gordon and Olshen [1980], Donoho et al. [1997], Biau et al. [2008], Scornet et al. [2015], Wager and Walther [2015]. Our posterior

1. A similar result is presented in Theorem 4 of Linero and Yang [2017], under a kernel-smoothed version of the BART prior, where the individual smoothness levels ν_t of the component functions are allowed to be greater than one.

concentration results break the path towards establishing other theoretical properties such as Bernstein-von Mises theorems (semi and non-parametric) and/or uncertainty quantification statements.

2.9 Proof of Theorem 2.7.1

The proof follows from Lemma 2.6.1, Lemma 2.5.1 and a modification of proof of Theorem 5.1 of RP20. Below, we outline the backbone of the proof and highlight those places where the proof of RP20 had to be modified. Our approach consists of establishing conditions (2.9), (2.10) and (2.11) for $\varepsilon_n = n^{-\nu/(2\nu+p)} \log^{1/2} n$. The first step requires constructing the sieve $\mathcal{F}_n \subset \mathcal{F}$. For a given $n \in \mathbb{N}$, $T \in \mathbb{N}$ and a suitably large integer k_n (chosen later), we define the sieve as follows:

$$\mathcal{F}_n = \bigcup_{\boldsymbol{K}: K^t \leq k_n} \bigcup_{\mathcal{E} \in \mathcal{VE}^{\boldsymbol{K}}} \mathcal{F}(\mathcal{E}), \tag{2.24}$$

where $\mathcal{F}(\mathcal{E})$ consists of all functions $f_{\mathcal{E},\boldsymbol{B}}$ of the form (2.2) that are supported on a δ-*valid* ensemble $\mathcal{E}$. All δ-valid ensembles consisting of T trees of sizes $\boldsymbol{K} = (K^1, \ldots, K^T)'$ are denoted with $\mathcal{VE}^{\boldsymbol{K}}$. The sieve (2.24) is different from the one in the proof of Theorem 5.1 of RP20. Their sieve consisted of all ensembles whose *total* number of leaves was smaller than k_n. Here, we allow for each tree individually to have up to k_n leaves.

Regarding Condition (2.9), RP20 in Section 9.1 obtain an upper bound on the covering number for $\mathcal{F}(\mathcal{E})$ as well as the cardinality of $\mathcal{VE}^{\boldsymbol{K}}$ which together yield (for some $D > 0$)

$$\begin{aligned} \log N\Big(\tfrac{\varepsilon}{36}, \Big\{ f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F}_n : \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n < \varepsilon \Big\}, \|.\|_n \Big) &< (k_n + 1) T \log(n\, p\, k_n) \\ &+ D\, T\, k_n \log\left(108\, \sqrt{T\, k_n} n^{1+\delta/2}\right). \end{aligned} \tag{2.25}$$

With the choice $k_n = \lfloor \widetilde{C} n \varepsilon_n^2 / \log n \rfloor \asymp n^{p/(2\alpha+p)}$ (for a large enough constant $\widetilde{C} > 0$), fixed $T \in \mathbb{N}$ and assuming $p \lesssim \log^{1/2} n$, the Condition 2.9 will be met.

Next, we wish to show that the prior assigns enough mass around the truth in the sense

that

$$\Pi(f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F} : \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n \leq \varepsilon_n) \geq \mathrm{e}^{-d\, n\varepsilon_n^2} \tag{2.26}$$

for some large enough $d > 2$. We establish this condition by finding a lower bound on the prior probability in (2.10), using only step functions supported on a single ensemble. According to Lemma 10.1 of RP20 there exists a 1-valid tree ensemble $f_{\widehat{\mathcal{E}},\widehat{\boldsymbol{B}}}$ that approximates f_0 well in the sense that

$$\|f_0 - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \leq ||f_0||_{\mathcal{H}^\nu} C\, p/\widehat{K}^{\nu/p} \tag{2.27}$$

for some $C > 0$, where $\|f_0\|_{\mathcal{H}^\nu}$ is the Hölder norm and where $\widehat{K} = 2^{s\,p}$ for some $s \in \mathbb{N}$. Next, we find the smallest $\widehat{K}$ such that $||f_0||_{\mathcal{H}^\nu} C\, p/\widehat{K}^{\nu/p} < \varepsilon_n/2$. This value will be denoted by a_n and it satisfies

$$\left(\frac{2C_0 p}{\varepsilon_n}\right)^{\frac{p}{\nu}} \leq a_n \leq \left(\frac{2C_0 p}{\varepsilon_n}\right)^{\frac{p}{\nu}} + 1. \tag{2.28}$$

Under the assumption $p \lesssim \log^{1/2} n$ we have $a_n \asymp n^{p/(2\nu+p)}$. Denote by $\widehat{\mathcal{E}}$ the approximating ensemble described in Section 2.6. Next, we denote with $\widehat{\boldsymbol{K}} = (\widehat{K}^1, \ldots, \widehat{K}^T)'$ the vector of tree sizes, where $\log_2 a_n + 1 \leq \widehat{K}^t \leq a_n$. Then we can lower-bound the left-hand side of (2.10) with

$$\pi(\widehat{\mathcal{E}})\Pi\left(f_{\widehat{\mathcal{E}},\mathcal{B}} \in \mathcal{F}(\widehat{\mathcal{E}}) : \|f_{\widehat{\mathcal{E}},\mathcal{B}} - f_0\|_n \leq \varepsilon_n\right), \tag{2.29}$$

where $\mathcal{F}(\widehat{\mathcal{E}})$ consists of all additive tree functions supported on $\widehat{\mathcal{E}}$. In Section 2.6 we show that $\pi(\widehat{\mathcal{E}}) > \mathrm{e}^{-c_2\, n\varepsilon_n^2}$. Moreover, RP20 in Section 10.2 show that, for some $C > 0$,

$$\Pi\left(f_{\widehat{\mathcal{E}},\mathcal{B}} \in \mathcal{F}(\widehat{\mathcal{E}}) : \|f_{\widehat{\mathcal{E}},\mathcal{B}} - f_0\|_n \leq \varepsilon_n\right) > \Pi\left(\mathcal{B} \in \mathbb{R}^{\widetilde{a}_n} : \|\mathcal{B} - \widehat{\mathcal{B}}\|_2 < \frac{\varepsilon_n}{2}\frac{1}{C\sqrt{\widetilde{a}_n}}\right),$$

where $\widetilde{a}_n = \sum_{t=1}^T \widehat{K}^t \leq T\, a_n$ and where $\widehat{\mathcal{B}} \in \mathbb{R}^{\widetilde{a}_n}$ are the steps of the approximating additive trees from Lemma 10.1 of RP20.

This can be further lower-bounded with

$$\mathrm{e}^{-\frac{\varepsilon_n^2}{8C^2\widetilde{a}_n}-a_n(C_2\|f_0\|_\infty^2+\log 2)}\left(\frac{\varepsilon_n^2}{4C^2\widetilde{a}_n}\right)^{\frac{\widetilde{a}_n}{2}}\left(\frac{2}{\widetilde{a}_n}\right)^{\widetilde{a}_n/2+1}. \tag{2.30}$$

Under the assumption $\|f_0\|_\infty \lesssim \log^{1/2} n$, this term is larger than $\mathrm{e}^{-D\,\widetilde{a}_n \log n}$ for some $D > 0$. Since $\widetilde{a}_n \lesssim n\varepsilon_n^2$, there exists $d > 0$ such that $\Pi(f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F} : \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n \leq \varepsilon_n) > \mathrm{e}^{-d\,n\varepsilon_n^2}$.

Lastly, Condition (2.11) entails showing that $\Pi(\mathcal{F}\backslash\mathcal{F}_n) = o(\mathrm{e}^{-(d+2)\,n\varepsilon_n^2})$ for d deployed in the previous paragraph. It suffices to show that

$$\Pi\left(\bigcup_{t=1}^{T}\{K^t > k_n\}\right)\mathrm{e}^{(d+2)\,n\varepsilon_n^2} \to 0.$$

Under the independent Galton-Watson prior on each tree partition, Corollary 2.5.2 implies that the probability above can be upper-bounded with $\sum_{t=1}^{T}\Pi(K^t > k_n) \lesssim T\mathrm{e}^{-C_K\,k_n\log k_n}$. With $k_n \asymp n\varepsilon_n^2/\log n$ and a fixed $T \in \mathbb{N}$, we have $T\mathrm{e}^{-C_K\,k_n\log k_n+(d+2)\,n\varepsilon_n^2} \to 0$ for C_K large enough.

CHAPTER 3

GENERALIZATIONS OF THEORY FOR BART

3.1 Introduction

In Chapter 2 we demonstrated that the Bayesian Additive Regression Trees (BART) estimator has a near-optimal posterior concentration rate for continuous regression, where the response variable is assumed to follow a Gaussian distribution and the underlying regression function is Hölder continuous. However as discussed in Section 2.1, the scope of BART and its various subsequent incarnations extend far beyond regression, to include many other specialized applications involving a wide range of response variables, such as binary classification [Denison et al., 1998, Chipman et al., 1998, 2010], regression on categorical and count responses [Murray, 2020] and regression on censored data [Sparapani et al., 2016], to name a few. Therefore a natural question to ask would be whether the optimality for BART models in terms of posterior concentration persists when the response distribution is not Gaussian and/or the regression function is not smooth.

In order to answer this question, we formulate a Generalized BART (henceforth referred to as G-BART) model, where the response variable is assumed to come from an exponential family distribution (and hence can be considered to be a semiparametric extension of Generalized Linear Models). Many prominent Bayesian CART and BART models used in practice [Denison et al., 1998, Chipman et al., 2010, Murray, 2020], including the regression model considered in Chapter 2, can be viewed as a special case of this generalized extension. Therefore theoretical properties of these conventional adaptations of BART can be studied as direct corollaries of analogous properties for the G-BART model. In particular, in this chapter we evaluate the posterior concentration rates of G-BART under the assumption that the parameters of the response distribution are unknown functions of the covariates and step functions supported on tree/forest partitions, equipped with the BART prior can be employed to estimate these parameters. We consider two different tree priors, one of them

being the BART prior proposed by Chipman et al. [1998] (considered in Chapter 2) and the other being the Bayesian CART prior proposed by Denison et al. [1998]. The theorems presented in this chapter extend the existing theoretical results on BART in three directions.

Firstly, previous theoretical results on BART [Rockova and Saha, 2019, Rockova et al., 2020, Linero and Yang, 2017] focus on continuous Gaussian responses, with the only exception of Jeong and Rockova [2020], who consider the problem of density estimation, alongside continuous regression and binary classification. We derive posterior concentration rates for BART when the response variable belongs to a class of exponential family distributions that includes Gaussian regression, both two-class (Bernoulli) and multi-class (Multinomial) classification / categorical response and count (Poisson) response variables, among others. We derive sufficient conditions on the response density, under which any BART model enjoys a near-minimax posterior concentration rate, under suitable regularity conditions on the underlying function space. We will demonstrate that the results for continuous regression discussed in Chapter 2 (e.g. Theorem 2.7.1) can be derived as a direct corollary of the theorems presented in this chapter (in particular Theorem 3.4.3).

Secondly, existing theoretical results on BART [Rockova et al., 2020, Rockova and Saha, 2019, Linero and Yang, 2017] build upon the assumption that the underlying regression function is Hölder continuous. Later Jeong and Rockova [2020] obtained similar statements for anisotropic Hölder functions. We present theoretical results for both step functions and monotone functions, alongside the set of all ν-Hölder continuous functions with $0 < \nu \leq 1$. The results on step functions are particularly important because posterior concentration rates for more general class of functions can be built upon these, aided by the "simple function approximation theorem" [Stein and Shakarchi, 2009]. Provisioned by these broader theoretical understanding, specifically Theorems 3.4.1 and 3.4.2, we can also prove that the theoretical optimality of BART demonstrated in Theorem 2.7.1 is not limited to Hölder continuous functions only and can be extended to include a variety of regression surfaces including step-functions and monotone functions.

Finally, Chipman et al. [2010] approximate the regression functions through step functions and assume that these step heights come from a Gaussian distribution. All subsequent theoretical and empirical developments have adopted this specification. In the G-BART setup we assume that the distribution of these step heights belong to a broader family of distributions that include both the Gaussian distribution and also some thicker tailed distributions like Laplace. We demonstrate that the BART model maintains a near-minimax posterior concentration rate, if the step heights come from any of the distributions belonging to this broader family, thus providing a wide range of distributional choices without sacrificing optimal posterior concentration. The theory also shows how important modelling choices such as link functions can impact performance of the posterior and hence can serve as a guide for empirical implementations as well.

Our Contributions

To summarize our previous discussion, we now briefly highlight our key contributions. The posterior concentration results discussed in this chapter, extend the existing theory on BART in three directions:

Response Distribution: We assume that the response variable comes from an exponential family distribution and derive sufficient conditions on the response density under which the posterior concentration rate of the BART model adapted to this particular response type would be almost equal to the minimax rate. This extends the existing theoretical results on BART for Gaussian regression.

Step Size Distribution: Instead of assigning any particular distribution on the step-heights (e.g. Gaussian distribution as in all existing literature) associated to the BART model, we impose sufficient conditions on the cumulative distribution function that guarantee a near-optimal posterior concentration rate.

Types of Functions: The objective of BART models is to estimate unknown functions that characterize the relationship between the response and the covariates. All existing results on BART assume this function to be Hölder continuous. We extend these results to the situations where the underlying function to be estimated is either a monotone function or a step function supported on an axes-paralleled partition.

This chapter is organized as follows. In Section 3.2 we describe the generalized BART model with the associated priors. Section 3.3 revisits the notion of posterior concentration, where we describe certain sufficient conditions for deriving posterior concentration rates. These conditions, although analogous to conditions (2.9)-(2.11) described in Chapter 2, have significantly wider applicability. We also highlight some key artifacts that distinguish the proof of Theorem 2.7.1 with the proofs of the main theoretical results on G-BART discussed in Section 3.4. Broader implications of these results are described in Section 3.5. In particular, Section 3.5.1 proves that the generalized Bayesian trees and G-BART models are resilient to overfitting and Section 3.5.2 describes some important adaptations of BART including classification and count regression and demonstrates how the results discussed in Section 3.4 can be employed to infer theoretical optimality of the posterior. We also demonstrate how some modelling choices can lead to better or worse posterior concentration rates, which might be insightful for practical implementations. Section 3.6 concludes with a discussion on the significance of G-BART, along with some possible future research directions. Proofs of major results are deferred till Section 3.7.

Notations:

For any two real numbers a and b, $a \vee b$ will denote the maximum of a and b. The notations $\gtrsim$ and $\lesssim$ will stand for "greater than or equal to up to a constant" and "less than or equal to up to a constant", respectively. For example $a \gtrsim b$ would be equivalent to saying there exists a constant $M > 0$ such that $a \geq Mb$.

The symbol P_f will abbreviate $\int f dP$ and $\mathbb{P}_f^{(n)} = \prod_{i=1}^n \mathbb{P}_f^i$ will denote the n-fold product measure of the n independent observations, where the i-th observation comes from the distribution P_f^i.

For a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ and a function $g : \mathbb{R}^q \to \mathbb{R}$, the notation $\nabla g(\boldsymbol{\beta})$ will denote the column vector of partial derivatives $\left(\frac{\partial g}{\partial \beta_1}, \ldots, \frac{\partial g}{\partial \beta_q}\right)^T$.

Let $h(f,g) = \left(\int (\sqrt{f} - \sqrt{g})^2 d\mu\right)^{1/2}$ and $K(f,g) = \int f \log(f/g) d\mu$ denote the Hellinger distance and the Kullback-Leibler divergence, respectively between any two non-negative densities f and g with respect to a measure μ. We define another discrepancy measure $V(f,g) = \int f \left(\log(f/g)\right)^2 d\mu$.

Finally, for any set of real vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^q$ of size n, define the average discrepancy measures $H_n(f,g) = \frac{1}{n}\sum_{i=1}^n H\left(f(\boldsymbol{X}_i), g(\boldsymbol{X}_i)\right)$, $K_n(f,g) = \frac{1}{n}\sum_{i=1}^n K\left(f(\boldsymbol{X}_i), g(\boldsymbol{X}_i)\right)$ and $V_n(f,g) = \frac{1}{n}\sum_{i=1}^n V\left(f(\boldsymbol{X}_i), g(\boldsymbol{X}_i)\right)$, where $f(\theta)$ and $g(\theta)$ denote the densities f and g with respect to parameter θ. Also, for any L_p norm $\|\cdot\|_p$, define the average norm $\|f - g\|_{p,n} = \frac{1}{n}\sum_{i=1}^n \|f - g\|_p$.

3.2 The Generalized BART Prior

The BART method of Chipman et al. [2010] is a prominent example of Bayesian ensemble learning, where individual shallow trees are entwined together into a forest, that is capable of estimating a multitude of nonlinear functions with exceptional accuracy, while simultaneously accounting for different orders of interactions among the covariates. Building upon BART, we describe a generalized framework, where the response variable is assumed to come from an exponential family distribution. For continuous Gaussian response variables, this generalized BART model reduces to the original BART prior of [Chipman et al., 2010].

The data setup under consideration consists of $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{ip})' \in \mathbb{R}^p$, a set of p-dimensional outputs, and $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iq})' \in [0,1]^q$, a set of q dimensional inputs for $1 \leq i \leq n$. We assume $\boldsymbol{Y}$ follows some distribution in the exponential family with density

of the following form:

$$P_{f_0}(\boldsymbol{Y} \mid \boldsymbol{X}) = h(\boldsymbol{Y}) g\left[f_0(\boldsymbol{X})\right] \exp\left[\eta\left(f_0(\boldsymbol{X})\right)^T T(\boldsymbol{Y})\right], \tag{3.1}$$

where $h : \mathbb{R}^p \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, $\eta : \mathbb{R}^p \to \mathbb{R}^J$, $T : \mathbb{R}^p \to \mathbb{R}^J$ for some integer J and $f_0 : \mathbb{R}^q \to \mathbb{R}^D$, for some integer D, are all real valued functions. Among these functions, h, g, η and T are usually *known* depending on the nature of the response $\boldsymbol{Y}$. The function f_0, connecting the input $\boldsymbol{X}$ with the output $\boldsymbol{Y}$, is the only unknown function and estimating this function is the primary objective of the G-BART estimator.

We assume that f_0 is an unconstrained function, i.e. the range of f_0 is the entire space $\mathbb{R}^D$ for some integer D. A suitable link function $\Psi(\cdot)$ is used to transform f_0 to the natural parameter of the distribution of $\boldsymbol{Y}$, which is often constrained. For example, in the continuous regression problem, $\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathcal{N}(\mu(\boldsymbol{X}), \sigma^2)$ for some $\sigma > 0$ [Chipman et al., 2010]. Since the natural parameter $\mu(\boldsymbol{X}) \in \mathbb{R}$ is unrestricted, we use $\Psi(z) = z$ as our link function. In contrast, for the binary classification problem, $\boldsymbol{Y} \sim \mathit{Bernoulli}\left(p(\boldsymbol{X})\right)$. Here the natural parameter $p(\boldsymbol{X}) \in (0, 1)$ is restricted and hence we can use $\Psi(z) = \frac{1}{1+\exp(-z)}$, the logistic function (or a probit function, as in Chipman et al. [2010]) to map the unconstrained function $f_0(\boldsymbol{X})$ to the natural parameter $p(\boldsymbol{X})$. There are usually several different choices for the link function. As we will see in Section 3.5.2, the BART estimator might have different posterior concentration rates depending on which link function is used to transform the function f_0 to the natural parameter of the response distribution.

The univariate regression and the two-class classification problem considered in the original BART paper [Chipman et al., 2010] and many of its important extensions, such as the multi-class classification and the log-linear BART [Murray, 2020] for categorical and count responses can be formulated as special cases of (3.1). The specific forms of the functions h, g, η and T for the continuous (regression), binary (two-class classification)), categorical (multi-class classification) and count response variables are given in Table 3.1.

Table 3.1: Univariate Regression (column 2), Two-class Classification (column 3), Multi-class Classification (column 4) and Count Regression (column 5), as special cases of the Generalized BART model. Ψ denotes the *logistic* function and $\mathcal{B}(\cdot)$ refers to the $Binomial(1,\cdot)$ distribution. Φ denotes the $Softmax$ function and $\mathcal{M}(\cdot)$ denotes the $Multinomial(1;\cdot)$ distribution. $\left(\{\mathbb{I}\{Y=i\}\}_{i=1}^{p}\right)'$ denotes the row vector where the i-th coordinate equals to one if $\boldsymbol{Y}$ belongs to class i and zero otherwise. $\mathcal{P}$ denotes the Poisson distribution.

Response ($\boldsymbol{Y}$)	Continuous	Binary	Categorical	Count
Dist.($\boldsymbol{Y}$)	$\mathcal{N}\left(f_0(\boldsymbol{X}),\sigma^2\right)$	$\mathcal{B}\left(\Psi(f_0(\boldsymbol{X}))\right)$	$\mathcal{M}\left(\Phi(f_0(\boldsymbol{X}))\right)$	$\mathcal{P}\left(\exp\left(f_0(\boldsymbol{X})\right)\right)$
$h(\boldsymbol{Y})$	$1/\sqrt{2\pi\sigma}$	1	1	$1/Y!$
$g\left(f_0(\boldsymbol{X})\right)$	$\exp\left(-f_0(\boldsymbol{X})^2/\sigma^2\right)$	$1-\Psi(f_0(\boldsymbol{X}))$	1	$\exp\left(-\exp\left(f_0(\boldsymbol{X})\right)\right)$
$\eta\left(f_0(\boldsymbol{X})\right)$	$(f_0(\boldsymbol{X}),1)$	$f_0(\boldsymbol{X})$	$f_0(\boldsymbol{X})$	$f_0(\boldsymbol{X})$
$T(\boldsymbol{Y})$	$\left(2Y/\sigma^2,-Y^2/\sigma^2\right)$	Y	$\left(\{\mathbb{I}\{Y=i\}\}_{i=1}^{p}\right)'$	Y
$f_0(\boldsymbol{X})$	$\mathbb{R}^q\to\mathbb{R}$	$\mathbb{R}^q\to\mathbb{R}$	$\mathbb{R}^q\to\mathbb{R}^{p-1}$	$\mathbb{R}^q\to\mathbb{R}$

Just as in the BART framework, in the generalized model also, a regression tree is used to reconstruct the unknown function $f_0:\mathbb{R}^q\to\mathbb{R}^D$ via a mapping $f_{\mathcal{T},\boldsymbol{\beta}}:[0,1]^q\to\mathbb{R}^D$ so that $f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{X})\approx f_0(\boldsymbol{X})$ for $\boldsymbol{X}\notin\{\boldsymbol{X}_i\}_{i=1}^{n}$. Each such mapping is essentially a step function of the form (2.1), supported on a tree-shaped partition $\mathcal{T}=\{\Omega_k\}_{k=1}^{K}$ and specified by a vector of step heights $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_K)'$. The vector $\beta_k\in\mathbb{R}^p$ represents the value of the expected response inside the k-th cell of the partition Ω_k. Just as in the classical BART model [Chipman et al., 2010], each partition $\mathcal{T}$ consists of rectangles obtained by recursively bisecting the predictor space by an axis-parallel surface. Partitioning is intended to increase within-node homogeneity (while simultaneously promoting inter-node heterogeneity) of outcomes. While depending on the number of offsprings, there are multiple tree topologies described in literature, we focus on only *binary* tree partitions, where each internal node (box) is split into exactly two children. A formal definition of a Binary Tree Partition is given in Definition 2.2.1 in Chapter 2.

Bayesian Additive trees consist of an ensemble of multiple shallow trees, each of which

is intended to be a weak learner [Zhou, 2012], geared towards addressing a slightly different aspect of the prediction problem. These trees are then woven into an *additive* forest mapping of the form (2.2), where each $f_{\mathcal{T}_t,\boldsymbol{\beta}_t}(\boldsymbol{x})$ is of the form (2.1), $\mathcal{E} = \{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$ is an ensemble of T trees and $\boldsymbol{B} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}'$ is a collection of jump sizes corresponding to the T trees. The next step is to assign a suitable prior distribution on the approximating function space given by equation (2.3) in Chapter 2.

Since each individual member of the approximating space is a step function of the form (2.2), supported on a Bayesian additive forest, the prior distribution should include three components: (i) a prior $\pi(T)$ on the number of trees T in the ensemble, (ii) a prior on individual tree partitions $\pi(\mathcal{T})$ and their collaboration within the ensemble and (iii) given a single tree partition a prior $\pi(\boldsymbol{\beta} \,|\, \mathcal{T})$ has to be imposed on the individual step heights.

In this chapter we follow the recommendation by Chipman et al. [2010] and assume the number of trees T to be fixed at a large value (e.g. $T = 200$ for regression and $T = 50$ for classification). This is equivalent to assigning a degenerate prior distribution on T, where all probability mass is concentrated on a single positive integer. Alternatively, one can also assign a prior with higher dispersion, as in Rockova et al. [2020] and Linero and Yang [2017] and replicate the steps of the proofs in Section 3.7 with minor modifications. Given the total number of trees in the ensemble, individual trees are assumed to be independent and identically distributed with some distribution $\pi(\mathcal{T})$. This reduces the prior on the ensemble to be of the form (2.4). The specific forms of the priors $\pi(\mathcal{T})$ and $\pi(\boldsymbol{\beta} \,|\, \mathcal{T})$ are described below. We will see that the scope of both these priors are wider than their analogous versions considered for the regression problem in Chapter 2.

Prior on Partitions $\pi(\mathcal{T})$

We consider two distinct prior distributions on the partitions $\pi(\mathcal{T})$. The first prior we consider is a variation of the Bayesian CART prior by Chipman et al. [1998]. We discussed the original form of this prior in Section 2.2. Starting with a single leaf (a root node)

$[0,1]^q$, every terminal node, say Ω_t, is split with a probability $p_{split}(\Omega_t)$, which is typically a decreasing function of $d(\Omega_t)$, the depth of the node Ω_t. In the last chapter we showed that the prior decay rate condition (C2) for near-optimal posterior concentration is not satisfied by (2.5), the splitting probability proposed in the original BART model [Chipman et al., 2010]. Therefore as per the recommendation in Chapter 2, we use $p_{split}(\Omega_t) \propto \alpha^{-d(\Omega_t)}$ instead, for some $0 < \alpha < 1/2$.

However one thing to keep in mind regarding this modification of the original BART prior is that the prior decay rate condition (C2) adapted from Ghosal et al. [2007] is only a *sufficient* but NOT a *necessary* condition for near-minimax posterior concentration. So the prior with (2.5) might still have a near-optimal posterior concentration rate, even if it hasn't been proved yet. In fact empirical evidence favors this proposition. Given that the node Ω_t splits, the splitting rule and the split point are chosen uniformly at random from the available directions and available data values, as described in Section 2.2.

Another alternative prior we consider is the Bayesian CART prior proposed by Denison et al. [1998]. The prior on individual Bayesian trees is assigned conditional on the number of terminal nodes/ leaves K and all prior probability is concentrated on the set of all *valid* tree partitions, as defined below (Definition 3.1 of Rockova et al. [2020]):

Definition 3.2.1. *Denote by* $\mathbf{\Omega} = \{\Omega\}_{k=1}^{K}$, *a partition of* $[0,1]^p$, *We say that* $\mathbf{\Omega}$ *is valid if*

$$\mu(\Omega_k) \geq \frac{C}{n} \quad \forall k = 1, \ldots, K \tag{3.2}$$

for some $C \in \mathbb{N} \setminus \{0\}$.

Valid partitions have non-empty cells, where the allocation does not need to be balanced. Now the prior on tree partitions is specified as follows:

1. The number of leaves in a tree K follows a Poisson distribution with parameter $\lambda > 0$

$$P(K) = \frac{\lambda^K}{(e^\lambda - 1)K!}, \qquad k = 1, 2, \ldots \tag{3.3}$$

2. Given the number of leaves K, a tree is chosen uniformly at random from the set of all available *valid* tree-partitions with K leaves. Number of valid tree partitions is given by

$$\Delta(V_K) = \frac{q^{K-1} n!}{(n-K+1)!} \tag{3.4}$$

This is a slightly modified version of the original prior proposed by Denison et al. [1998]. This modified version was used by Rockova et al. [2020] to derive posterior concentration rates for the BART estimator under this prior.

3. At each node, the splitting rule consists of picking a split variable j uniformly at random from the available directions $\{1, \ldots, q\}$ and picking a split point c, also uniformly at random from the available data values $x_{1j}, \ldots, x_{nj}$.

Prior on Step Heights $\pi(\boldsymbol{\beta} \,|\, \mathcal{T})$

We impose a broad class of priors on the step heights that incorporate the corresponding component of the classical BART model as a special case. Given a tree partition $\mathcal{T}_t$ with K_t steps, Chipman et al. [2010] considers identically distributed independent Gaussian jumps: with mean 0 and variance σ^2, after shifting and rescaling the response variables $\boldsymbol{Y}_i$'s so that the observed transformed values range from -0.5 to 0.5. Then they impose a conjugate Gaussian prior $\beta_{tj} \sim N(0, \sigma^2)$, where $\sigma = 0.5/k\sqrt{T}$ for some suitable value of k. The objective is to ensure that the prior assigns substantial probability to the observed range of the $\boldsymbol{Y}_i$'s. In the G-BART set-up we assume that the step heights $\beta_{tj} \overset{i.i.d}{\sim} F_\beta$, where F_β is any general distribution with the following property: For some constants C_1, C_2, C_3 such that $C_1 > 0$, $0 < C_2 \leq 2$ and $C_3 > 0$,

$$F_\beta(\|\beta\|_\infty \leq t) \gtrsim \left(e^{-C_1 t^{C_2}} t\right)^p \quad \text{for } 0 < t \leq 1 \tag{3.5}$$

$$F_\beta(\|\beta\|_\infty \geq t) \lesssim e^{-C_3 t} \quad \text{for } t \geq 1 \tag{3.6}$$

where $\|\|\|_\infty$ represents the L_∞ norm and $F_\beta(\|\beta\|_\infty \geq t)$ denotes the tail probability of the distribution on the step heights $\beta \in \mathbb{R}^p$. Both the multivariate Gaussian $\mathcal{N}_p(\mathbf{0}, \mathbb{I}_q)$ and the multivariate Laplace $\mathcal{L}_p(\mathbf{0}, \mathbb{I}_q)$ distribution come from this family of distributions and so do any sub-Gaussian distributions. A proof of this statement is provided in the supplement. We will see in Section 3.4.1 and Section 3.4.3 that these conditions are *sufficient* to guarantee that the G-BART estimator has a near-optimal posterior concentration rate. Student's t-distributions with small degrees of freedom usually violate condition (3.6).

However we should note that the conditions (3.5)-(3.6), although *sufficient*, are not *necessary* conditions and distributional assumptions on the step sizes that do not satisfy these conditions, might still guarantee a near-optimal posterior concentration rate. For instance, in a binary classification problem, the response $\boldsymbol{Y} \mid \boldsymbol{X} \sim \text{Bernoulli}(\pi(\boldsymbol{X}))$. Suppose we approximate $\pi(\boldsymbol{X})$ by a step function $\pi(\boldsymbol{X}) = \sum_{k=1}^{K} \pi_k \mathbb{I}\{\boldsymbol{X} \in \Omega_k\}$ on a tree-partition $\{\Omega_k\}_{k=1}^{K}$ and assign prior $\pi_k \overset{i.i.d}{\sim} \text{Beta}(2, 2)$ on the step-heights. We will see in Section 3.5.2, that this estimator has a near-optimal posterior concentration rate, even though $\text{Beta}(2, 2)$ violates condition (3.5). A proof of this statement is given in Section 3.7.

3.3 Posterior Concentration Revisited

As discussed in Chapter 2, posterior concentration statements are a prominent artifact in Bayesian nonparametrics, where the primary motivation is to examine the quality of a Bayesian procedure, by studying the learning rate of its posterior (defined in (2.6) in Chapter 2), i.e. the rate at which the posterior distribution, centralizes around the truth as the sample size $n \to \infty$. Ideally under a suitable prior, the posterior should put most of its probability mass around a small neighborhood of the true function and as the sample size increases, the diameter of this neighborhood should go to zero at a fast pace. Formally speaking, for a given sample size n, if we examine an ε_n-neighborhood of the true function $\mathcal{A}_{\varepsilon_n}$ (such as in (2.7) for the regression scenario), where $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, we should

expect

$$\Pi(\mathcal{A}_{\varepsilon_n}^c \mid \boldsymbol{Y}^{(n)}) \to 0 \quad \text{in } \mathbb{P}_{f_0}^{(n)}\text{-probability as } n \to \infty, \tag{3.7}$$

where $\mathcal{A}_{\varepsilon_n}^c$ denotes the complement of the neighborhood $\mathcal{A}_{\varepsilon_n}$.

In the context of G-BART, given observed data $\boldsymbol{Y}^{(n)} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)'$, we are interested in evaluating whether the posterior concentrates around the true likelihood $\mathbb{P}_{f_0}^{(n)} = \prod_{i=1}^n P_{f_0}^i$ at a near-minimax rate, where $P_{f_0}^i = P_{f_0}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ is of the form (3.1), for $i = 1, \ldots, n$. Following the suggestions of Ghosal et al. [2007], we look at the smallest H_n-neighborhoods around $\mathbb{P}_{f_0}^{(n)}$ that contain the bulk of the posterior probability, where H_n denotes the Hellinger distance between the true and estimated density, averaged over the observed sample (defined in Section 3.1). Specifically, for a diameter $\varepsilon > 0$ define

$$\mathcal{A}_\varepsilon = \{f \in \mathcal{F} : H_n(P_f, P_{f_0}) \leq \varepsilon\} \tag{3.8}$$

In Chapter 2 we considered the case where $\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathcal{N}(f_0(\boldsymbol{X}), \sigma^2)$, which reduces the H_n-neighborhood of $P_{f_0}^{(n)}$ to an L_2-neighborhood of f_0. However for a general exponential family density of the form (3.1), these two neighborhoods are not always equivalent.

Just as in Chapter 2, we will follow the strategy proposed by Ghosal et al. [2007] for deriving the posterior concentration rate of G-BART. Let us first define a sequence of approximating sieves $\mathbb{F}_n \subseteq \mathbb{F}_{n+1}$, such that $\mathbb{F}_n \to \mathcal{F}$ as $n \to \infty$. For a given $n \in \mathbb{N} \setminus \{0\}$, define

$$\mathbb{F}_n = \mathcal{F}_n \cap \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty \leq C_\beta^n\}, \tag{3.9}$$

where $\mathcal{F}_n$ is as defined in (2.24) and $\{C_\beta^n\}_{n \geq 1}$ is an increasing positive sequence such that $C_\beta^n \lesssim n^M$ for some $M > 0$. The construction of sieves will depend on which exponential family distribution the response $\boldsymbol{Y}$ comes from. For a given n, the approximating sieve (3.9) considered here is a subset of the approximating sieve (2.24) considered in Chapter 2. The additional constraint is required to control the rate of variation of the likelihood $\mathbb{P}_f^{(n)}$ within

each sieve. We will justify this construction under "Assumption 1" in Section 3.4.

Theorem 4 of Ghosal et al. [2007] demonstrates that the statement (3.7) can be proved by verifying three sufficient conditions, which albeit being analogous to the conditions (2.9)-(2.11), are significantly wider in scope. The first condition, henceforth referred to as the "entropy condition" specifies that

$$\sup_{\varepsilon > \varepsilon_n} \log N\left(\tfrac{\varepsilon}{36}; \mathbb{F}_n \cap \mathcal{A}_\varepsilon; H_n\right) \lesssim n\,\varepsilon_n^2, \tag{C1}$$

where $N(\varepsilon; \Omega; d)$ denotes the ε-covering number of a set Ω for a semimetric d, i.e. the minimal number of d-balls of radius ε needed to cover the set Ω.

The second condition requires that the prior puts enough mass around the true likelihood $\mathbb{P}_{f_0}^{(n)}$, meaning that for a given sample size $n \in \mathbb{N} \setminus \{0\}$ and for some $d > 2$,

$$\Pi(f \in \mathcal{F} : K_n(f, f_0) \vee V_n(f, f_0) \leq \varepsilon_n^2) \gtrsim e^{-d\,n\,\varepsilon_n^2}, \tag{C2}$$

where K_n and V_n are the Kullback-Leibler divergence and the variation, averaged over the observed data points, as defined in Section 3.1.

The final condition, referred to as the "prior decay rate condition" stipulates that the sequence of sieves $\mathbb{F}_n \uparrow \mathcal{F}$ captures the entire parameter space with increasing accuracy, in the sense that the complementary space $\mathcal{F} \backslash \mathbb{F}_n$ has negligible prior probability mass for large values of n.

$$\Pi(\mathcal{F} \backslash \mathbb{F}_n) = o(e^{-(d+2)\,n\,\varepsilon_n^2}) \tag{C3}$$

The conditions (C1), (C2) and (C3) are generalized versions of the conditions (2.9), (2.10) and (2.11) respectively described in Section 2.3. The key distinction is that unlike the conditions in Chapter 2, the above statements are formulated in terms of the average Hellinger distance H_n, the Kullback-Leibler divergence K_n and the variation V_n between the estimated likelihood $\mathbb{P}_f^{(n)}$ and the true likelihood $\mathbb{P}_{f_0}^{(n)}$ of the response variable $\boldsymbol{Y}$. For

continuous regression, all three of these divergences can be reduced in terms of the average L_2 distance between the true and estimated mean parameters of the model, thus giving rise to the conditions described in Section 2.3. In essence, the conditions (C1)-(C3) described above can potentially be employed to study the posterior concentration rates of *any* density estimator, while the analogous conditions considered in the previous chapter are specific to the Gaussian regression setup.

3.4 Main Results

In this section we describe our main theoretical findings, which describe the posterior concentration rates of the generalized Bayesian trees and their additive ensembles (G-BART), when the true function f_0 connecting the response $\boldsymbol{Y}$ with the covariates $\boldsymbol{X}$, is either (a) a step function (Theorem 3.4.1), or (b) a monotone function (Theorem 3.4.2), or (c) a ν-Hölder continuous function with $0 < \nu \leq 1$ (Theorem 3.4.3). We start by proving that the Generalized Bayesian *tree* estimator, described in Section 3.2, has a near-optimal rate of posterior concentration, under suitable conditions. These results are then extended to the G-BART estimator on *additive tree ensembles or forests*. We make two important assumptions: the first assumption (subsequently referred to as Assumption 1), given below restricts the distribution of the response variable $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\} \in \mathbb{R}^p$ to a specific class of exponential family distributions while the second assumption (subsequently referred to as Assumption 2) concerns the spread of the covariates $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \in \mathbb{R}^q$.

Assumption 1: Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \sim P_f$, where P_f denotes a probability density function of the form (3.1), such that, $\eta(z) = z$ and there exists strictly increasing positive sequences $\{C_g^n\}_{n\geq 1}$ and $\{C_\beta^n\}_{n\geq 1}$, such that

$$\left|\frac{\nabla g(\boldsymbol{\beta})}{g(\boldsymbol{\beta})}\right| \leq C_g^n \mathbf{1}_p, \quad \forall \boldsymbol{\beta} \in B_n = \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty \leq C_\beta^n\right\}, \tag{3.10}$$

where $\mathbf{1}_p = (1, \ldots, 1) \in \mathbb{R}^p$ denotes a p-dimensional vector of ones. We assume $\{C_g^n\} \vee \{C_\beta^n\} \lesssim n^M$ for some $M > 0$. The significance is that the function $g(\cdot)$ should not change too rapidly, and the higher the sample size the larger the rate of change is allowed. The above assumption is satisfied by most distributions commonly used in the regression and classification settings, as will be demonstrated in Section 3.5.2.

The second assumption is the same as the design condition described in Section 2.6 and hence omitted. This condition ensures that for a k-d tree partition, the maximum diameter of the cells does not exceed a constant multiple of the average cell diameters.

3.4.1 Results on Step-Functions

Let us suppose f_0 is a step function supported on an axes-paralleled partition $\{\Omega_k\}_{k=1}^{K_0}$. For any such step function f_0, we define the *complexity of* f_0, as follows:

Definition 3.4.1. *Complexity of f_0 is defined as the smallest K such that there exists a partition $\{\Omega_k\}_{k=1}^K$ with K cells, for which the step function $f(x) = \sum_{k=1}^K \beta_k \mathbb{I}\{x \in \Omega_k\}$ can approximate f_0 without any error, for some step heights $(\beta_1, \ldots, \beta_K) \in \mathbb{R}^K$.*

This complexity number, denoted by K_{f_0}, depends on the true number of step K_0, the diameter of the intervals $\{\Omega_k\}_{k=1}^{K_0}$, and the number of covariates q. An illustration of such approximation by a k-d tree can be found in van der Pas and Rockova [2017] for $q = 1$. The actual minimax rate for approximating such piecewise-constant functions f_0 with $K_0 > 2$ pieces, is $n^{-1/2}\sqrt{K_0 \log(n/K_0)}$ [Gao et al., 2017]. The following theorem shows that the posterior concentration rate of G-BART is almost equal to the minimax rate, except that K_0 gets replaced by K_{f_0}. The discrepancy is an unavoidable consequence of the fact that the true number of steps K_0 is unknown. Had this information been available, the G-BART estimator would have attained the exact minimax rate.

Theorem 3.4.1. *If we assume that the distribution of the step-sizes satisfies* (3.5) *and* (3.6)*, then under Assumptions 1 and 2 with $q \lesssim \sqrt{\log n}$, the Bayesian Tree estimator satisfies the*

following property:

If f_0 is a step-function, supported on an axes-paralleled partition, with complexity $K_{f_0} \lesssim \sqrt{n}$ and $\|f_0\|_\infty \lesssim \sqrt{\log n}$, then with $\varepsilon_n = n^{-1/2}\sqrt{K_{f_0}\log^{2\gamma}\left(n/K_{f_0}\right)}$ and $\gamma > 1/2$,

$$\Pi\left(f \in \mathcal{F} : H_n(\mathbb{P}_f, \mathbb{P}_{f_0}) > \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

in $\mathbb{P}_{f_0}^{(n)}$-probability, as $n, q \to \infty$.

The above statement is true for both tree priors considered in literature: the prior by Denison et al. [1998] and a modified version of the prior by Chipman et al. [1998] with $p_{split}(\Omega_t) = \alpha^{d(\Omega_t)}$ for some $1/n \leq \alpha < 1/2$.

Proof. Proof is given in Section 3.7. □

From the above theorem we see that the Generalized BART prior adapts to the *unknown* complexity of the true function f_0, without imposing any additional constraints on the number of leaves in the estimating partition space.

3.4.2 Results on Monotone Functions

An important implication of Theorem 3.4.1 is that posterior concentration results on step functions can potentially build the foundation for similar results on broader class of functions, aided by the "simple function approximation theorem" [Stein and Shakarchi, 2009], which states that for any measurable function f on $\mathcal{E} \subseteq \mathbb{R}^q$, there exists a sequence of step functions $\{f_k\}$ which converges point-wise to f almost everywhere [Stein and Shakarchi, 2009]. As a corollary to this theorem, we can derive the following result on the set of all monotone functions. A function $f_0 : \mathbb{R}^q \to \mathbb{R}$ is defined as monotone increasing (or decreasing) if $f_0(\boldsymbol{x}_1) \geq f_0(\boldsymbol{x}_2)$ (or, $f_0(\boldsymbol{x}_1) \leq f_0(\boldsymbol{x}_2)$) for all $\boldsymbol{x}_1, \boldsymbol{x}_2$ such that every coordinate of $\boldsymbol{x}_1$ is greater than or equal to the corresponding coordinate of $\boldsymbol{x}_2$.

Lemma 3.4.1. *Any* **monotone** *bounded function f_0 can be approximated with arbitrary precision ε, by a step function supported on a k-d tree partition with number of leaves*

$K_{f_0}(\varepsilon) \geq \lceil 1/\varepsilon \rceil$. *We define* $K_{f_0}(\varepsilon)$ *to be the complexity of the monotone function* f_0 *with respect to* $\varepsilon > 0$.

The complexity $K_{f_0}(\varepsilon)$ also depends on the dimension of the domain q as well as on the magnitude of the true function $\|f_0\|_\infty$. This paves the way for deriving the posterior concentration rate of G-BART when the true function $f_0(\cdot)$ connecting the covariates $\boldsymbol{X}$ with a univariate response $\boldsymbol{Y}$ is a monotone function. The minimax rate of estimation for such densities is $n^{-1/(2+q)}$ [Biau and Devroye, 2003]. The following theorem states that the posterior concentration rate of G-BART equals to this optimum rate up to a logarithmic function, provided that the magnitude of the true function f_0 is not "too large".

Theorem 3.4.2. *If we assume that the distribution of the step-sizes satisfies* (3.5) *and* (3.6)*, then under Assumptions 1 and 2 with* $q \lesssim \sqrt{\log n}$*, the Bayesian Tree estimator satisfies the following property:*

If the true function $f_0 : \mathbb{R}^q \to \mathbb{R}$ *is monotonic on every coordinate, with* $\|f_0\|_\infty \lesssim \sqrt{\log n}$*, then with* $\varepsilon_n = n^{-1/(2+q)}\sqrt{\log n}$*,*

$$\Pi\left(f \in \mathcal{F} : H_n(\mathbb{P}_f, \mathbb{P}_{f_0}) > \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

in $\mathbb{P}_{f_0}^{(n)}$*-probability, as* $n, q \to \infty$.

The above statement is true for both tree priors considered in literature: the prior by Denison et al. [1998] and a modified version of the prior by Chipman et al. [1998] with $p_{split}(\Omega_t) = \alpha^{d(\Omega_t)}$ *for some* $1/n \leq \alpha < 1/2$.

Proof. The first step of the proof involves finding an approximating step-function $\widehat{f}_0$ by Lemma 3.4.1, such that $\|f_0 - \widehat{f}_0\|_{2,n} < \varepsilon_n/2$. The rest of the proof follows by retracing the steps as in the proof of Theorem 3.4.3 given in Section 3.7. □

The above result demonstrates that the Generalized BART model adapts to monotonic patterns in the true function f_0, without any additional prior assumptions. In fact this the-

oretical observation is in line with the empirical evidence obtained in Section 4.4. Chipman et al. [2016] proposed a Monotone-BART model, where the step-heights are constrained to exhibit a monotone pattern over consecutive cells, geared towards high dimensional monotone function estimation. They demonstrated that in multiple simulated and benchmark datasets, the unconstrained BART model has comparable performance with the specially designed Monotone BART model. Theorem 3.4.2 provides a frequentist theoretical justification for this phenomenon.

3.4.3 Results on Hölder Continuous Functions

This section describes the posterior concentration results on G-BART when the true function f_0 connecting $\boldsymbol{X}$ with $\boldsymbol{Y}$ is a ν-Hölder continuous function with $0 < \nu \leq 1$. Rockova et al. [2020] and Rockova and Saha [2019] proved that the posterior concentration rates of the BART model (under the priors of Denison et al. [1998] and Chipman et al. [2010] respectively) are equal to $n^{-\alpha/(2\alpha+q)}$, the minimax rate of estimation for such functions [Stone, 1982], except for a logarithmic factor. These results can be derived as direct corollaries of the following theorem for G-BART, when $\boldsymbol{Y}$ is a univariate continuous response and the step-sizes are assumed to follow a Gaussian distribution. Thus the following result can be treated as a generalization of Theorem 2.7.1 given in Chapter 2, for higher dimensions and wider class of distributions on the response as well as the approximating step heights.

Theorem 3.4.3. *If we assume that the distribution of the step-sizes satisfies* (3.5) *and* (3.6)*, then under Assumptions 1 and 2 with* $q \lesssim \sqrt{\log n}$*, the Bayesian Tree estimator satisfies the following property:*

If f_0 *is a* ν*-Hölder continuous function with* $0 < \nu \leq 1$*, where* $\|f_0\|_\infty \lesssim \sqrt{\log n}$*, then with* $\varepsilon_n = n^{-\alpha/(2\alpha+q)}\sqrt{\log n}$*,*

$$\Pi\left(f \in \mathcal{F} : H_n(\mathbb{P}_f, \mathbb{P}_{f_0}) > \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

in $\mathbb{P}_{f_0}^{(n)}$*-probability, as* $n, q \to \infty$.

The above statement is true for both tree priors considered in literature: the prior by Denison et al. [1998] and a modified version of the prior by Chipman et al. [1998] with $p_{split}(\Omega_t) = \alpha^{d(\Omega_t)}$ *for some* $1/n \leq \alpha < 1/2$.

Proof. Proof is given in Section 3.7. □

3.4.4 Extensions to Tree Ensembles

Theorems 3.4.1, 3.4.2 and 3.4.3 demonstrate that generalized Bayesian trees have a near-optimum posterior concentration rate under suitable conditions. These results can be easily extended to tree ensembles or generalized BART models. For G-BART, one needs to find an approximating *tree ensemble* and show that it has enough prior support. The approximating ensemble can be constructed by Lemma 10.1 of Rockova et al. [2020]. This consists of $\widehat{\mathcal{E}} = \{\widehat{\mathcal{T}}_1, \ldots, \widehat{\mathcal{T}}_T\}$ tree partitions obtained by chopping of branches of $\widehat{\mathcal{T}}$, the approximating tree considered in the proofs of Theorems 3.4.1, 3.4.2 and 3.4.3. The number of trees T is fixed and assumed to satisfy $1 \leq T \leq \widehat{K}/2$. The default BART choice $T = 200$ [Chipman et al., 2010] safely satisfies this as long as $p > 9$. The little trees $\widehat{\mathcal{T}}_t$ have $\widehat{K}^t$ leaves and satisfy $\log_2 \widehat{K} + 1 \leq \widehat{K}^t \leq \widehat{K}$ (depending on the choice of T). Using the fact that the trees are independent a-priori (from (2.4)) and that T is fixed, we then obtain analogous theorems for G-BART following similar steps as in the proofs of Theorems 3.4.1, 3.4.2 and 3.4.3.

However one thing to keep in mind is that in most empirical applications tree ensembles/forests provide significantly superior out-of-sample predictive performance compared to a single tree. Even though not apparent in the previous paragraph, this distinction can also be demonstrated theoretically by comparing the posterior concentration rates for single trees and forests when the true function f_0 is of additive nature and the individual component functions have different degrees of smoothness (e.g. Theorem 6.1 of Rockova et al. [2020] and Theorem 4 of Linero and Yang [2017]). This issue has been discussed in Remark 2.7.1.

3.5 Implications

The primary significance of Theorems 3.4.1, 3.4.2 and 3.4.3 is that these results provide a frequentist theoretical justification for superior empirical performance of generalized Bayesian trees, claiming that the posterior concentrates around the truth at a near-optimal learning rate. Since this generalized framework encompasses most prominent regression and classification tasks, the theoretical implications transcend to these examples as well. Another important consequence of these results is that, as a direct corollary, we can show that the generalized Bayesian trees and forests favor simpler partitions and hence are resilient to overfitting. Another implication of posterior concentration statements as in Theorems 3.4.1, 3.4.2 and 3.4.3 is that they can build the foundation for obtaining uncertainty quantification statements such as semiparametric Bernstein-von Mises type theorems.

3.5.1 Parsimony of G-BART

In Chapter 2 we demonstrated that the BART model for regression is resilient to overfitting (Corollary 2.7.1). As a by-product of the theoretical results discussed in Section 3.4, we can also obtain similar statements for the generalized Bayesian trees and its additive ensembles (G-BART). In fact, Corollary 2.7.1 can be derived as a special case of part (iii) of the following corollary, when both the response variable and the step-heights are restricted to follow univariate Gaussian distributions.

Corollary 3.5.1.

(i) Under the assumptions of Theorem 3.4.1 we have $\Pi\left(K \gtrsim K_{f_0} \mid \boldsymbol{Y}^{(n)}\right) \to 0$ *in* $\mathbb{P}_{f_0}^{(n)}$*-probability, as* $n, q \to \infty$.

(ii) Under the assumptions of Theorem 3.4.2 we have $\Pi\left(K \gtrsim n^{q/(2+q)} \mid \boldsymbol{Y}^{(n)}\right) \to 0$ *in* $\mathbb{P}_{f_0}^{(n)}$*-probability, as* $n, q \to \infty$.

(iii) Under the assumptions of Theorem 3.4.3 we have $\Pi\left(K \gtrsim n^{q/(2\nu+q)} \mid \boldsymbol{Y}^{(n)}\right) \to 0$ *in*

$\mathbb{P}_{f_0}^{(n)}$-*probability, as* $n, q \to \infty$.

Proof. The proofs of (i), (ii) and (iii) follow from Lemma 1 of Ghosal et al. [2007], in conjunction with the proofs of Theorems 3.4.1, 3.4.2 and 3.4.3 respectively. □

In other words, the posterior distribution on the number of leaves in a generalized Bayesian tree does not exceed the optimal number of splits by more than a constant multiple, when the true function f_0 is either a step-function, a monotone function or a ν-Hölder continuous function with $0 < \nu \leq 1$. These results can be easily generalized to tree ensembles, i.e for G-BART following the technique discussed in Section 3.4.4. The only exception is that $\Pi\left(K \gtrsim z \mid \boldsymbol{Y}^{(n)}\right)$ gets replaced by $\Pi\left(\cup_{t=1}^{\mathcal{T}}\{K_t \gtrsim z\} \mid \boldsymbol{Y}^{(n)}\right)$ for all three parts.

3.5.2 Some Examples and Exceptions

In this section we demonstrate the breadth of applicability of Theorems 3.4.1, 3.4.2 and 3.4.3 in proving analogous theoretical results for a wide range of BART models. As direct corollaries to these theorems, we show that the original BART model [Chipman et al., 2010], along with some of its commonly used variants (such as BART for multi-class classification and regression on count data) has a near-optimal posterior concentration rate.

Continuous Regression:

For the univariate regression setup, G-BART reduces to the original BART [Chipman et al., 2010] model when the step-heights are assumed to follow a Gaussian distribution. In Chapter 2, built upon the results of Rockova and Saha [2019], we showed that the posterior concentration rate of BART equals to the minimax rate, up to a logarithmic factor, when the function f_0 connecting the input $\boldsymbol{X}$ with the output $\boldsymbol{Y}$ is a ν-Hölder continuous function. In this section we demonstrate that the same result (Theorem 2.7.1) can be obtained as a direct corollary of Theorem 3.4.3. In addition, we show that similar posterior optimality statements are also valid for multivariate regression, when the true function f_0 is either a

step function, a monotone function or a ν-Hölder continuous function with $0 < \nu \leq 1$. For modelling a (multivariate) continuous response $\boldsymbol{Y}$, assume

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}(\boldsymbol{X}), \Sigma),$$

where Σ is a known positive definite matrix. As a special case, for univariate regression, Chipman et al. [2010] assumes $\sigma = 0.5/k$ for some suitable value of k. Since the natural parameter $\mu(\boldsymbol{X})$ is unconstrained, we use an identity link to connect $f_0(\boldsymbol{X})$ to $\boldsymbol{\mu}(\boldsymbol{X})$. Therefore $g(f_0(\boldsymbol{X})) = g(\boldsymbol{\mu}) = e^{-\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}/2}$, which satisfies (3.10) with $B_n = [-n, n]^p$ and $C_g^n = n\lambda(\Sigma)$, where $\lambda(\Sigma)$ denotes the maximum eigenvalue of Σ. Hence from Theorems 3.4.1, 3.4.2 and 3.4.3, we can conclude that for the (multivariate) continuous regression problem, the G-BART estimator has a near-minimax posterior concentration rate, provided that the true function f_0 connecting the input $\boldsymbol{X}$ with the output $\boldsymbol{Y}$ is either a step function, a monotone function or a ν-Hölder continuous function with $0 < \nu \leq 1$. This can be suitably adapted, with minor modifications, to demonstrate the near-optimaliy of the Monotone BART [Chipman et al., 2016] model which is specifically designed for high dimensional monotone function estimation.

Classification with Gaussian Step Heights:

For a multi-class classification problem with p classes, where the response variable $\boldsymbol{Y}$ is a categorical random variable with p categories, $\boldsymbol{Y}$ can be written as a p dimensional binary vector that has 1 at the l-th coordinate if $\boldsymbol{Y}$ belongs to category $l \in \{1, \ldots, p\}$ and 0 elsewhere. We can assume

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \text{Multinomial}(1; \boldsymbol{\pi}(\boldsymbol{X}))$$

for some $\boldsymbol{\pi} : \mathbb{R}^q \in (0,1)^p$ such that $\boldsymbol{\pi}'\mathbf{1}_p = 1$. The unrestricted function $f_0(\boldsymbol{X})$ can be transformed to the natural parameter $\pi(\boldsymbol{X})$ by a logistic (softmax) or an inverse-probit link

function [Chipman et al., 2010] denoted by $\Psi(\cdot)$, so that $\pi(\boldsymbol{X}) = \Psi(f_0(\boldsymbol{X}))$. In either case, the function $g(f_0(\boldsymbol{X})) = 1$ trivially satisfies condition (3.10). Hence from Theorem 3.4.1 and Theorem 3.4.3, we can conclude that for the multi-class classification problem, the G-BART estimator has a near-optimal posterior concentration rate, provided that the true function f_0 connecting the input $\boldsymbol{X}$ with the output $\boldsymbol{Y}$ is either a step function supported on an axes paralleled partition, or a ν-Hölder continuous function with $0 < \nu \leq 1$.

Classification with Dirichlet Step-Heights

For the same multi-class classification problem with p classes described above, an alternative prior specification is recommended by Denison et al. [1998]. Assume

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \text{Multinomial}(1; \boldsymbol{f_0}(\boldsymbol{X})), \tag{3.11}$$

where $\boldsymbol{f_0} = \big(f_{01}, \ldots, f_{0p}\big)' : \mathbb{R}^q \to (0,1)^p$ is a constrained function with $\boldsymbol{f_0}(\boldsymbol{X})'\mathbf{1}_p = 1$ for any $\boldsymbol{X} \in \mathbb{R}^q$. Each $f_{0l}(\cdot)$ can be approximated by a step function of the form

$$f_{\mathcal{T},P}(\boldsymbol{x}) = \sum_{k=1}^{K} P_k \mathbb{I}(\boldsymbol{x} \in \Omega_k) \tag{3.12}$$

on a tree-partition $\{\Omega_k\}_{k=1}^K$. Denison et al. [1998] assumes

$$P_k = \big(P_{k1}, \ldots, P_{kp}\big) \overset{i.i.d}{\sim} \text{Dirichlet}(\alpha_1, \ldots, \alpha_p), \tag{3.13}$$

where $\alpha_l > 0, \quad \forall l \in \{1, \ldots, p\}$.

For example, in a binary classification ($p = 2$) problem, we can assign prior $P_k \overset{i.i.d}{\sim} \text{Beta}(2,2)$ on the step-heights. The prior $\text{Beta}(2,2)$ violates condition (3.5). But we can show that this estimator has a near-optimal posterior concentration rate, even if we cannot conclude this from the results discussed in Section 3.4. A proof is given in Section 3.7. This demonstrates that the assumptions we make in Section 3.4 are merely *sufficient* but not *nec-*

essary conditions for proving that the generalized Bayesian tree (or G-BART) estimator has a near-optimum posterior concentration rate, and this optimality property goes far beyond the class of estimators discussed in this chapter.

Count Regression:

For count response variable, $\boldsymbol{Y} \sim Poisson\,[\lambda(\boldsymbol{X})]$ with $\lambda(\boldsymbol{X}) > 0$. There are several choices for the link function $\Psi(\cdot)$ to map the unconstrained function $f_0(\boldsymbol{X})$ to the constrained parameter $\lambda(\boldsymbol{X})$. The posterior concentration rate of the Generalized Bayesian tree estimator might differ depending on which link function is used. For example, if we use $\Psi(z) = \log(1+\exp(z))$, the Softplus link function, then $g(f_0(\boldsymbol{X})) = 1/(1+\exp(f_0(\boldsymbol{X}))$, trivially satisfies condition (3.10) and we can conclude that the generalized tree estimator has a near-minimax concentration rate from Theorems 3.4.1, 3.4.2 and 3.4.3.

In contrast, if we use $\Psi(z) = \exp(z)$ as the link function, then $g(f_0(\boldsymbol{X})) = \exp(-\exp(f_0(\boldsymbol{X})))$ does not satisfy the condition (3.10), when the true function f_0 is a ν-Hölder continuous function. Therefore we cannot apply Theorem 3.4.3 anymore to imply that the generalized tree estimator has a near-optimal rate of posterior concentration. When f_0 is a step function with complexity K_{f_0}, the condition (3.10) is satisfied with $B_n = [-K_{f_0}\log n, K_{f_0}\log n]$ and $C_g^n = n^{K_{f_0}}$. The posterior concentration rate becomes $\varepsilon_n = n^{-\frac{1-\alpha}{2}}\sqrt{K_{f_0}\log^{2\eta}(n/K_{f_0})}$ under the assumption $K_{f_0} \lesssim n^{\alpha}$ for some $0 < \alpha < 1$. This is slower than the near-optimal concentration rate $n^{-\frac{1}{2}}\sqrt{K_{f_0}\log^{2\eta}(n/K_{f_0})}$, if we use $\Psi(z) = \log(1+\exp(z))$, the Softplus link function, instead. This demonstrates the need for choosing suitable link functions in empirical applications.

3.6 Discussion

So far in this chapter we have examined G-BART, a general framework for Bayesian Additive Regression Tree Models that encapsulates a broad range of regression and classification

tasks, where the response variables are allowed to follow any distribution belonging to a wide subset of the exponential family. The near-optimal posterior concentration rate of G-BART proved in this paper supports the empirical success of BART and its variants, from a theoretical perspective. The significance of our work is three-fold: firstly it extends the theoretical understanding on BART beyond the univariate Gaussian regression set-up. As direct corollaries of the main results discussed in Section 3.4, we have shown that the Bayesian trees and forests have a near-optimal posterior concentration rate for a wide range of regression and classification problems. Moreover the posterior concentration statements for step-functions discussed in this chapter pave the way for deriving similar statements for broader class of functions, beyond the realm of Hölder continuity, that has till date dominated the theory for BART. Theoretical results on monotone functions discussed in this section demonstrates the potential of this approach. Finally, these results also build the foundation for Bernstein-von-Mises-type theorem and /or uncertainty quantification statements for a wide variety of BART models, opening up several interesting avenues for future research. Among empirical implications, we have established the need for careful modeling choices such as selecting appropriate link functions. The theoretical results also substantiate the scope of a wider variety of distributions on approximating step-heights, that can prove advantageous for applications where extreme values are of special importance. These theoretical findings also provide strong motivation for exploring novel application areas for flexible BART-like models. In the later parts of this chapter, following the proof of main results discussed in the next section, we will discuss adaptations of BART to some interesting practical problems.

3.7 Proofs of Main Results

This section contains the proofs of the main results regarding posterior concentration, along with some additional results required for these proofs. We start by discussing some preliminary lemmas, followed by the proof of Theorems 3.4.1 and 3.4.3. Next we move on to the posterior concentration rate (with derivation) for the Bayesian classification tree with

Dirichlet step heights [Denison et al., 1998] discussed in Section 3.5.2.

Preliminary Results with Proof

Lemma 3.7.1. *The multivariate Gaussian $\mathcal{N}_p(\mathbf{0}, \mathbb{I}_p)$ and the multivariate Laplace $\mathcal{L}_p(\mathbf{0}, \mathbb{I}_p)$ distribution belong to the general family of distributions with CDF F_β that has the following property: For some $C_1 > 0$, $0 < C_2 \leq 2$ and $C_3 > 0$ and any $t > 0$,*

$$F_\beta(\|\beta\|_\infty \leq t) \gtrsim \left(e^{-C_1 t^{C_2}} t\right)^p \quad \text{for } t > 0 \tag{3.14}$$

$$F_\beta(\|\beta\|_\infty \geq t) \lesssim e^{-C_3 t} \quad \text{for } t \geq 1 \tag{3.15}$$

Proof. If $F_\beta = \mathcal{N}_p(\mathbf{0}, \mathbb{I}_p)$, then for any $t > 0$,

$$F_\beta(\|\beta\|_\infty \leq t) \gtrsim \left(e^{-t^2/2} \int_{-t}^{t} d\beta\right)^p \gtrsim e^{-pt^2/2} t^p$$

For $t \geq 1$

$$F_\beta(\|\beta\|_\infty \geq t) \lesssim \left(e^{-t^2/4} 2 \int_{t}^{\infty} e^{-z^2/4} dz\right)^p \lesssim e^{-C_3 t}$$

If $F_\beta = \mathcal{L}_p(\mathbf{0}, \mathbb{I}_p)$, then for any $t > 0$,

$$F_\beta(\|\beta\|_\infty \leq t) \gtrsim \left(e^{-t} \int_{-t}^{t} d\beta\right)^p \gtrsim e^{-pt} t^p$$

Also, for any $t \geq 0$,

$$F_\beta(\|\beta\|_\infty \geq t) = \frac{e^{-pt}}{2} < e^{-pt}$$

□

Lemma 3.7.2. *Let f and f_0 denote step functions of the form $f(\boldsymbol{X}) = \sum_{k=1}^{K} \beta_k \mathbb{I}(\boldsymbol{X} \in \Omega_k)$ and $f_0(\boldsymbol{X}) = \sum_{k=1}^{K} \beta_k^0 \mathbb{I}(\boldsymbol{X} \in \Omega_k)$ respectively, on a tree-shaped partition $\{\Omega_k\}_{k=1}^{K}$. Let P_f and P_{f_0} denote two probability densities belonging to an Exponential family distribution of*

the form

$$P_f(\boldsymbol{Y} \mid \boldsymbol{X}) = h(\boldsymbol{Y}) g\left[f(\boldsymbol{X})\right] \exp\left[\eta\left(f(\boldsymbol{X})\right)^T T(\boldsymbol{Y})\right], \tag{3.16}$$

with parameters f *and* f_0 *respectively. If* $\left|\frac{\nabla^T g(\boldsymbol{\beta})}{g(\boldsymbol{\beta})}\right| \leq C_g^n \mathbf{1}_p$*, for some positive sequence* $\{C_g^n\}_{n\geq 1}$*, then*

$$K_n(P_f, P_{f_0}) \vee V_n(P_f, P_{f_0}) \lesssim C_g^n \sum_{k=1}^{K} \left\|\beta_k - \beta_k^0\right\|_1 \tag{3.17}$$

$$\text{and} \quad H_n(P_f, P_{f_0}) \lesssim C_g^n \sum_{k=1}^{K} \left\|\beta_k - \beta_k^0\right\|_1 \tag{3.18}$$

Proof. Denoting $f_i = f(\boldsymbol{X}_i)$ and $f_{i0} = f_0(\boldsymbol{X}_i)$, we can write

$$\begin{aligned} K_n(P_f, P_{f_0}) &= \frac{1}{n}\sum_{i=1}^{n} g(f_i) \int h(\boldsymbol{Y}) \exp\left(f_i T(\boldsymbol{Y})\right) \left[\log \frac{g(f_i)}{g(f_{i0})} + \exp\left[(f_i - f_{i0})^T T(\boldsymbol{Y})\right]\right] d\boldsymbol{Y} \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[\log \frac{g(f_i)}{g(f_{i0})} + (f_i - f_{i0})^T \mathbb{E}\left[T(\boldsymbol{Y})\right]\right] \\ &= \sum_{k=1}^{K} \mu(\Omega_k) \left[\log \frac{g(\beta_k)}{g(\beta_k^0)} - \frac{\nabla^T g(\beta_k)}{g(\beta_k)}\left(\beta_k - \beta_k^0\right)\right] \end{aligned}$$

By triangle inequality and Taylor series approximation of $\log \nabla g(\beta_k)$ about β_k^0, we get

$$K_n(P_f, P_{f_0}) \lesssim \sup\left|\frac{\nabla^T g(\cdot)}{g(\cdot)}\right| \sum_{k=1}^{K} \left\|\beta_k - \beta_k^0\right\|_1 = C_g^n \sum_{k=1}^{K} \left\|\beta_k - \beta_k^0\right\|_1,$$

Similar technique works for $V_n(P_f, P_{f_0})$

Also, Since Hellinger metric is bounded from above by Kullback-Leibler divergence, H_n satisfies,

$$H_n(P_f, P_{f_0}) \lesssim C_g^n \sum_{k=1}^{K} \left\|\beta_k - \beta_k^0\right\|_1$$

□

Lemma 3.7.3. *Any bounded* **monotone** *function* f_0 *can be approximated with arbitrary*

precision ε_n, by a step function supported on a k-d tree partition with $\widehat{K} \geq \lceil 1/\epsilon_n \rceil$ leaves.

Proof. Without loss of generality, assume $0 \leq f_0(\cdot) \leq 1$. Partition interval $[0,1]$ by $0 = y_0 < y_1 < \cdots < y_k < \cdots < y_{K-1} < y_K = 1$, with $K = \lceil 1/\epsilon_n \rceil$. Then $|\, y_k - y_{k-1} \,| < \varepsilon_n$ and we can approximate $f_0(\boldsymbol{X})$ by the step function:

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} y_k \mathbb{I}\{\boldsymbol{X} \in \Omega_k\}, \quad \text{where} \quad \Omega_k = f^{-1}[y_{k-1}, y_k]$$

If f is monotone, $\Omega_k = \prod_{j=1}^{q} \{x_j \in \mathrm{I}_j\}$, where I_j is an interval and x_j denotes the j-th coordinate of $\boldsymbol{X} \in \mathbb{R}^q$.

Since any step function supported on an axis-paralleled partition has an equivalent step function supported on a k-d tree, we can approximate the axis paralleled partition $\{\Omega_k\}_{k=1}^{K}$ by a recursive binary tree partition $\{\widehat{\Omega}_k\}_{k=1}^{\widehat{K}}$ with number of leaves $\widehat{K} \geq K$. □

Proofs of Theorems 3.4.1 and 3.4.3

In this section we prove Theorem 3.4.1 and Theorem 3.4.3. Most steps in the proofs are identical and hence for simplicity we describe the common steps of the proofs together and mark the steps that are different by the corresponding theorem number. As discussed in Section 3.3, in order to prove the theorems regarding posterior concentration, it suffices to individually prove three conditions: entropy condition (C1), prior concentration condition (C2) and prior decay rate condition (C3). Below we describe the necessary steps required for proving each of these three conditions.

Entropy Condition (C1)

Since $\|\boldsymbol{z}\|_1 \leq Kp \|\boldsymbol{z}\|_\infty$ for any $\boldsymbol{z} \in \mathbb{R}^{Kp}$, by the bounds (3.17) and by definition of $\mathbb{F}_n$,

$$N\left(\frac{\varepsilon_n}{36}, \mathbb{F}_n \cap A_\varepsilon, H_n\right) \lesssim \sum_{K=1}^{k_n} N\left(\frac{\varepsilon_n}{36 C_g^n K p}, \{\beta : \|\beta\|_\infty \leq C_\beta^n\}, \|\cdot\|_\infty\right)$$

The RHS of the above can in turn be simplified to reduce the condition further:

$$N\left(\frac{\varepsilon_n}{36}, \mathbb{F}_n, H_n\right) \lesssim \sum_{K=1}^{k_n} \left(\frac{36 C_\beta^n C_g^n K q}{\varepsilon_n}\right)^{Kq}$$

Therefore the LHS of (C1) can be bounded from above by

$$(k_n + 1)p \left[\log 36 + \log(C_\beta^n C_g^n) + \log k_n + \log p - \log \varepsilon_n\right]$$

Since $C_\beta^n, C_g^n \lesssim n^M$ for some $M > 0$, ignoring smaller terms, proving condition (C1) reduces to proving

$$(k_n + 1)p \log n \lesssim n\varepsilon_n^2 \tag{3.19}$$

(Theorem 3.4.1): When f_0 is a step function with complexity K_{f_0} we can prove (3.19) by replacing $\varepsilon_n = n^{-1/2}\sqrt{K_{f_0} \log^{2\eta}(n/K_{f_0})}$ and $k_n \propto \frac{n\varepsilon_n^2}{p \log(n/K_{f_0})} = K_{f_0} \log^{2\theta-1}(n/K_{f_0})$ for some $\theta > 1/2$.

(Theorem 3.4.3): When f_{0l} is a ν-Hölder continuous function with $0 < \nu \leq 1$ for all $l = 1, \ldots, p$, replacing $\varepsilon_n = n^{-\nu/(2\nu+q)}\sqrt{\log n}$ and $k_n \propto \frac{n\varepsilon_n^2}{\log n} = n^{q/(2\nu+q)}$ proves (3.19).

Prior Concentration Condition (C2)

Let $\widetilde{f}_0 = \left(f_{\mathcal{T},\boldsymbol{B}_1^0}(\boldsymbol{x}), \ldots, f_{\mathcal{T},\boldsymbol{B}_p^0}(\boldsymbol{x})\right)$ denote the projection of f_0 onto a balanced k-d tree partition with a_n leaves, where a_n is chosen so that $\left\|f_0 - \widetilde{f}_0\right\|_{2,n} < \varepsilon_n/2$.

(Theorem 3.4.1:) If f_0 is a step function, $a_n = K_{f_0}$

(Theorem 3.4.3:) If f_0 is a ν-Hölder continuous function, a_n is chosen by the following lemma, which is analogous to Lemma 3.2 of Rockova et al. [2020].

Lemma 3.7.4. *Denote $f = \{f_l\}_{l=1}^p$ and assume $f_l \in \mathcal{H}^{\nu_l}$ where $\nu_l \leq 1$ for all $l = 1, \ldots, p$ and $\mathcal{X}$ is regular. Then there exists tree structured step functions $\hat{f} = \{f_{\mathcal{T},\boldsymbol{B}_l}\}_{l=1}^p \in \mathcal{F}_K$ for some given tree partition $\mathcal{T}$ with $K \in \mathbb{N}$ leaves such that for some constant $C > 0$,*

$$\left\|\hat{f} - f\right\|_{2,n} \leq Cd \sum_{l=1}^{p} \left(\frac{1}{K^{\nu_l/q}} \|f_l\|_{\mathcal{H}^{\nu_l}} \right) \leq C \frac{q}{K^{\nu/q}} \sum_{l=1}^{p} \left(\|f_l\|_{\mathcal{H}^{\nu_l}} \right),$$

where $\nu = \min_{l=1}^p \nu_l$.

As a corollary, replacing $C_0 = C\left(\sum_{l=1}^p \|f_l\|_{\mathcal{H}^\nu}\right)$, a_n satisfies

$$\left(\frac{2C_0 q}{\varepsilon_n}\right)^{q/\nu} \leq a_n \leq \left(\frac{2C_0 q}{\varepsilon_n}\right)^{q/\nu} + 1 \tag{3.20}$$

Using (3.17) and by triangle inequality, we can bound the LHS of (C2) from below by

$$C\pi(a_n)\Pi\left(\beta \in B_n^{a_n} : \left\|\beta - \beta^0\right\|_1 \leq \frac{\epsilon_n^2}{2C_g^n}\right)$$

For the prior by Chipman et al. [2010], $C = 1$ and $\pi(a_n) \gtrsim e^{-a_n \log a_n}$ (by Corollary 5.2 of Rockova and Saha [2019]).

For the prior by Denison et al. [1998], $C = \frac{1}{|F_{a_n}|} > (a_n d n)^{-a_n} > e^{-a_n \log a_n}$ (by Lemma 3.1 of Rockova et al. [2020]) and $\pi(a_n) \gtrsim e^{-a_n \log a_n}$ (by proof of Theorem 4.1 of Rockova et al. [2020]). Thus for both priors $C\pi(a_n) \gtrsim e^{-2a_n \log a_n}$.

For any $\mathbf{v} \in \mathbb{R}^M$, for some $M \in \mathbb{N} \setminus \{0\}$, we have $\|\mathbf{v}\|_1 \leq M \|\boldsymbol{\beta}\|_\infty$. Hence by using the definition of $B_n^{a_n}$ we can write

$$\Pi\left(\beta \in B_n^{a_n} : \left\|\beta - \beta^0\right\|_1 \leq \frac{\epsilon_n^2}{2C_g^n}\right) \geq \Pi\left(\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty \leq C_\beta^n, \quad \left\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\right\|_\infty \leq \frac{\epsilon_n^2}{2a_n q C_g^n}\right).$$

Since C_g^n and C_β^n both are increasing with n, for sufficiently large n, we will have $C_\beta^n \geq \frac{\epsilon_n^2}{2a_n q C_g^n}$, which would imply that the set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty \leq C_\beta^n, \quad \left\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\right\|_\infty \leq \frac{\epsilon_n^2}{2a_n q C_g^n}\}$ is exactly equal to the set $\{\boldsymbol{\beta} : \left\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\right\|_\infty \leq \frac{\epsilon_n^2}{2a_n q C_g^n}\}$ for all n large enough.

Thus the above expression gets bounded below by

$$\Pi\left(\boldsymbol{\beta} : \left\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\right\|_\infty \leq \frac{\varepsilon_n^2}{2a_npC_g^n}\right) \gtrsim e^{-C_1a_np\left(\|\beta_0\|_\infty+\frac{\varepsilon_n^2}{2a_npC_g^n}\right)_2^C}\left(\|\beta_0\|_\infty+\frac{\varepsilon_n^2}{2a_npC_g^n}\right)^{a_np}$$

Since $\varepsilon_n^2 \to 0$ and both a_n and C_g^n are both increasing with n, assuming $\|f_0\|_\infty \lesssim \sqrt{\log n}$, the above bound reduces to

$$e^{-C_1a_np\log^{C_2/2}n}\,\|\beta_0\|_\infty^{a_np/2} \gtrsim \log\left[-C_1a_np\log^{C_2/2}n\right]$$

We can prove $e^{-a_n\log n} \gtrsim e^{-n\varepsilon_n^2}$ for Theorem 3.4.1 and Theorem 3.4.3 separately by replacing appropriate values of ε_n. Since $C_2 \leq 2$, this would complete the proof.

Prior Decay Rate Condition (C3)

(Theorem 3.4.1:) When f_0 is a step-function with complexity K_{f_0},

$$\begin{aligned}\Pi(\mathcal{F}\setminus\mathbb{F}_n) &\leq \Pi(\mathcal{F}\setminus\bigcup_{K=1}^{k_n}F_K)+\Pi(\bigcup_{K\leq k_n}\{f\in F_K : \|\beta\|_\infty > C_\beta^n\})\\ &\leq \Pi(\bigcup_{K>k_n}F_K)+e^{-K_{f_0}\log n/2}\end{aligned}$$

When the true function f_0 connecting the input $\boldsymbol{X}$ with the output $\boldsymbol{Y}$ is a step-function with complexity K_{f_0}, we get $C_\beta^n \gtrsim K_{f_0}\log n$ and the second term becomes negligible. Thus the above condition reduces to proving

$$\Pi(\mathcal{F}\setminus\mathbb{F}_n) \lesssim \Pi(\bigcup_{K>k_n}F_K)+o(e^{-n\varepsilon_n^2})$$

When f_0 is a smooth function, similar simplification steps proceed as follows:

(Theorem 3.4.3:) For a ν-Hölder continuous function f_0, the LHS of condition (C3) can be bounded from above by

$$\Pi(\mathcal{F} \setminus \mathbb{F}_n) \leq \Pi(\mathcal{F} \setminus \bigcup_{K=1}^{k_n} F_K) + \Pi(\bigcup_{K \leq k_n} \{f \in F_K : \|\beta\|_\infty > C_\beta^n\})$$

By simplifying the second term, we get

$$\Pi(\mathcal{F} \setminus \mathbb{F}_n) \leq \Pi(\bigcup_{K > k_n} F_K) + \sum_{K=1}^{k_n} \Pi(\{\beta : \|\beta\|_\infty > C_\beta^n\})$$

Thus for proving condition (C3) it suffices to individually bound the two terms in the RHS of the above expression.

By condition (3.6), we can bound the term $\Pi(\{\beta : \|\beta\|_\infty > C_\beta^n\})$ and get

$$\begin{aligned}
\Pi(\mathcal{F} \setminus \mathbb{F}_n) &\lesssim \Pi(\bigcup_{K > k_n} F_K) + \sum_{K=1}^{k_n} e^{-C_\beta^n} \\
&\leq \Pi(\bigcup_{K > k_n} F_K) + k_n e^{-C_\beta^n} \\
&= \Pi(\bigcup_{K > k_n} F_K) + o(e^{-n\varepsilon_n^2})
\end{aligned}$$

The last line is due to the fact $C_\beta^n \gtrsim n$, when f_0 is a ν-Hölder continuous functions. Thus the second term in the RHS of the above expression can be ignored and it suffices to suitably bound the first term only.

Therefore it is enough to show that

$$\Pi(\bigcup_{K > k_n} F_K) \lesssim e^{-n\varepsilon_n^2}$$

This condition is satisfied for both priors under consideration. This follows from Section 8.3 of Rockova et al. [2020] for the prior by Denison et al. [1998] and from Corollary 5.2 of

Rockova and Saha [2019] for the prior by Chipman et al. [2010].

Classification with Dirichlet Step-Heights

The following theorem demonstrates that the conditions (3.5) and (3.6) are *sufficient* but not *necessary* conditions for guaranteeing that the posterior concentration rate for G-BART is near-minimax.

Theorem 3.7.1. *If we assume that the distribution of the step-sizes satisfies* (3.13)*, then under Assumptions 1 & 2 described in Section 4 of the manuscript, the Bayesian Tree estimator satisfies the following property,:*

(i) If f_0 *is* ν*-Hölder continuous with* $0 < \nu \leq 1$ *where* $\|f_0\|_\infty \lesssim \log^{1/2} n$*, then with* $\varepsilon_n = n^{-\alpha/(2\alpha+p)} \log^{1/2} n$*, and*

(ii) If f_0 *is step-function with complexity* $K_{f_0} \lesssim \sqrt{n}$*, then with* $\varepsilon_n = n^{-1/2}\sqrt{K_{f_0} p \log^{2\nu}\left(n/K_{f_0}p\right)}n$*,*

$$\Pi\left(f \in \mathcal{F} : H_n(\mathbb{P}_f, \mathbb{P}_{f_0}) > \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

in $\mathbb{P}_{f_0}^{(n)}$*-probability, as* $n, p \to \infty$*.*

The above statement is true for both tree priors considered in this paper: the prior by Denison et al. [1998] and a modified version of the prior by Chipman et al. [1998] with $p_{split}(\Omega_t) = \alpha^{d(\Omega_t)}$ *for some* $1/n \leq \alpha < 1/2$*.*

Proof. We need to prove three conditions: entropy condition (C1), prior concentration condition (C2) and prior decay rate condition (C3). Among these (C1) and (C3) can be proved by the same technique as in Section 3.7. Therefore we will only prove Condition (C2) here. We need to show, for some $c > 0$

$$\Pi\left(f \in \mathcal{F} : max\{K_n(f, f_0), V_n(f, f_0)\} \leq \varepsilon_n^2\right) \gtrsim e^{-cn\varepsilon_n^2} \tag{3.21}$$

Let $\widetilde{f}_0 = \left(f_{\mathcal{T},P_1^0}(\boldsymbol{x}), \ldots, f_{\mathcal{T},P_q^0}(\boldsymbol{x})\right)$ denote the projection of f_0 onto a balanced k-d tree partition $\mathcal{T}$ with a_n leaves, where a_n is chosen so that $\left\|f_0 - \widetilde{f}_0\right\|_{2,n} < \varepsilon_n/2$. If f_0 is a step function, $a_n = K_{f_0}$. If f_0 is a ν-Hölder continuous function, a_n is chosen by Lemma 3.2 of Rockova et al. [2020], where replacing $C_0 = C\left(\sum_{l=1}^p \|f_l\|_{\mathcal{H}^\nu}\right)$ we get

$$\left(\frac{2C_0 q}{\varepsilon_n}\right)^{q/\nu} \leq a_n \leq \left(\frac{2C_0 q}{\varepsilon_n}\right)^{q/\nu} + 1 \tag{3.22}$$

$f_{\mathcal{T},P_l^0}(\boldsymbol{x})$ is of the form (3.12) for some tree topology $\mathcal{T}$ with a_n leaves and $P_l^0 = \{P_{kl}^0\}_{k=1}^{a_n}$ for $l = 1, \ldots, p$. We assume there exists some $\delta_0 > 0$ such that $\min f_{0l} > \delta_0$ for all $l = 1, \ldots, q$. This implies $P_{lk}^0 > \delta_0$ for all $l = 1, \ldots q$ and all $k = 1, \ldots, K$. Therefore by (3.17), we can bound the LHS of (3.21) from above by

$$C\pi(a_n)\Pi\left(P \in [0,1]^{a_n p} : \left\|P - P^0\right\|_1 \leq \delta_0 \varepsilon_n^2/2\right)$$

For the prior by Chipman et al. [1998], $C = 1$ and for the prior by Denison et al. [1998], $C = \frac{1}{|\,F_{a_n}\,|} > (a_n d n)^{-a_n} > e^{-a_n \log a_n}$ (by Lemma 3.1 of Rockova et al. [2020]). By Corollary 5.2 of Rockova and Saha [2019] for the prior by Chipman et al. [1998] and by proof of Theorem 4.1 of Rockova et al. [2020] for the prior by Denison et al. [1998], we can show $\pi(a_n) \geq e^{-a_n \log a_n}$. Thus for both priors,

$$C\pi(a_n) > e^{-2a_n \log a_n} \tag{3.23}$$

Since $P_k \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_p)$ for all $k = 1, \ldots, K$ and $P_{lk}^0 > \delta_0$, for all $l = 1, \ldots, p$ and all $k = 1, \ldots, K$, we can bound $\Pi\left(P \in [0,1]^{a_n q} : \left\|P - P^0\right\|_1 \leq \delta_0 \epsilon_n^2/2\right)$ from above by

$$\Pi\left(P \in [0,1]^{a_n p} : \left\|P - P^0\right\|_\infty \leq \frac{\delta_0 \epsilon_n^2}{2a_n p}\right) \gtrsim C_\alpha (\frac{\delta_0 \epsilon_n^2}{a_n p})^{a_n p}, \tag{3.24}$$

where C_α is a constant that depends on the Dirichlet parameters $\alpha = (\alpha_1, \ldots, \alpha_q)$. Com-

bining (3.23) and (3.24) completes the proof. □

This concludes the discussion on the posterior concentration properties of Generalized Bayesian trees and G-BART models discussed in this chapter. The most important implication of these results is that they demonstrate that the theoretical optimality of BART models can be extended beyond the Gaussian regression setup discussed in Chapter 2. These theoretical strength in conjunction with extraordinary empirical success already demonstrated in a decade's worth of Bayesian machine learning literature, provide strong motivation for future research on discovering novel empirical applications for BART-like models.

CHAPTER 4

APPLICATIONS OF BART TO DISCRETE CHOICE DATA

4.1 Introduction

Discrete choice models are useful in situations where a rational agent makes a choice between several alternative options (e.g. when a voter tries to decide between several candidates in a democratic election) and the goal of the researcher is to analyze how different characteristics of the agent (voter) and/or the choice alternatives (candidates) impact the choice decision. These models have been extremely popular in marketing literature for modeling demand and brand loyalty [Keane, 1997, Kim et al., 2020, Lockshin et al., 2006, Sammer and Wüstenhagen, 2006] for consumer products. The seminal work of Domencich and McFadden [1975] on the urban transportation choice demonstrated the efficacy of employing choice data to estimate the consumer utility functions [Hausman and Wise, 1978]. The estimated model of consumer utility can be implemented to interpret counterfactual scenarios, such as the potential impact of a change in price on sales. The discrete choice models have been effective for analyzing not only consumer product market, but also transportation and tourism [Bates, 1987, Golob and Regan, 2002, Albaladejo and Díaz-Delfa, 2020, Truong and Hensher, 1985, Morey et al., 1991, Hackbarth and Madlener, 2013], labor market [Kornstad and Thoresen, 2007], political science [Poole and Rosenthal, 1985], criminology [Bernasco and Block, 2009], psychometry [Morikawa et al., 2002] and social behavioral analysis [Soetevent and Kooreman, 2007], to mention a few.

Conventional discrete choice models [Chintagunta and Nair, 2011, Keane et al., 2013, Hanemann, 1984] depend on strong parametric assumptions to predict consumer demand of particular products, depending on product characteristics and consumer attributes. These models are also an efficient tool for analyzing the influence of various factors in determining product sales, often dynamically across time [Keane et al., 2013] and across various demographics. The resulting inference often provides useful insights into customer segmentation

and targeted promotion of particular brands [Ruiz et al., 2020]. Counterfactual inference regarding inclusion of new products [Donnelly et al., 2019] is another important objective of large scale choice models. Beyond discrete choice, multiple-constraint models with non-exclusive preferences [Satomura et al., 2011] are another important component of choice model literature. Model for ordered alternatives [Small, 1987] have also been studied using latent variable formulations. Researchers have applied the discrete choice framework to a wide range of data sets, including aggregate, market-level data [Berry et al., 1995, Nevo, 2001, Petrin, 2002], as well as data from individual choices for a cross-section of individuals.

Recently a new avenue of research has emerged that employ latent factor models for predicting large-scale consumer demand in both single and multiple product categories [Ruiz et al., 2020, Görür et al., 2006, Athey et al., 2018]. These models, often dependent on infinitely many latent factors, have been extremely successful in demand forecast, as well as counterfactual inference such as examining price sensitivity [Donnelly et al., 2019, Wan et al., 2017]. Machine learning models such as deep neural networks have also been used for discrete choice modeling [Munro, 2018, Sifringer et al., 2018, Kanodia and Ganeriwal]. However these models lack interpretability despite good predictive performance.

In this chapter we describe a semiparametric model of consumer choice using Bayesian Additive Regression Trees (BART) [Chipman et al., 2016]. We consider situations, where given several alternatives, each customer chooses exactly one product. Our objective is to predict preference of the customer using information on the customer's characteristics (e.g. age, income etc) and attributes of the different product varieties (e.g. price, color, size etc). We implement a multinomial BART model using existing software (the `BART` R Package) that partitions the entire consumer space based on various product features and customer attributes and then models individual preferences locally, within each partition, using a multinomial distribution.

BART provides several advantages for large-scale consumer choice modeling. First BART is an efficient tool for estimating nonlinear utility functions. Most conventional discrete

choice models discussed above (except the deep learning models) assume a linear relationship between the choice response and the customer and product characteristics and thus often miss interesting nonlinear patterns and interactions. In contrast BART employs step functions to accommodate nonlinearity and interactions among covariates. This is particularly suitable for nonhomothetic choice situations with asymmetric switching [Rossi et al., 2012] between different products. BART incorporates different degrees of interactions between customer and/or product attributes, thus facilitating heterogeneity in price-sensitivity and income-dependence across varieties. The entire consumer space is divided into smaller homogeneous segments based on customer attributes and their relative sensitivity to different product features, that might facilitate targeted marketing. Moreover BART relaxes some of the strong assumptions implicit in using multinomial choice models, such as "independence of irrelevant alternatives" [McFadden, 1974] which states that if a certain product variety is removed from the market, the customers preference for the other products remain unaffected. This assumption is often unrealistic and we will see in Section 4.2 that BART can alleviate the limitation by incorporating competing product attributes in the partitioning scheme.

This chapter is organized as follows. Section 4.2 describes the multinomial BART model for discrete choice data. Section 4.3 compares BART with some popular alternatives: multinomial probit models with varying coefficients and/or mixed effects and a multinomial log-linear model fitted with neural networks, on ten benchmark consumer choice datasets. In Section 4.4 we discuss how the BART model can be adapted to account for monotonic dependence of customer preference on certain covariates. Finally Section 4.5 concludes with a discussion on the efficacy of BART for consumer choice modelling along with possible directions of future research.

4.2 Adapting BART for Discrete Choice Modeling

The data setup under consideration consists of a categorical response $\boldsymbol{Y}$ that records which alternative is chosen by the consumer and several covariates $\boldsymbol{X}$ recording customer and/or

product attributes. Specifically, if a customer chooses to buy the i-th product out of 'q' available alternatives $\{1, \ldots, q\}$, the response $\boldsymbol{Y}$ can be formulated as a q-vector which has 1 at the i-th coordinate and 0 elsewhere. Since the customer is assumed to choose exactly one option (i.e. the alternatives are mutually exclusive), the response vector $\boldsymbol{Y}$ has exactly one non-zero entry. The objective is to predict customer choice $\boldsymbol{Y}$ based on d_1 product attributes (such as price, display activity etc) $\boldsymbol{X}_1 \in \mathbb{R}^{d_1}$ and d_2 customer attributes (such as income, age etc) $\boldsymbol{X}_2 \in \mathbb{R}^{d_2}$. Denote the set of all covariates by $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \in \mathbb{R}^d$ where $d = d_1 + d_2$ is the total number of covariates.

Each such covariate is recorded for every choice instance. For example, if the choice situation involves a customer choosing one out of two brands of bread, every instance of the data might record which brand the customer chose ($\boldsymbol{Y}$), prices of the two brands at the time of shopping ($\boldsymbol{X}_1$) and the customer's yearly income ($\boldsymbol{X}_2$). The covariate values may change over each shopping instance, even for the same customer, if the prices of one or both brands change and/or the income of the customer changes.

We can assign a Multinomial distribution to the response variable, while treating the individual parameters as functions of the covariates. Subsequently we can approximate these unknown functions with Generalized BART models (Section 3.2), equipped with suitable link functions. Specifically, we assume that

$$\boldsymbol{Y} \sim Multinomial\left(1; \gamma(\boldsymbol{X})\right), \tag{4.1}$$

where the parameter $\gamma(\boldsymbol{X}) = (\gamma_1(\boldsymbol{X}), \ldots, \gamma_q(\boldsymbol{X}))$ is a q-dimensional function of the covariates $\boldsymbol{X}$, such that $\gamma(\boldsymbol{X})'\mathbf{1}_q = 1$ for any given $\boldsymbol{X}$. Here $\mathbf{1}_q$ denotes a q-dimensional vector with all entries equal to 1. Assume $\gamma_j(\boldsymbol{X}) = \Phi_j\left(f_1(\boldsymbol{X}), \ldots, f_q(\boldsymbol{X})\right)$ for some unconstrained functions $\{f_j\}_{j=1}^q$, where the transformation functions $\Phi_j(z_1, \ldots, z_q) = \frac{e^{z_j}}{\sum_{l=1}^q e^{z_l}}$ is assumed to be the softmax link function. Our goal is to estimate these unknown functions $\{f_j\}_{j=1}^q$.

We can impose independent BART priors on each of f_j, for $j = 1, \ldots, q$, and estimate

them using step-functions supported on binary tree/forest partitions. Consider an approximating tree ensemble consisting of T trees where the t-th tree has K_t leaves that generates a partition $\{\Omega_k^t\}_{k=1}^{K_t}$ of the covariate space. Then each f_j can be approximated by a step-function of the form $\widehat{f}_j = \sum_{k=1}^{K} \beta_{jk} \mathbb{I}\{X \in \Omega_k\}$ which in turn reduces the estimate of each parameter γ_j to be of the form:

$$\gamma_j = \frac{\sum_{t=1}^{T} \sum_{k=1}^{K_t} \beta_{jk}^t \mathbb{I}\{\boldsymbol{X} \in \Omega_k^t\}}{\sum_{l=1}^{q} \sum_{t=1}^{T} \sum_{k=1}^{K_t} \beta_{lk}^t \mathbb{I}\{\boldsymbol{X} \in \Omega_k^t\}}, \tag{4.2}$$

The multinomial logit model described above is equivalent to the solution for a linear utility function with extreme value distribution of the form $p(u) = e^{-u}e^{-e^{-u}}$ [McFadden, 2001]. The solution is always to choose the product with maximum marginal utility. If instead the error distribution is assumed to be Gaussian, this gives rise to the multinomial probit model.

Note that every f_j depends on *all* the available covariates. For example suppose a customer tries to choose between three different brands of yogurt and the only covariate information available is the prices of the three brands. Let p_i denote the price of brand i. So the covariate vector is $\boldsymbol{X} = (p_1, p_2, p_3) \in \mathbb{R}^3$ and each of the three functions $\{f_1, f_2, f_3\}$ would depend on *all* three prices. This would mean that the estimating partition for every tree $\{\Omega_k^t\}_{k=1}^{K_t}$ can potentially depend on the prices of all three brands. Thus the probability that the customer chooses brand j would depend on the price of the j-th brand, as well as the prices of all the competing brands. Therefore if a particular brand is removed from the market, the relative preferences of the remaining brands change as well. Thus the assumption of "independence of irrelevant alternatives" is no longer valid. The model (4.1)-(4.2) is in fact same as the BART model for multiclass classification and hence can be implemented using existing software (the `BART` R package of McCulloch [2017]).

In case of a quantitative choice situation, i.e. when information is available on the quantities purchased, a hierarchical model can be specified where the first level involves a qualitative model, such as the one specified above, that selects the preferred alternative

and the second level models the logarithm of the purchased quantity as a Gaussian random variable, where the parameters might depend on the selected variety.

4.3 Empirical Results

In this section we compare the empirical predictive performance of the BART model (4.1) with popular alternatives: multinomial logistic regression and its variants [Train and Croissant, 2012], over ten benchmark data sets (Table 4.1), each of which considers some discrete consumer choice situation where a consumer chooses exactly one product out of several alternatives. We consider two competing models: (i) popular variants of the multinomial logit model commonly used for marketing data analysis [Train and Croissant, 2012]: equipped with varying coefficients where the slope and intercepts are allowed to vary across different product categories (choice alternatives) and mixed effects with random intercept for every customer (referred subsequently as "MLogit") and (ii) the traditional multinomial logistic regression fitted with a neural network (referred subsequently as "NNet").

The Multinomial logit model with varying coefficient and/or mixed effects are implemented through the R package `mlogit` [Croissant and Croissant, 2020], following the instructions provided by Train and Croissant [2012]. The multinomial logistic regression is implemented through the function "multinom" in the widely used R package `nnet` [Ripley et al., 2016]. This function fits a multinomial log-linear model via neural networks. In contrast to the "MLogit" model that includes a customer specific random intercept, the "Multinom" function considers each individual observation as independent, even if repeated choice instances are reported for the same consumer. Finally the multinomial BART model (4.1) is implemented through the R package `BART`.

Now let us describe the ten discrete choice datasets used for the empirical comparison in Table 4.1. Among these ten datasets, the data on "Electricity" is obtained from the R package `mlogit` [Croissant and Croissant, 2020], the data on "OrangeJuice" is obtained from the R package `ISLR` [James et al., 2017], while all the other eight datasets are from the R

package `Ecdat` [Croissant and Graves, 2020]. The package `Ecdat` contains a wide range of Econometrics data sets, which are also available in the Journal of Business Economics and Statistics web site `https://amstat.tandfonline.com/loi/ubes20`.

For every dataset considered here, consumers choose exactly one option out of several alternatives (e.g. one out of four brands of ketchup or one out of two travel mode: train and car). For every choice instance the decision of the customer is recorded along with some characteristics of the customer (e.g. age) and also some attributes of the choice alternatives (e.g. price). Customer preferences are transformed into binary vectors. For example if the customer chooses the 'i'-th option out of 'p' alternatives, the response is encoded as a 'p'-vector which has the 'i'-th entry equal to one and the other entries equal to zero. Below we give a brief description of the corresponding choice situation, along with the number of recorded choice instances (the sample size), number of choice alternatives and available information on product and/or customer attributes (the covariates).

Car: This is a dataset on 4654 customers who choose to buy a vehicle among 6 possible alternatives. Every choice instance records information on different choice attributes such as type of vehicle, type of fuel required, fuel range, acceleration, highest attainable speed, emission, size and space capacity of the vehicle, cost/mile of travel and the fraction of stations that can refuel/recharge the vehicle. Every observation also contains information on individual customer such as whether they received college education, size of the household and their average amount of commute per day. Information on choice attributes are recorded for every choice instance and they might not be same for every customer, meaning that two different customers facing the exact same choice decision between the same five cars, might have different covariate vectors if the 'cost/mile of travel' of a particular car changes.

Cracker: This is a panel data containing 3292 choice instances for some customers who choose among one of the four brands of cracker: sunshine, kleebler, nabisco and private. Information is available on three attributes for each of the brands: their price, a binary

indicator recording whether there was a display for that brand and another binary indicator which reports whether there was a newspaper feature advertisement for that brand. Individual customer identifiers are also available. These customer identifiers can be coded as a factor and supplied as a covariate while training the BART model. However including customer identifiers in the partitioning scheme would mean that we will not be able to predict the choice decision of a customer who is not in the training data. Therefore we choose to ignore these customer identifiers while implementing the multinomial BART model (4.1).

Electricity: This is a dataframe describing the choice of Electricity provider among 4 possible alternatives (marked $1, 2, 3, 4$) for 2308 households. Information is recorded on the individual household index, as well as several attributes of the supplier such as fixed price at a stated cents per kWh, with the price varying over suppliers and experiments, for scenario i (= 1, 2, 3, 4), the length of contract that the supplier offered, in years (such as 1 year or 5 years.) During this contract period, the supplier guaranteed the prices and the buyer would have to pay a penalty if the customer switched to another supplier. The supplier could offer no contract in which case either side could stop the agreement at any time, and the contract length is recorded as 0. There are two binary indicators that denote whether the supplier is a local company and if the supplier is well-known. Two other covariates are available: a time-of-day rate under which the price is 11 cents per kWh from 8 am to 8 pm and 5 cents per kWh from 8pm to 8am. These TOD prices did not vary over suppliers or experiments. And finally a seasonal rate under which the price is 10 cents per kWh in the summer, 8 cents per kWh in the winter, and 6 cents per kWh in the spring and fall.

Fishing: This is a data on the fishing mode choice for 1182 consumers each of whom choose a preferred mode of recreation from 4 available options: beach, pier, private boat and charter boat. Information on the price and catch rate for every mode option is available, along with the monthly income of the consumer.

Heating: This dataset contains information on the choice of heating mechanism for 900 households, each of which chooses a preferred heating system out of 5 possible alternatives: gas central, gas room, electric central, electric room and heat pump. The installation cost and annual operating cost for each alternative is recorded. Customer attributes such as the annual income of the household, age of the household head and numbers of rooms in the house are also available.

Ketchup: This is a panel data on 4956 choice instances for several customers, each of whom choose to buy a preferred brand of ketchup from 4 possible options: heinz, hunts, del monte and store brand. Price of every brand for each shopping trip is available, along with identifying information for every customer.

ModeChoice: This is a data on preferred travel mode choice for 840 consumers, each of whom choose to travel by one of the 2 alternatives: car vs train. Travel cost and travel time are recorded for the preferred mode for each trip, along with household income and party size travelling through the chosen mode.

OrangeJuice: The data set contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid orange juice. A number of characterstics of the customer and product are recorded (18 variables), such as price, discount, list price and sale price corresponding to both brands, a binary indicator denoting whether there is a special promotion on a particular brand at that time, Customer attributes such as brand loyalty, the week of purchase and identifying information for the store where the sale took place.

Tuna: This is a choice data on 13705 shopping instances where customers choose to buy one of the five possible brands of canned tuna: Starkist (in water), Chicken of the sea (in water), a store–specific private label (in water), Starkist (in oil) and Chicken of the sea (in oil). Price of every brand and the customer identifier for each shopping trip is recorded.

Data	Sample	Choices	MLogit	NNet	BART
Car	4654	6	0.6266	0.6477	0.6867
Cracker	3292	4	0.4551	0.6652	0.8891
Electricity	2308	4	0.7307	0.7397	0.7500
Fishing	1182	4	0.5128	0.9120	0.9787
Heating	900	5	0.4662	0.6387	0.7362
Ketchup	4956	4	0.4553	0.7235	0.8100
ModeChoice	840	2	0.7045	0.7285	0.9810
OrangeJuice	1070	2	0.6684	0.8742	0.8868
Tuna	13705	5	0.7612	0.7855	0.8926
Yogurt	2412	4	0.3903	0.6774	0.9091

Table 4.1: Area Under the ROC Curve (AUC) [Hand and Till, 2001] computed by averaging over AUC obtained from a 4-fold cross validation. Comparison is conducted over three methods: (i) a varying coefficient and/or random intercept multinomial logistic model (referred to as MLogit), (ii) the multinomial logistic regression fitted through a neural network (referred to as NNet) and (iii) a multinomial logistic BART model (referred to as BART).

Yogurt: This data records the choice of yogurt brand over 2412 shopping trips by multiple consumers. There are 4 available yogurt brands the customers can choose from: yoplait, dannon, hiland and weight watcher. Two product characteristics are recorded for every brand: their price and whether there is a newspaper feature advertisement for that brand at the time of purchase. Individual customer identifiers are also recorded.

Table 4.1 records a predictive performance comparison between the multinomial BART model (4.1) and two popular alternatives: the multinomial logit model (referred to as 'MLogit') with varying coefficient and customer specific random effects, where applicable [Croissant and Croissant, 2020] and a multinomial log-linear model fitted with a neural network (referred to as 'NNet'). We consider the ten benchmark datasets described above. Each data is split into four cross-validation segments. For every segment, the models are trained on the three remaining segments and individual choice probabilities are predicted on the left-out segment based on the trained model. Then the area under the ROC curve (AUC) [Hand and Till, 2001] is computed for each of the four segments and the average AUC is recorded in Table 4.1. From the table, we see that the BART model demonstrates

significant predictive superiority over the multinomial logistic regression with varying coefficients and/or mixed effects (MLogit) for all the ten data sets under consideration. BART also outperforms the multinomial log-linear model fitted with neural networks (NNet), even though the performance gain is less dramatic compared to the "MLogit" model. Now let us dig deeper into the specific examples considered in Table 4.1 in order to discover possible causes for such huge discrepancy between BART and linear parametric models like MLogit, in terms of out-of-sample prediction.

Interestingly, the average AUC values for BART and the multinomial logit model are almost equal for choice situations that are less frequent in a customer's lifetime and /or are "more pricey" investments, such as choosing an electricity provider or buying a car. The cost of switching between different alternatives is higher but the alternatives themselves do not differ much in price, e.g. the cost of different cars in the data do not differ much but the switching cost is high, meaning that if after buying a car the customer does not like it and wants to exchange it for a different car, the entire procedure is expensive. Also, the product options vary greatly in terms of specific attributes. For example different cars might differ a lot in emission and attainable speed. While some customers might put more importance on the highest attainable speed, others might prefer an alternative with low emission. In these kinds of situations, customers tend to choose the option that meets their specific needs and a change in price of other alternatives might not influence the customer's decision at all, which means that the interactions between the features of one alternative with another might not have much influence on the customer's decision.

In contrast, the highest gain in predictive accuracy for BART, compared to the multinomial logit (MLogit) model is observed for products that are bought more often and are "less pricey", such as yogurt, cracker, ketchup etc. For these frequently bought products, how much of one brand the customer prefers might be heavily influenced by the price of the competing brands or how much advertisement was observed for the competing brand. So there tends to be much more interaction between features of competing alternatives. It is much

harder to account for different degrees of interactions among the covariates while using parametric linear models such as multinomial logit. In comparison, BART easily accommodates different degrees of interactions and exhibits superior predictive performance.

4.4 Adapting to Monotonicity:

Consumer theory states that for most product categories, a rational consumer chooses to buy less of a product if its price increases [Deaton and Muellbauer, 1980]. Sales might also decrease when income decreases and the customer chooses to spend less. There are many such product or customer attributes that have a monotonic effect on sales volume. Such monotonic constraints might be incorporated into discrete choice models discussed above by suitably adapting a Monotone version of the BART model, proposed by Chipman et al. [2016]. This model assumes a similar prior as the classical BART model [Chipman et al., 2010] on tree partitions $\Pi(\mathcal{T})$. Unlike BART, the monotone BART model imposes a monotonic constraint on the step heights $\Pi(\beta)$. A brief description of the monotone BART prior [Chipman et al., 2016] is given in the appendix of this chapter.

For discrete choice situations, the monotone BART model needs to be implemented with caution. For example, in the 'Orange Juice' data described in the "Empirical Results" section below, the average Area under the ROC Curve (AUC) over a four-fold cross validation equals 0.8868 for BART and 0.886 for Monotone BART. However the Monotone BART prior can exhibit poor performance if the sign of the covariates are not adjusted properly to have an increasing relationship with $P(\boldsymbol{Y} = 1)$. For example, if we use the covariates in the orange juice data without any alterations, the average AUC drops to 0.7612. In another example, in the 'Tuna' data described in the "Empirical Results" section, customers choose one out of five brands of canned tuna. If we assume the probability of buying a particular brand has a monotonic decreasing relationship with its price and a monotonic increasing relationship with the price of its alternatives, which is in line with consumer theory, the Monotone BART model gives an average AUC = 0.7539, over a four-fold cross validation, whereas the

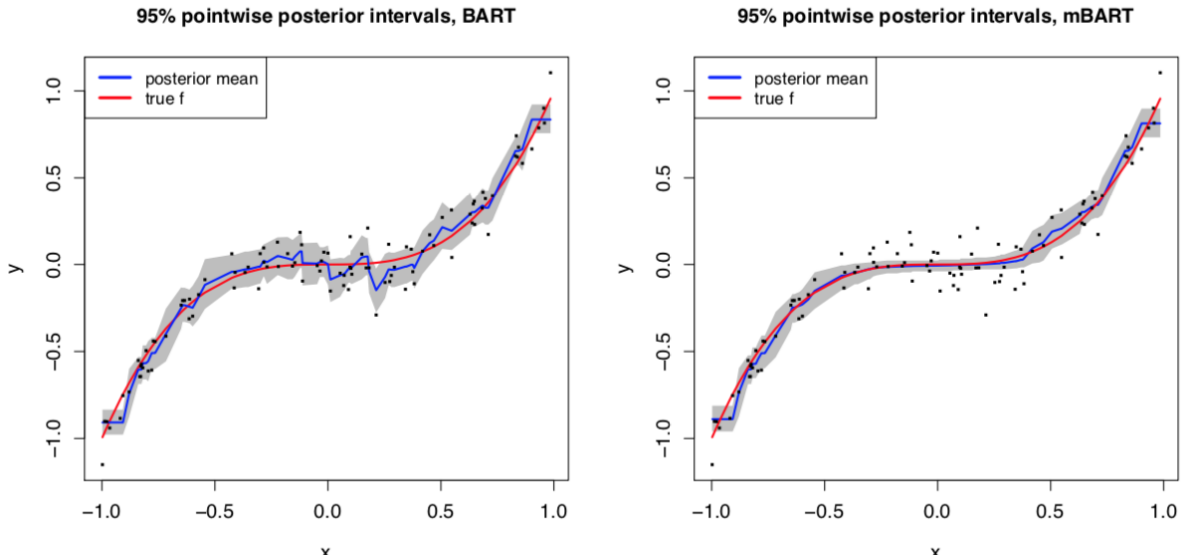


Figure 4.1: Comparing BART and Monotone BART estimates for a monotone one-dimensional function $f(z) = z^3$. Even though the Monotone BART fits are better throughout, BART manages to recover the overall monotonicity of the function, without imposing any additional assumption [Chipman et al., 2016].

average AUC for the BART model is 0.8926. This indicates that for real data, anticipating the monotonic nature of relationship between sales and prices is often not simple. In such situations applying Monotone BART can significantly diminish predictive accuracy.

In fact a similar phenomenon was pointed out by [Chipman et al., 2016], who demonstrated that even though for a truely monotonic function (Figure 4.1), the Monotone BART model performs slightly better than regular BART, in terms of squared error, BART still manages to recover the underlying monotonicity fairly well without imposing any additional constraints. However if the true function is not monotonic (Figure 4.2), the Monotone BART model performs poorly and the estimated function grossly deviates from the truth.

Consumer theory states that *keeping other covariates constant*, sales volume of a particular product decreases as its price increases and/or the price of a *close* substitute drops down. Therefore it might seem reasonable to impose a suitable monotonicity assumption on the relationship between sales and price. However quite often for a real data, the re-

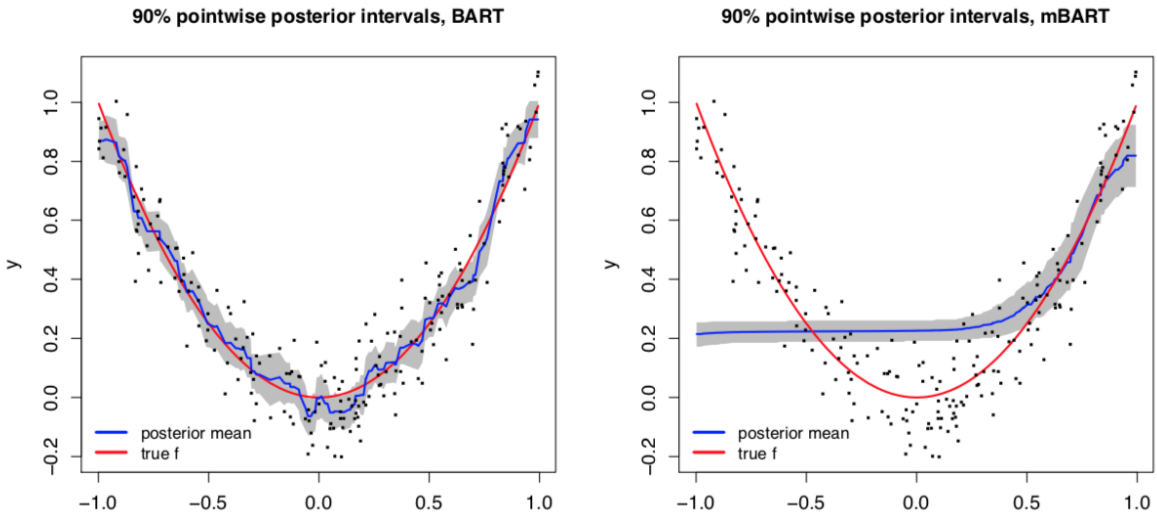


Figure 4.2: Comparing BART and Monotone BART estimates for a non-monotone one-dimensional function $f(z) = z^2$. The monotone BART estimator suffers poorer performance due to the lack of monotonicity of the true function [Chipman et al., 2016].

lationship between price and sales is not monotone, unless accounted for relative product quality. For example, if price of an inferior substitute drops, the sales of other available options might remain the same. Since the information on relative product quality is often unavailable and accounting for all possible confounders that might interact with price is almost always impossible, it is safer to use BART instead of Monotone BART, because implementing Monotone BART requires calibrating the signs of each covariate so that they have a monotone increasing relationship with the response. Guessing the nature of such relationships is often cumbersome, to say the least. BART on the other hand requires negligible tuning while providing comparable predictive accuracy, as observed from both empirical applications [Chipman et al., 2016] and theoretical findings (Theorem 3.4.2).

4.5 Further Comments

In this section we have demonstrated how Bayesian tree ensembles can be suitably adapted to model discrete choice situations. BART provides a flexible tool that can efficiently model nonlinear dependence of choice probabilities on customer and/or product attributes, simul-

taneously accounting for possible interactions among different covariates. BART can potentially eliminate many stringent assumptions such as "independence of irrelevant alternatives" and "constant marginal utility", that often torment conventional linear models for discrete choice data. Moreover, BART does not require any variable transformation unlike most parametric linear models for discrete choice data. In Section 4.3 we see that the BART model exhibits impressive predictive performance over several benchmark datasets, compared to popular existing alternatives. However several research questions remain unexplored.

We have observed the empirical performance of the BART model only for situations where the rational agent (e.g. consumer) chooses exactly one out of several alternatives (Model 4.1). However in many real life situations customers often choose multiple alternatives at the same time. For example, for aggregated goods like "food" and "shelter", consumers might choose all the alternatives by some amount. For differentiated product offerings such as 'yogurt' or 'soft drinks' the customer might choose more than one flavor at the same shopping instance. These kinds of choice situations might be modelled using a "multivariate probit" [Rossi et al., 2012] adaptation of BART. Quantity information, if present can also be utilized by using a hierarchical extension of Bayesian trees where individual quantities are assumed to come from a log Normal (for continuous quantity information such as in produce) / Poisson distribution (for discrete quantity information such as in packs of soft drinks), conditioned on the alternatives chosen.

Another interesting direction would be to explore the scope of Bayesian trees and forests for panel data, where customers are tracked over multiple shopping instances over time. A simple extension would be to include random customer-specific intercepts to the step-heights, thus enabling information sharing among correlated observations. This would be a multinomial extension of the random intercept BART model proposed by Tan et al. [2016]. However this method ignores the time variability in consumer behavior and a dynamic extension of BART connecting past choice behavior with current preference, might be worth exploring. A common feature in time series data on consumer choice is that certain brands

might become unavailable and new brands might get introduced over time. In this context, a dynamic spike-and-slab (Chapter 5) extension of the individual choice probabilities might be interesting, where a point-mass spike distribution accounts for the situation where the corresponding product variety is unavailable.

An important avenue for future research would be to design Bayesian tree ensembles that are more interpretable and/or suitably adapt existing BART models that facilitate efficient counterfactual inference for discrete choice data, thus making flexible machine learning models like BART more amenable to marketing research applications. A very important objective of discrete choice models is to quantify the influence of individual covariates on the volume of sales of a certain product, for linear models which amounts to measuring the individual gradients of the linear function with respect to the covariate of interest. However BART being a sum of step functions, measuring contributions of individual covariates is almost unattainable, even though a relative ranking of "covariate-importance" can be obtained [Sparapani et al., 2019]. An interesting development in this direction is "BART with targeted smoothing", proposed by Starling et al. [2020], that can be used to model functional response variables that are smooth over a covariate, such as time. A potential approach would be to design a multivariate analog of this model, without constraining the estimator space too much and hence avoiding significant loss in predictive accuracy. Another important extension would be to build a hierarchical BART model that can incorporate multiple product categories at the same time, such as in grocery store data [Donnelly et al., 2019]. Such models often benefit from accounting for unobserved inter and intra-consumer heterogeneity [Krueger et al., 2020]. Moran [2019] propose a factor analysis framework with BART that can be potentially adapted for incorporating latent consumer and product characteristics. Evaluating computational feasibility and empirical performance of these adaptations would provide a promising avenue for future research.

4.6 Appendix

The Monotone BART Prior [Chipman et al., 2016]

The tree generation process for the monotone BART model [Chipman et al., 2016] is same as in the unconstrained BART model [Chipman et al., 2010] described in Section 2.2. The two models primarily differ in terms of the prior on the step heights $\pi(\boldsymbol{\beta} \mid \mathcal{T})$.

Let $\boldsymbol{Y}$ denote the response and $\boldsymbol{X} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ denote the covariates, among which a subset of the covariates $\mathcal{S} = \{x_{i_1}, \ldots, x_{i_M}\}$ has a monotonic relationship with the response and the dependence of the response on the other covariates are unrestricted. The monotone BART model puts positive prior on only those trees that are monotone in coordinates $\mathcal{S}$. Such a monotone tree is defined as follows.

For a tree each terminal node region will be a rectangular region of the form

$$\Omega_k = \{\boldsymbol{X} : x_i \in [L_{ik}, U_{ik}), i = 1, \ldots, d\},$$

where the interval $[L_{ik}, U_{ik})$ for each x_i is determined by the sequence of splitting rules leading to Ω_k. Two nodes Ω_k and $\Omega_{k'}$ are called separated if $U_{ik} < L_{ik'}$ or $U_{ik'} < L_{ik}$ for some i. If Ω_k and $\Omega_{k'}$ are not separated, Ω_k is called an above-neighbor of $\Omega_{k'}$ if $L_{ik} = U_{ik'}$ for some i and it is called a below-neighbor of $\Omega_{k'}$ if $L_{ik'} = U_{ik}$ for some i. A tree function will be monotone in coordinate x_i if the step-height of each of its terminal node is

(i) not greater than the minimum level of all of its above-neighbor regions in the x_i direction, and

(ii) not less than the maximum level of all of its below-neighbor regions in the x_i direction.

Let C be the set of all trees $\mathcal{T}$ which satisfy these monotonicity constraints on each $x_i \in \mathcal{S}$. The monotone BART prior puts all prior mass on the trees in this set C.

Next, given a monotone tree partition $\mathcal{T}$ with K steps, the distribution of the step heights are assumed to be normal, just as in the unconstrained BART [Chipman et al.,

2010] model but with different prior variance choices depending on whether or not the step-height is constrained by the set C. If the k-th step height β_k is unconstrained, assume $\beta_k \sim \mathcal{N}(0, \sigma_\beta^2)$. On the other hand if β_k is constrained by C, assume $\mathcal{N}(0, c^2\sigma_\beta^2)$ with the choice $c = \frac{\pi}{\pi - 1} \approx 1.4669$. For a detailed justification of this choice, please refer to Section 3.3 of Chipman et al. [2016].

Software

All empirical results presented in this chapter have been implemented using the opensource software `R`. All the relevant code are available in the following GitHub repository:

`https://github.com/Enakshi-Saha/Discrete-Choice-Model`

CHAPTER 5

DYNAMIC SPARSE FACTOR ANALYSIS

5.1 Introduction

The premise of dynamic factor analysis (DFA) is fairly straightforward: there are unobservable commonalities in the variation of observable time series, which can be exploited for interpretation, forecasting, and decision making. Dating back to, at least, Burns and Mitchell [1947], the fundamental idea that a small number of indices drive co-movements of many time series has found plentiful empirical support across a wide range of applications including economics [Stock and Watson, 2002, Bai and Ng, 2002, Bernanke et al., 2005, Baumeister et al., Cheng et al., 2016], finance [Diebold and Nerlove, 1989, Aguilar et al., 1998, Pitt and Shephard, 1999, Aguilar and West, 2000, Carvalho et al., 2011], and ecology [Zuur et al., 2003], to name just a few. More notably, in their seminal work on DFA, Sargent et al. [1977] showed that two dynamic factors could explain a large fraction of the variance of U.S. quarterly macroeconomic variables. Motivated by a similar (but significantly larger) application, we develop scalable Bayesian DFA methodology to glean insights into the hidden drivers of the U.S. macroeconomy before, during and after the Great Recession.

With large-scale cross sectional data becoming readily available, the need for developing scalable and reliable tools adept at capturing complex latent dynamics have spurred in both statistics and econometrics [Kaufmann and Beyeler, 2018, Kaufmann and Schumacher, 2017, Frühwirth-Schnatter and Lopes, 2018, Nakajima et al., 2017]. A wide variety of factor-type models exist with varying degrees of modeling flexibility. One popular class is factor stochastic volatility models [Pitt and Shephard, 1999, Aguilar and West, 2000, Kastner et al., 2017] which, in their simplest form, assume (a) constant loadings, (b) independent factors, and (c) time-varying structures on residual variances and factor variances. Extensions to time-varying loadings that allow for more flexible correlation modeling have been considered in, e.g., Aguilar et al. [1998], Aguilar and West [1998, 2000], Lopes and Carvalho [2007] and later

extended by Nakajima and West [2013,], Nakajima et al. [2017] to sparse factor models via latent thresholding. Sparsity has also been a key component in dynamic covariance estimation models, such as in Kastner [2019], who proposes a factor stochastic volatility model in combination with a global–local shrinkage prior. This prior is a generalization of the Bayesian Lasso [Park and Casella, 2008] and has also been adopted in the context of Bayesian vector autoregressive (VAR) models [Huber and Feldkircher, 2019, Kastner and Huber, 2020] that are capable of handling vast-dimensional time series. Other developments include large-scale Bayesian VAR methods [Bańbura et al., 2010, Koop and Korobilis, 2013, Korobilis, 2013, Giannone et al., 2014, 2015, Kuschnig and Vashold, 2019]. More recently, Koop et al. [2019] and Aunsri and Taveeapiradeecharoen [2020] extended random compression dynamic regression methods [Guhaniyogi and Dunson, 2015] to the VAR framework giving rise to the Bayesian Compressed VAR (BCVAR) model that exhibits an impressive forecasting performance in high dimensions. VAR models have also been integrated within dynamic factor structures in factor augmented vector autoregressive (FAVAR) models [Bernanke et al., 2005, Stock and Watson, 2005] and in their recently introduced sparse extension [Kaufmann and Beyeler, 2018]. These FAVAR models have been particularly effective in high-dimensional macroeconomic applications [Wagan et al., 2019, Evgenidis et al., 2019, Potjagailo, 2017, Daniele and Schnaitmann, 2019]. A detailed discussion on FAVAR models in macroeconomics can be found in Stock and Watson [2016].

Sparsity is an indispensable tool in high dimensional inference situations where the number of parameters exceed the number of observations by a large extent. The fundamental goal of our research is to build a dynamic factor analysis method that discovers a dynamic sparse factor structure with an unknown and possibly time-varying number of latent factors and with factor loading matrices that evolve somewhat smoothly over time. There are three important ingredients of dynamic sparsity that reside at the core of our methodology.

Firstly, the latent factor loadings should account for time-varying patterns of sparsity. In (macro-)economics and finance, the sequentially observed variables may go through multiple

periods of shocks, expansions, and contractions [Hamilton, 1989]. It is thus expected that the underlying latent structure changes over time– either gradually or suddenly– where some factors might be active at all times, while others only at certain times. For example, in our empirical analysis we find that certain factors exert influence on some series only during a crisis and later permeate through different components of the economy as the shock spreads.

Dynamic sparsity plays a very compelling role in capturing and characterizing such dynamics. Recent developments in sparse factor analysis reflect this direction of interest [West, 2003, Carvalho et al., 2008, Yoshida and West, 2010, Lopes et al., 2010]. More recently, Nakajima and West [2013], Nakajima et al. [2017] deployed the latent threshold approach of Nakajima and West [2013] in order to induce zero loadings dynamically over time. Our methodological contribution builds on this development, but poses less practical limitations on the dimensionality of the data.

Related to the previous point is the question of selecting the number of factors. This modeling choice is traditionally determined by a combination of *a-priori* knowledge, a visual inspection of the scree plot [Onatski, 2009], and/or information criteria [Bai and Ng, 2002, Hallin and Liska, 2007]. In the presence of model uncertainty, the Bayesian approach affords the opportunity to assign a probabilistic blanket over various models. Bayesian nonparametric approaches have been considered for estimating the factor dimensionality using sparsity inducing priors [Bhattacharya and Dunson, 2011, Rockova and George, 2016]. The added difficulty stemming from time series data, however, is that the number of factors *may change over time* [Bai and Ng, 2002]. As a remedy, we turn to dynamic sparsity as a compass for determining the number of factors without necessarily committing to one fixed number ahead of time.

The third essential requirement is accounting for structural instabilities over time with time-varying loadings and/or factors. One seemingly simple solution has been to deploy rolling/extending window approaches to obtain pseudo-dynamic loadings. These estimates, however, lack any supporting probabilistic structure that would induce smoothness and/or

capture sudden dynamics. Recent DFA developments [Del Negro and Otrok, 2008, Nakajima and West, 2013, Kaufmann and Schumacher, 2019] have treated both the factors and loadings as stochastic and dynamic. Adopting this point of view, we blend smoothness with sparsity via Dynamic Spike-and-Slab (DSS) priors on factor loadings [Rockova et al., 2020]. This prior regards factor loadings as arising from a mixture of two states: an inactive state represented by very small loadings and an active state represented by smoothly evolving large loadings. The mixing weights between these two states themselves are time-varying, reflecting past information to prevent from erratic regime switching. The DSS priors allow latent factors to effectively, and smoothly, appear or disappear from each series, tracking the evolution of sparsity over time.

In this work, we develop methodology for sparse dynamic factor analysis that is built on the three principles mentioned above. Using this methodology, we examine a large-scale balanced panel of macroeconomic indices that span multiple corners of the U.S. economy from 2001 to 2015. Our method helps understand how the economy evolves over time and how shocks affect its individual components. In particular, examining the latent factor structure before, during, and after the Great Recession, we obtain insights into the channels of dependencies and we assess permanence of structural changes.

To ensure that our implementation scales with large datasets, we propose an EM algorithm for MAP estimation that recovers evolving sparse latent structures in a fast and potent manner. An important consideration for any factor analysis tool is the interpretability of the latent factors. While interpretation can be achieved with ex-post rotations [Bai and Ng, 2013, Kaufmann and Schumacher, 2017, 2019], here we deploy parameter expansion, with rotations to sparsity *inside* the EM algorithm (Section 5.3.1) along the lines of Rockova and George [2016] to (a) accelerate convergence and (b) obtain better oriented sparse solutions. We also provide a more traditional estimation strategy using MCMC (Section 5.3.2) using the conventional lower triangular identification constraint [Nakajima and West, 2013,] on the factor loading matrices.

The chapter is structured as follows. Section 5.2 outlines the dynamic sparse factor model. Section 5.3 summarizes our estimation strategy with a parameter expanded EM algorithm, followed by an alternative MCMC implementation technique. A detailed simulation study that highlights the interpretability of our strategy relative to other methods is in Section 5.4, followed by an empirical study on a large-scale macroeconomic dataset in Section 5.5. In Section 5.6, we demonstrate the forecasting accuracy of our method, compared to some key competitors, on simulated and real datasets. We conclude the paper with additional comments in Section 5.7. Details of the implementation are in the appendix.

5.2 Dynamic Sparse Factor Models

The data setup under consideration consists of a matrix of high-dimensional multivariate time series $\boldsymbol{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T] \in \mathbb{R}^{P \times T}$, where each vector $\boldsymbol{Y}_t \in \mathbb{R}^P$ contains a snapshot of continuous measurements at time t. Dynamic factor models are built on the premise that there are only a few latent factors that drive the co-movements of $\boldsymbol{Y}_t$. Evolving covariance patterns of time series can be captured with the following state space model:

$$\boldsymbol{Y}_t = \boldsymbol{B}_t \boldsymbol{\omega}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{ind}{\sim} \mathcal{N}_P(\mathbf{0}, \boldsymbol{\Sigma}_t), \tag{5.1}$$

$$\boldsymbol{\omega}_t = \boldsymbol{\Phi} \boldsymbol{\omega}_{t-1} + \mathbf{e}_t, \quad \mathbf{e}_t \overset{ind}{\sim} \mathcal{N}_K(\mathbf{0}, \mathbb{I}_K), \tag{5.2}$$

which extends the more standard dynamic factor models [Sargent et al., 1977, Geweke, 1977] in at least two ways. First, the observation equation (5.1) links $\boldsymbol{Y}_t$ to a vector of factors $\boldsymbol{\omega}_t$ through multivariate regression with loadings $\boldsymbol{B}_t \in \mathbb{R}^{P \times K}$ and with residual variances $\boldsymbol{\Sigma}_t = \text{diag}\{\sigma_{1t}^2, \ldots, \sigma_{Pt}^2\}$, where *both* $\boldsymbol{B}_t$ and $\boldsymbol{\Sigma}_t$ are *dynamic*, i.e. are allowed to evolve over time. In this section, we tacitly assume that any location shifts in $\boldsymbol{Y}$ have been standardized away and thereby we omit an intercept in (5.1). The (dynamic) intercept can be however included, as we demonstrate in Section 5.5. Second, the transition equation (5.2) describes the unobserved factors $\boldsymbol{\omega}_t$ as following a stationary autoregressive process

with a transition matrix $\mathbf{\Phi} = diag(\widetilde{\phi}_1, \ldots, \widetilde{\phi}_K)$ with $0 < \widetilde{\phi}_k < 1$ for $k = 1, \ldots, K$ and with Gaussian disturbances $\mathbf{e}_t$. As is customary with state-space models of this type, we assume that $\boldsymbol{\omega}_t$, $\mathbf{e}_t$ and $\boldsymbol{\epsilon}_t$ are cross-sectionally independent.

A related approach was proposed in Aguilar and West [2000] and Lopes and Carvalho [2007], who also permit time-varying loadings, but do not impose the AR(1) process on the factors. Instead, their factors are cross-sectionally independent and linked over time through a stochastic volatility evolution of their idiosyncratic variances. Bai and Ng [2002] and Stock and Watson [2010], on the other hand, assume that factors follow vector autoregression, but the loadings are constant over time. As in Nakajima and West [2013], our model (5.1) and (5.2) differs from these more standard dynamic factor model formulations because it combines the AR(1) factor aspect together with dynamic loadings. A few remarks are in place. The assumption of independent and homoscedastic factor innovations may be unnecessarily restrictive. Estimating the factor covariance matrix in our framework is precluded due to lack of identification. This is because our auxiliary covariance matrices $\boldsymbol{A}_t$ (in the expanded model, to be described in Section 5.2.2) are not linked over time. We use parameter expansion to intentionally over-parametrize (Section 5.3.1) as a computational trick rather than as an attempt to model the co-volatilities. However, in related work Zhou et al. [2014] assume the factor covariance matrices $\boldsymbol{A}_t$ to be non-diagonal and time varying. These matrices can be reduced to diagonal matrices $\mathbf{\Psi}_t$ by pre and post multiplying by lower triangular matrices $\mathbf{L_t}$, with diagonal elements equal to one, which are in fact the Cholesky factors of the matrices $\boldsymbol{A}_t$. Suitable dynamic priors are then imposed on individual elements of both $\mathbf{\Psi}_t$ and $\boldsymbol{L}_t$. The model is made identifiable by assuming the factor loading matrices $\boldsymbol{B}_t$ to be lower-triangular. Another way to makes the identification problem less severe would be assuming certain dynamics on $\boldsymbol{A}_t$ with identifiability inherited from the initial condition $\boldsymbol{A}_0$.

The equations (5.1) and (5.2) imply that, marginally, $\boldsymbol{Y}_t \sim \mathcal{N}_P(0, \widetilde{\mathbf{\Sigma}}_t)$, where $\widetilde{\mathbf{\Sigma}}_t = \boldsymbol{B}_t \mathbf{V} \boldsymbol{B}_t' + \mathbf{\Sigma}_t$ with $\mathbf{V}$ being the stationary autoregressive covariance matrix of the latent

factors.[1] This decomposition provides a fundamental justification for factor-based dynamic covariance modeling. The information in high-dimensional vectors $\boldsymbol{Y}_t$ is distilled through latent factors into lower-dimensional factor loadings matrices $\boldsymbol{B}_t$, which *completely* characterize the movements of covariances over time. Other authors [Del Negro and Otrok, 2008, Lopes and Carvalho, 2007] consider a stochastic volatility (SV) evolution (either log-AR(1) or Bayesian discounting) on the variance of the latent factors and/or the innovations $\boldsymbol{\epsilon}_t$ in (5.1). While both are feasible within our framework, here we impose Bayesian discounting SV formulation on the innovation variances: $\sigma_{jt} = \sigma_{jt-1}\delta/\upsilon_{jt}$, where $\delta \in (0, 1]$ is a discount parameter and where $\upsilon_{jt} \sim \mathcal{B}(\delta\eta_{t-1}/2, (1-\delta)\eta_{t-1}/2)$ with $\eta_t = \delta\eta_{t-1} + 1$. We use this stochastic discounting model due to its computational convenience (with Kalman filtering equation) as explained in, for example, Chapter 10 of West and Harrison [1997] and Chapter 4 of Prado and West [2010].

Parsimonious covariance estimation is only one of the objectives of dynamic factor modeling. The more traditional objective is disentangling the covariance structure and understanding its driving forces and how they change over time. *Sparse* modeling has been indispensable for both of these objectives, where fewer estimable coefficients yield far more stable covariance estimates and where nonzero patterns in $\boldsymbol{B}_t$ yield superior interpretable characterizations [Carvalho et al., 2008, Yoshida and West, 2010]. Next, we explore the role of dynamic sparsity in DFA.

5.2.1 Dynamic Sparsity with Shrinkage Process Priors

No assumption has been as pervasive in the analysis of high-dimensional data as the one of sparsity. Sparsity is a practical modeling choice that facilitates high-dimensional inference and/or computation. In factor model contexts, it can also be used to anchor on identifiable parametrizations [Frühwirth-Schnatter and Lopes, 2009] and/or for estimating factor dimensionality [Rockova and George, 2016, Bhattacharya and Dunson, 2011]. The potential

1. $\boldsymbol{V}$ is the implicit solution to $\boldsymbol{V} = \boldsymbol{\Phi}\boldsymbol{V}\boldsymbol{\Phi}' + \mathbb{I}_K$, e.g. when $\boldsymbol{\Phi} = \widetilde{\phi}\mathbb{I}_K$, $\boldsymbol{V} = \frac{1}{1-\widetilde{\phi}^2}\mathbb{I}_K$.

of sparsity in dynamic factor models has begun to be recognized [Nakajima and West, 2013, Kaufmann and Schumacher, 2017, 2019, Kaufmann and Beyeler, 2018].

In this work, we complement the factor model formulation (5.1) with dynamic sparsity priors on the factor loadings $\boldsymbol{B}_t$ for $1 \leq t \leq T$. In other words, rather than imposing a dense model by assigning a random walk (or a stationary autoregressive) prior on the loadings [such as Stock and Watson, 2002, Del Negro and Otrok, 2008], we allow for the possibility that the loadings are zero at certain times.

We will write $\boldsymbol{B}_t = (\beta_{jk}^t)_{j,k=1}^{P,K}$ and impose a shrinkage process prior on individual time series $\{\beta_{jk}^t\}_{t=1}^T$ for each (j,k). Several authors have reported on the benefits of dynamic variable selection in the analysis of macroeconomic data [Frühwirth-Schnatter and Wagner, 2010, Bitto and Frühwirth-Schnatter, 2019, Lopes et al., 2010, Nakajima and West, 2013, Koop et al., 2010]. We build on one of the more recent developments, the Dynamic Spike-and-Slab (DSS) priors proposed by Rockova et al. [2020].

DSS priors are dynamic extensions of spike-and-slab priors for variable selection [George and McCulloch, 1993, Rockova and George, 2018]. Each coefficient in DSS is thought of as arising from two latent states: (1) an *inactive* state, where the coefficient meanders randomly around zero, and (2) an *active* state, where the coefficient walks on an autoregressive path. The switching between these two states is driven by a *dynamic* mixing weight which depends on past values of the series, making the states less erratic over time.

As DSS processes are a mixture of states, it relates closely to the broader framework of mixture autoregressive (MAR) processes, where the mixing weights are time dependent, e.g. Markov switching models, [Wong and Li, 2000, 2001, Kalliovirta et al., 2015, Wood et al., 2011]. Here, we utilize DSS priors in the context of dynamic latent factor modeling. A key feature of DSS priors is that, compared to latent thresholding, it yields benchmark continuous spike-and-slab priors (such as the Spike-and-Slab LASSO of Rockova and George 2018) as its marginal stationary distribution, guaranteeing marginal stability in the selection/shrinkage dynamics and probabilistic coherence.

We begin by reviewing the conditional specification of the DSS prior. For each coefficient β_{jk}^t, we have a binary indicator $\gamma_{jk}^t \in \{0,1\}$, which encodes the state of β_{jk}^t (the "spike" inactive state for $\gamma_{jk}^t = 0$ and the "slab" active state for $\gamma_{jk}^t = 1$). Given γ_{jk}^t and a lagged value β_{jk}^{t-1}, we assume a conditional mixture prior (independently for each (j,k)):

$$\pi(\beta_{jk}^t|\gamma_{jk}^t, \beta_{jk}^{t-1}) = (1-\gamma_{jk}^t)\psi_0(\beta_{jk}^t|\lambda_0) + \gamma_{jk}^t\psi_1\left(\beta_{jk}^t \,|\, \mu(\beta_{jk}^{t-1}), \lambda_1\right), \tag{5.3}$$

where

$$\mu(\beta_{jk}^{t-1}) = \phi_0 + \phi_1(\beta_{jk}^{t-1} - \phi_0) \quad \text{with} \quad |\phi_1| < 1 \tag{5.4}$$

and

$$\mathbb{P}(\gamma_{jk}^t = 1|\beta_{jk}^{t-1}) = \theta_{jk}^t. \tag{5.5}$$

The conditional prior (5.3) is a mixture of two components: (i) a spike Laplace/Gaussian density $\psi_0(\beta|\lambda_0)$ that is concentrated around zero and (ii) a Gaussian slab density $\psi_1(\beta_t|\mu(\beta_{jk}^{t-1}), \lambda_1)$, which is moderately peaked around its mean $\mu(\beta_{jk}^{t-1})$ with variance λ_1. This mixture formulation is an extension of existing continuous spike-and-slab priors [George and McCulloch, 1993, Ishwaran et al., 2005, Rockova, 2018], allowing the mean $\mu(\beta_{jk}^{t-1})$ of the non-negligible coefficients to evolve smoothly over time (through a stationary autoregressive process of order 1).[2] The spike distribution $\psi_0(\beta_t|\lambda_0)$, on the other hand, *does not* depend on β_{jk}^{t-1}, effectively shrinking the negligible coefficients towards zero. In this regard, the conditional prior in (5.3) can be seen as a "multiple shrinkage" prior [George, 1986,] with two centers of gravity.

In time series data (as will be seen from our empirical study), it is reasonable to expect that some factors are active only for some periods of time. Such "pockets of predictability" [Farmer et al., 2018] can be captured with spike/slab memberships γ_{jk}^t that evolve somewhat

2. While our framework can be extended to higher order autoregressive processes, we outline our methodology for first order autoregression with $\phi_0 = 1$ due to its universal usage in practice ([West and Harrison, 1997, Prado and West, 2010])

smoothly. This behavior can be encouraged with dynamic mixing weights θ_{jk}^t (defined in (5.5)) that reflect past information. Thus, in a context where we expect the latent structure to change somewhat smoothly over time, it is important that the sequence of slab probabilities, $\theta_{1:T}$, seamlessly evolve over times as well. Because each series $\{\theta_{jk}^t\}$ is a key driving force behind the sparsity in $\{\beta_{jk}^t\}$, it is crucial that it be (marginally) stable. To this end, we deploy the deterministic construction of Rockova et al. [2020] defined, for some global balancing parameter $0 < \Theta < 1$, as follows

$$\theta_{jk}^t \equiv \theta(\beta_{jk}^t) = \frac{\Theta\psi_1^{ST}\left(\beta_{jk}^{t-1}|\lambda_1, \phi_0, \phi_1\right)}{\Theta\psi_1^{ST}\left(\beta_{jk}^{t-1}|\lambda_1, \phi_0, \phi_1\right) + (1-\Theta)\psi_0\left(\beta_{jk}^{t-1}|\lambda_0\right)}, \tag{5.6}$$

given $(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1)$. This mixing weight has an interesting interpretation. It is defined as the *marginal* inclusion probability $\mathbb{P}(\gamma_{jk}^{t-1} = 1 \mid \beta_{jk}^{t-1})$ for classifying β_{jk}^{t-1} as arising from the *stationary* slab distribution $\psi_1^{ST}\left(\beta_{jk}^{t-1}|\lambda_1, \phi_0, \phi_1\right)$, as opposed to the stationary spike distribution $\psi_0\left(\beta_{jk}^{t-1}|\lambda_0\right)$, under the prior $\mathbb{P}(\gamma_{jk}^{t-1} = 1) = \Theta$. As θ_{jk}^t's evolve over time, they project the latent state (active/inactive) of the past value onto the next values. These weights induce marginal stability in the sense that each coefficient β_{jk} has a *marginal spike-and-slab distribution*, i.e. $\pi(\beta_{jk}) = \Theta\psi_1^{ST}\left(\beta_{jk}^t|\lambda_1, \phi_0, \phi_1\right) + (1-\Theta)\psi_0\left(\beta_{jk}^t|\lambda_0\right)$, which follows from the theorem by Rockova et al. [2020] given below:

Theorem 5.2.1. *Assume* $\{\beta_t\}_{t=1}^T \sim DSS(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1)$ *with* $|\phi_1| < 1$. *Then* $\{\beta_t\}_{t=1}^T$ *has a stationary distribution characterized by the following univariate marginal distributions:*

$$\pi^{ST}(\beta|\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1) = \Theta\,\psi_1^{ST}(\beta \mid \lambda_1, \phi_0, \phi_1) + (1-\Theta)\psi_0\left(\beta \mid \lambda_0\right), \tag{5.7}$$

where $\psi_1^{ST}(\beta \mid \lambda_1, \phi_0, \phi_1)$ *is the stationary slab distribution.*

Having introduced the DSS priors, we can now fully specify our dynamic latent factor model with (5.1)-(5.5). It is possible to extend our model to non-stationary random walk slab process, (obtained with $\phi_1 = 1$) by allowing transition weights θ_{jk}^t to be random, equal

to some deterministic sequence (e.g. as in Nakajima and West [2013]) or to a fixed value $\theta_{jk}^t = \widetilde{\theta}$ for $1 \leq t \leq T$. When treated as random, the weights may be prone to transitioning too often between the spike/slab states creating instabilities over time.

Our sparse dynamic factor model is related to the approach of Nakajima and West [2013], who zero out loadings whenever their autoregressive path drops below a certain threshold [see Rockova et al., 2020, for comparisons]. Another related approach is by Kaufmann and Beyeler [2018], who induce a point-mass spike and slab prior on the loadings. However, their approach (a) does not link the inclusion indicators and loadings over time, and (b) MCMC is deployed for calculations. Here, we develop both MCMC and an EM estimation procedure which does not rely on strict identifiability constraints.

5.2.2 Identifiability Considerations

Factor models are not free from identifiability problems owing to the fact that the model (5.1) and (5.2) is observationally equivalent to $\boldsymbol{Y}_t = \boldsymbol{B}_t^\star \boldsymbol{\omega}_t^\star + \boldsymbol{\epsilon}_t$ and $\boldsymbol{\omega}_t^\star = \boldsymbol{\Phi}\boldsymbol{\omega}_{t-1}^\star + \boldsymbol{e}_t$, where $\boldsymbol{\omega}_t^\star = \boldsymbol{A}_t\boldsymbol{\omega}_t$ and $\boldsymbol{B}_t^\star = \boldsymbol{B}_t\boldsymbol{A}_t'$ for any orthonormal matrix $\boldsymbol{A}_t$. Identification restrictions are particularly important for Bayesian analysis with MCMC, where meaningful interpretation of $\boldsymbol{B}_t$ could be prevented by averaging over various model orientations in the Markov Chain. To ensure identifiability, it is customary to restrict $\boldsymbol{B}_t$ to be lower-triangular, with ones on the diagonal [Nakajima and West, 2013, Aguilar and West, 2000, Lopes and West, 2004, Lopes and Carvalho, 2007, Zhou et al., 2014] or some variant of this form [Frühwirth-Schnatter and Lopes, 2009]. Identifiability in *sparse* factor models is even more delicate [Frühwirth-Schnatter and Lopes, 2009]. Nevertheless, these constraints render the analysis dependent on the ordering of the responses. Even without strict identifiability constraints, one needs to verify ex-post that the estimated sparse loadings satisfy identifiability constraints (as discussed e.g. by Frühwirth-Schnatter and Lopes [2009]). Bayesian ex-post MCMC strategies have been proposed that *do not* deploy identifiability constraints during the estimation stage [Kastner et al., 2017, Kaufmann and Schumacher, 2019]. Instead, posterior draws coming

from potentially very different orientations (identification schemes) are rotated ex-post.

For implementing our EM algorithm, we also do not impose any strict identifiability constraints on our model. Instead, we induce soft identifiability through sparsity priors and we let the EM optimization strategy converge towards one sparse posterior mode. Unlike with MCMC (an averaging strategy mixing over various sparse orientations), the EM output is conditional on one particular orientation and can be interpreted as such. To accelerate convergence and improve the chances of reaching better local modes, we use parameter expansion with automatic rotations to sparsity, as implemented by Rockova and George [2016]. Unlike the ex-post rotations deployed in Frühwirth-Schnatter and Lopes [2009], our rotations are performed *inside* the algorithm to gear the EM trajectory towards promising modes. This corresponds to a variant of the PX-EM algorithm of (Liu et al. [1998] and the one-step late PX-EM of Van Dyk and Tang [2003] for Bayesian factor analysis, where the augmented data log likelihood is maximized as a function of the augmented parameter within each EM iteration. This is in contrast to conditional data augmentation of Meng and Van Dyk [1998], where one seeks an optimal value of the augmented parameter before starting the EM algorithm. Similar data augmentation strategies can also be used to speed-up MCMC covergence, as demonstrated by the conditional and marginal data augmentation approaches of Meng and Van Dyk [1999]. In the context of Bayesian factor analysis, Ghosh and Dunson [2009] proposed a prior specification through parameter expansion that facilitates posterior computation. Yu and Meng [2011] proposed an ancillarity-sufficiency interweaving strategy for speeding-up MCMC convergence. This strategy was applied in the context of factor models in Kastner et al. [2017]. For our MCMC implementation, we will impose the usual constraints on loading matrices with a block lower triangular structure and with diagonal elements strictly positive [Geweke and Zhou, 1996, Aguilar and West, 2000, Lopes and Migon, 2002, Lopes and Carvalho, 2007]. Similar identifiability constraints has also been adapted by Nakajima and West [2013,], Zhou et al. [2014], in conjunction with latent thresholding.

5.2.3 Estimating Factor Dimensionality

The factor model (5.1) and (5.2) is formulated conditionally on the number of factors $K \in \mathbb{N}$. As noted by Bai and Ng [2002], "the correct specification of the number of factors is central to both the theoretical and empirical validity of factor models." The authors propose a criterion and show that it is consistent for estimating K in high-dimensional setups. In another strand of research, sparsity has been exploited for determining the effective factor dimensionality [Frühwirth-Schnatter and Lopes, 2009]. In particular, Bayesian non-parametric formulations have been proposed [Bhattacharya and Dunson, 2011, Rockova and George, 2016], where K is extended to infinity, while making sure that the number of *nonzero* columns in $\boldsymbol{B}_t$ is finite with probability one. Treating the number of *active* factors as unknown and bounded by K in this way, the posterior output under our spike-and-slab priors can be used to determine the effective dimensionality. We adopt a similar approach to Rockova and George [2016], where K in (5.1) is purposefully over-estimated and the number of *nonzero* columns obtained under strict sparsity priors will indicate how many *effective* factors there are.

5.3 Estimation Strategy

To estimate the proposed dynamic latent factor model with DSS priors, we develop two computational methods: an EM algorithm [Dempster et al., 1977], which allows for fast identification of posterior modes by iteratively maximizing the conditional expectation of the log posterior, and a standard MCMC implementation that is comparatively slower and thereby less appealing for large data applications. We describe both approaches in the following subsections.

5.3.1 EM Algorithm

The EM algorithm is well-suited for latent variable models, such as factor analysis, where it has been deployed by multiple authors including Rubin and Thayer [1982], Watson and

Algorithm: *EM algorithm for Automatic Rotations to Sparsity*		
	Initialize $\boldsymbol{\Delta} = (\boldsymbol{B}_{0:T}, \boldsymbol{\Sigma}_{1:T})$	
	Repeat the following E-Step, M-Step and Rotation step until convergence	
	The E-Step	
		For $t = 1, \ldots, T$
E1:	Latent Features:	Get $\boldsymbol{\omega}_{t\,|\,T}, \boldsymbol{V}_{t\,|\,T}$ and $\boldsymbol{V}_{t,t-1\,|\,T}$ from the Kalman filter and smoother
E2:	Latent Indicators	Compute $\langle \gamma_{jk}^t \rangle$ for $j = 1, \ldots, P$, $k = 1, \ldots, K$,
		$\langle \gamma_{jk}^0 \rangle = \frac{\Theta \psi_1(\beta_{jk}^0 \vert 0, \frac{\lambda_1}{1-\phi^2})}{\Theta \psi_1(\beta_{jk}^0 \vert 0, \frac{\lambda_1}{1-\phi^2}) + (1-\Theta)\psi_0(\beta_{jk}^0 \vert 0, \lambda_0)}$
		$\langle \gamma_{jk}^t \rangle = \frac{\theta_{jk}^t \psi_1(\beta_{jk}^t \vert \phi \beta_{jk}^{t-1}, \lambda_1)}{\theta_{jk}^t \psi_1(\beta_{jk}^t \vert \phi \beta_{jk}^{t-1}, \lambda_1) + (1-\theta_{jk}^t)\psi_0(\beta_{jk}^t \vert 0, \lambda_0)}$ (5.8)
	The M-Step	
M1:	Loadings	For $t = 0, \ldots, T$
		Update β_{jk}^{t*}, for $j = 1, \ldots, P$, $k = 1, \ldots, K$ following equation (5.15) in the Appendix A.
M2:	Rotation Matrix	Set $\boldsymbol{A}_0 = \boldsymbol{I}_K$
		For $t = 1, \ldots, T$
		Update $\boldsymbol{A}_t = \boldsymbol{M}_{1t} - \boldsymbol{M}_{12t} - \boldsymbol{M}_{12t}' + \boldsymbol{M}_{2t}$, where
		$\boldsymbol{M}_{1t} = \phi^2 \left[\boldsymbol{\omega}_{t-1\,|\,T} \boldsymbol{\omega}_{t-1\,|\,T}' + \boldsymbol{V}_{t-1\,|\,T} \right]$
		$\boldsymbol{M}_{12t} = \phi \left[\boldsymbol{\omega}_{t-1\,|\,T} \boldsymbol{\omega}_{t\,|\,T}' + \boldsymbol{V}_{t,t-1\,|\,T} \right]$
		$\boldsymbol{M}_{2t} = \boldsymbol{\omega}_{t\,|\,T} \boldsymbol{\omega}_{t\,|\,T}' + \boldsymbol{V}_{t\,|\,T}$
M3:	Idiosyncratic Variance	Compute $\boldsymbol{\Sigma}_{1:T}$ using Forward Filtering Backward Smoothing
	The Rotation Step	
R:	Rotation	For $t = 0, \ldots, T$
		Get Cholesky decomposition $\boldsymbol{A}_t = \boldsymbol{A}_{tL} \boldsymbol{A}_{tL}'$
		Rotate $\boldsymbol{B}_t = \boldsymbol{B}_t^* \boldsymbol{A}_{tL}$

Table 5.1: Parameter Expanded EM algorithm for sparse Bayesian dynamic factor analysis

Engle [1983], Zuur et al. [2003] and, more recently, Rockova and George [2016]. EM can be motivated by two simple facts. First, if we knew the missing data, standard estimation techniques can be deployed to estimate model parameters. Second, once we update our beliefs about model parameters we can make a much better educated guess about the missing data. Iterating between these two steps provides a fast way of obtaining maximum likelihood estimates and posterior modes.

Our EM algorithm has a few extra features that make it particularly attractive for dynamic factor analysis. First, the DSS priors (with a Laplace spike at zero) create spiky posteriors with sparse modes at coordinate axes. These modes yield interpretable latent factor structures that are anchored on sparse representations without arbitrary identifiability constraints. Second, the number of *active* factors does not have to be pre-specified and can be inferred from the dynamically evolving sparse structure.

As we discussed in Section 2.2, the model is invariant under rotation of factor loading

matrices. While this lack of identifiability has been regarded as a setback, it can also be regarded as an opportunity. Rotational invariance creates ridge-lines in the posterior that connect posterior modes and that can guide optimization trajectories [Rockova and George, 2016]. We follow the parameter expansion approach [see also Liu et al., 1998, Liu and Wu, 1999] that intentionally over-parametrizes the model and takes advantage of the lack of identification to speed up convergence. Similarly as Rockova and George [2016], we work with the expanded model discussed in Section 5.2.2. We assume the initial condition $\boldsymbol{\omega}_0 \sim \mathcal{N}_K(\mathbf{0}, 1/(1-\widetilde{\phi}^2)\boldsymbol{I}_K)$ which is the stationary distribution of the latent factors when $\boldsymbol{\Phi} = \widetilde{\phi}\mathbb{I}_K$ for some $0 < \widetilde{\phi} < 1$. We impose the DSS prior on the individual entries of the *rotated* matrix $\boldsymbol{B}_t^\star = \boldsymbol{B}_t\boldsymbol{A}_{tL}^{-1}$. The idea is to rotate towards sparse orientations throughout the iterations of the EM algorithm. The key observation is as follows: while matrices $\boldsymbol{A}_t$ for $1 \leq t \leq T$ *cannot* be identified from the observed data $\boldsymbol{Y}$, they can be identified from the complete data. Denoting $\boldsymbol{\Gamma}_t = \left\{\gamma_{jk}^t\right\}_{j,k}$, we treat both $\boldsymbol{\Omega} = [\boldsymbol{\omega}_0, \ldots, \boldsymbol{\omega}_T]$ and $\boldsymbol{\Gamma} = \{\boldsymbol{\Gamma}_0, \ldots, \boldsymbol{\Gamma}_T\}$ as missing data. The reduced model is obtained by setting $\boldsymbol{A}_t = \boldsymbol{I}_K$ for all $1 \leq t \leq T$.

All the unknown components of our model can be divided into three categories: missing data ($\boldsymbol{\Omega}$, $\boldsymbol{\Gamma}$), estimated parameters ($\boldsymbol{B}_0$, $\boldsymbol{B}_{1:T}$, $\boldsymbol{\Sigma}_{1:T}$) and parameters whose values are pre-specified ($\widetilde{\phi}$, ϕ_1, λ_0, λ_1, $\boldsymbol{\Theta}$ and δ). Among the pre-specified parameters we set the AR coefficients $\widetilde{\phi}$ and ϕ_1 close to (i.e. slightly smaller than) 1. This is because we want the AR processes to be stationary but, at the same time, we do not want the current values to deviate too far away from the past. We recommend setting values in the range $[0.9, 1]$ for the AR parameters. For similar reasons, the discount factor δ is also set close to unity (0.95 to be precise). However, instead of being treated as fixed values, the AR coefficients ϕ_0 and ϕ_1 *can* be easily estimated by assigning suitable priors, as demonstrated in the MCMC implementation discussed in Section 3.2. Following the recommendation of [Rockova et al., 2020], we set a moderate penalty for the spike distribution $\lambda_0 = 0.9$ and a comparatively large slab variance $\lambda_1 = 10(1-\phi_1^2)$. This is to ensure that the penalty on the factor loadings is unimodal. The marginal importance weight $\boldsymbol{\Theta} = 0.9$ is chosen to be large because smaller

values provide an overwhelming support towards zero factor loadings.

Let us denote by $\boldsymbol{\Delta} = (\boldsymbol{B}_0, \boldsymbol{B}_{1:T}, \boldsymbol{\Sigma}_{1:T})$ the model parameters. The matrix $\boldsymbol{B}_0$ contains the initial conditions that are assumed to arise from the stationary spike-and-slab prior distribution and $\boldsymbol{B}_{1:T}$ denotes all matrices $\boldsymbol{B}_t$ for $1 \leq t \leq T$. The goal of the EM algorithm is to find parameter values $\widehat{\boldsymbol{\Delta}}$, which are most likely (*a posteriori*) to have generated the data, i.e. $\widehat{\boldsymbol{\Delta}} = \arg\max_{\boldsymbol{\Delta}} \log \pi(\boldsymbol{\Delta} \mid \boldsymbol{Y})$. This is achieved indirectly by iteratively maximizing the expectation of the augmented log-posterior, treating the hidden factors $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ as missing data. Starting with an initialization $\boldsymbol{\Delta}^{(0)}$, the $(m+1)^{st}$ step of the EM algorithm outputs $\boldsymbol{\Delta}^{(m+1)} = \arg\max_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta} \mid \boldsymbol{\Delta}^{(m)})$, where $Q(\boldsymbol{\Delta} \mid \boldsymbol{\Delta}^{(m)}) = \mathbb{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega}|\boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}[\log \pi(\boldsymbol{\Delta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega} \mid \boldsymbol{Y})]$ with $\mathbb{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega}|\boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}(.)$ denoting the conditional expectation given the observed data and current parameter estimates at the m^{th} iteration. The EM algorithm iterates between the E-step (obtaining the conditional expectation of the log-posterior) and the M-step (obtaining $\boldsymbol{\Delta}^{(m+1)}$). The *parameter-expanded EM* works in a slightly different manner.

The E-step of the *parameter-expanded version* operates in the reduced space (keeping $\boldsymbol{A}_t = \boldsymbol{I}_K$), while the M-step operates in the expanded space (allowing for general $\boldsymbol{A}_t$). Namely, the E-step computes the expectation $Q(\boldsymbol{\Delta} \mid \boldsymbol{\Delta}^{(m)})$ with respect to the conditional distribution of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ under the *original model* anchoring on $\boldsymbol{B}_t$ and $\boldsymbol{A}_t = \boldsymbol{I}_K$, rather than on $\boldsymbol{B}_t^{\star}$ and unrestricted $\boldsymbol{A}_t$. Thus, the updated $\boldsymbol{A}$ is *not* carried forward throughout the iterations. Instead, each E-step is anchored on $\boldsymbol{A} = \mathbb{I}_K$. The M-step obtains updates of values Δ by first computing the solution in the expanded space and by rotating back to the original space to obtain $\Delta^{(m+1)}$. The steps are detailed in the Appendix. The M-step, on the other hand, is performed in the *expanded* parameter space, where optimization takes place over $\boldsymbol{B}_{0:T}^{\star}$, $\boldsymbol{\Sigma}_{1:T}$, and $\boldsymbol{A}_{1:T}$. Updating $\boldsymbol{B}_{0:T}^{\star(m+1)}$ boils down to solving a series of independent penalized dynamic regressions. The idiosyncratic variances $\boldsymbol{\Sigma}_t = \text{diag}\{\sigma_{1t}^2, \ldots, \sigma_{Pt}^2\}$ for $t = 1, \ldots, T$ are estimated in the M-step using Forward Filtering Backward Smoothing (Table A2 in the Appendix) [Ch. 4.3.7 Prado and West, 2010] using the discount SV specification (as discussed in the Supplemental Material). Since $\boldsymbol{A}_{1:T}$ can be inferred from the complete

data, one can estimate these matrices in the M-step to leverage the information in the missing data. Nevertheless, the updated matrices $\boldsymbol{A}_{1:T}$ are *not* carried forward towards the next E-step (which uses $\boldsymbol{A}_t = \boldsymbol{I}_K$), but are used to *rotate* the solution $\boldsymbol{B}_{0:T}^{\star(m+1)}$ back towards the reduced space via $\boldsymbol{B}_t^{(m+1)} = \boldsymbol{B}_t^{\star(m+1)}\boldsymbol{A}_{tL}$. The steps of the algorithm are carefully explained in Section 5.8.2. The computations are summarized in Table 5.1. The convergence of the EM algorithm with parameter expansion is provably faster [Liu et al., 1998, Rockova and George, 2016].

5.3.2 MCMC

This section describes an MCMC algorithm for our dynamic factor model with dynamic spike-and-slab priors. The *DSS* prior specification here is slightly different from the setup considered for the EM algorithm. The Laplace spike distribution $\psi_0(\beta|\lambda_0) = \lambda_0/2\mathrm{e}^{-\lambda_0|\beta|}$ yields sparse posterior modes, a favorable feature for the EM algorithm. However, MCMC ultimately reports the posterior mean which is non-sparse even under the Laplace prior. We will therefore assume a Gaussian spike (instead of Laplace) to utilize its direct conditional conjugacy for posterior updating. In particular, we assume the following spike density for $\lambda_0 << \lambda_1$

$$\psi_0(\beta \,|\, \lambda_0) = \exp\{-\beta^2/(2\lambda_0)\}/\sqrt{2\pi\lambda_0}. \tag{5.9}$$

This yields the following conditional Gaussian distribution for individual factor loadings β_{jk}^t

$$\beta_{jk}^t \,|\, \gamma_{jk}^t, \beta_{jk}^{t-1} \sim \mathcal{N}\left(\gamma_{jk}^t\mu_{jk}^t \,,\, \gamma_{jk}^t\lambda_1 + (1-\gamma_{jk}^t)\lambda_0\right)$$

and transition weights θ_{jk}^t in (5.6) with the Gaussian stationary spike distribution $\psi_0^{ST}(\beta_{jk}^{t-1}|\lambda_0) = \psi_0(\beta \,|\, \lambda_0)$. An extension to the Laplace spike is possible with an additional augmentation step, casting the Laplace distribution as a scale mixture of Gaussians with an exponential mixing distribution [Park and Casella, 2008]. The MCMC algorithm has a Gibbs structure, sampling iteratively from the conditional posteriors of the regression co-

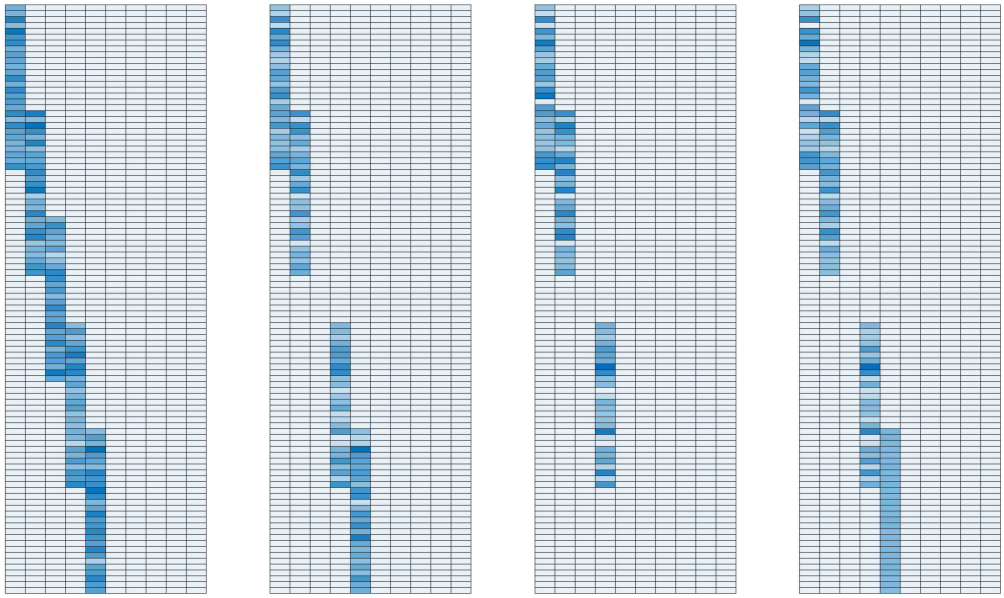


Figure 5.1: Simulation Study: The true latent factor loadings $\boldsymbol{B}_t^0$ at $t = 1, 101, 201, 301$.

efficients $\boldsymbol{B}_{0:T}$, latent indicators $\boldsymbol{\Gamma_{0:T}}$ and variances $\boldsymbol{\Sigma}_{0:T}$ (Frühwirth-Schnatter 1994; Prado and West 2010, Sect 4.5).

For the stationary DSS prior, we assume that the autoregressive parameter $|\phi_1| < 1$ is assigned the following beta prior (as in [Rockova et al., 2020])

$$\pi(\phi_1) \propto \left(\frac{1+\phi_1}{2}\right)^{a0-1} \left(\frac{1-\phi_1}{2}\right)^{b0-1} \mathbb{I}(|\phi_1| < 1) \quad \text{with } a_0 = 20 \text{ and } b_0 = 1.5, \qquad (5.10)$$

implying a prior mean of $2a_0/(a_0 + b_0) - 1 = 0.86$. We will update ϕ_1 with a Metropolis step, using a uniform proposal density on the interval $[0.8, 1]$. While we assume $\phi_0 = 0$ throughout, one can update ϕ_0 in a similar fashion. Table B1 in the appendix gives a step-by-step summary of the MCMC algorithm. A small demonstration on a simulated dataset is given in Appendix B.

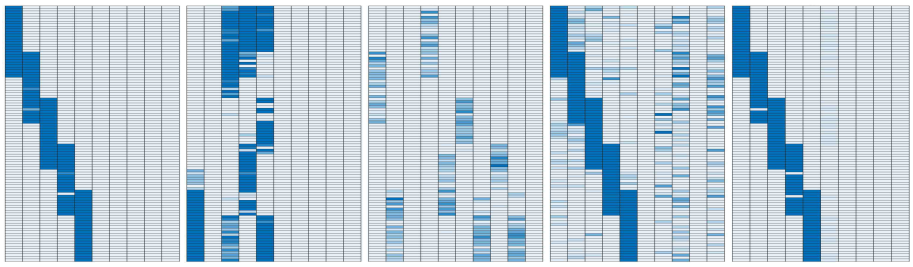
5.4 Simulation Study

We illustrate the usefulness of our proposed approach, relative to multiple existing methods, on synthetic data reflecting the following characteristics that can occur in real applications:

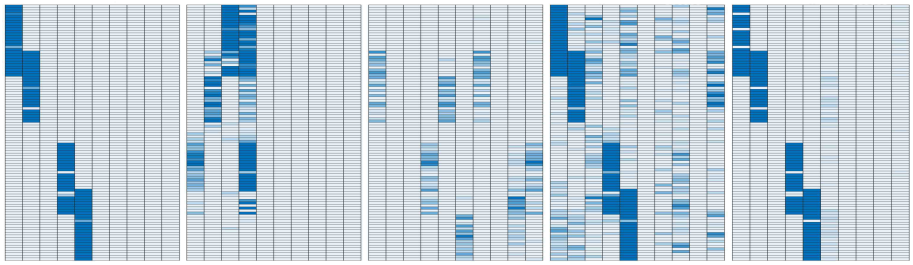
dynamic patterns of sparsity, smoothness, and a time-varying factor dimension.

First, we generate a single dataset with $P = 100$ responses, $K = 10$ candidate latent factors, and $T = 400$ time series observations (extra 100 data points are generated as training data for the rolling window analysis, as will be described below). The dimensionality of this example is already beyond practical limits of many Bayesian procedures. The elements of latent factors $\boldsymbol{\Omega}_t$ and idiosyncratic errors $\boldsymbol{\epsilon}_t$ are generated from a standard Gaussian distribution. Only the first five factors are potentially active over time, with the latter five being always inactive. We now describe the true loading matrices $\boldsymbol{B}^0 = [\boldsymbol{B}^0_1, \ldots, \boldsymbol{B}^0_T]$, which were used to generate the data, where $\boldsymbol{B}^0_t = \{\beta^{0t}_{jk}\} \in \mathbb{R}^{P\times K}$. At time $t = 1$, the active latent factor loadings form a block diagonal structure with 28 active loadings per factor, of which 10 overlap with another factor. In other words, we have 60 series with only one active factor, and 40 with two active factors (see the leftmost image in Figure 5.1). The sparsity pattern changes structurally over time where (a) at time $t = 101$ the loadings of the third factor become inactive, (b) at $t = 201$ the loadings of the fifth factor become inactive, and (c) at $t = 301$ the loadings of the fifth factor are re-introduced and active until $T = 400$ (Figure 5.1). The true nonzero loadings are smooth and arrive from an autoregressive process, i.e. $\beta^{0t}_{jk} = \phi\beta^{0t-1}_{jk} + v^t_{jk}$ with $v^t_{jk} \overset{iid}{\sim} \mathcal{N}(0, 0.0025)$ for $\phi = 0.99$, initiated at $\beta^{01}_{jk} = 2$ for all $1 \leq j \leq P$ and $1 \leq k \leq 5$. When loadings β^{0t}_{jk} become inactive, they are thresholded to zero. The true factor loadings are thereby smooth until they suddenly drop out and can emerge.

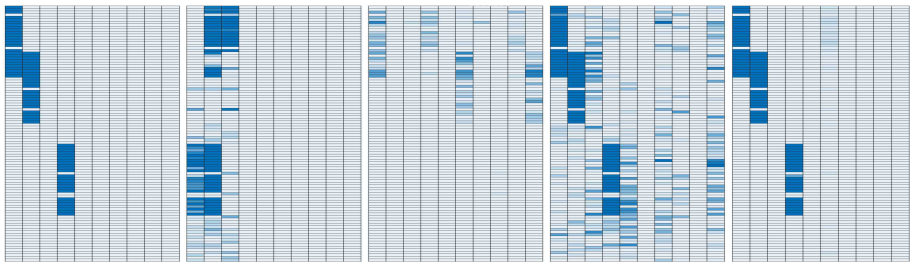
We compare our proposed dynamic spike-and-slab factor selection with three other approaches. The first one is the "rolling window" version of the static factor analysis with rotations to sparsity by Rockova and George [2016] using $K = 10$ (i.e. overshooting the true factor dimensionality). The rolling window size is set to 100, where we generate extra 100 data points as a training set, using the same sparsity pattern $\boldsymbol{B}^1_0$ as the first 100 observations. The rolling window strategy allows one to capture quasi-dynamics by iteratively incorporating new data in the training set. We compare this approach with "Adaptive PCA" of Bai and Ng [2002], which corresponds to a rolling-window principal component analysis



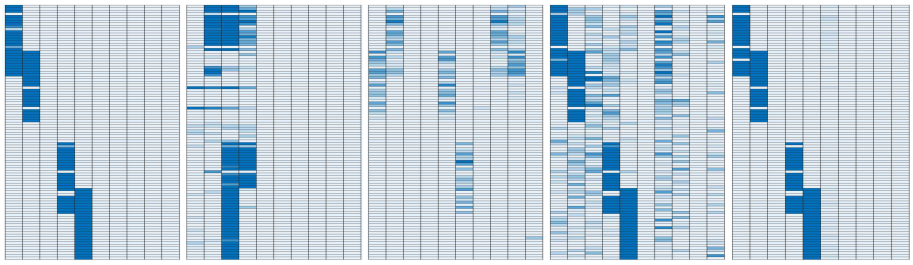
(a) $t = 100$



(b) $t = 200$



(c) $t = 300$



(d) $t = 400$

Figure 5.2: Simulated Example: Heatmaps of true and estimated factor loadings at $t = \{100, 200, 300, 400\}$. Comparisons are made between (from left to right), the true factor loadings, "Adaptive PCA," "Sparse PCA" ($K = 10$), rolling window spike-and-slab factor analysis ($K = 10$), and our dynamic spike-and-slab factor analysis. The first three methods are estimated using a rolling window of 100 data points. Factor loadings are absolute and capped at 0.5 for visibility.

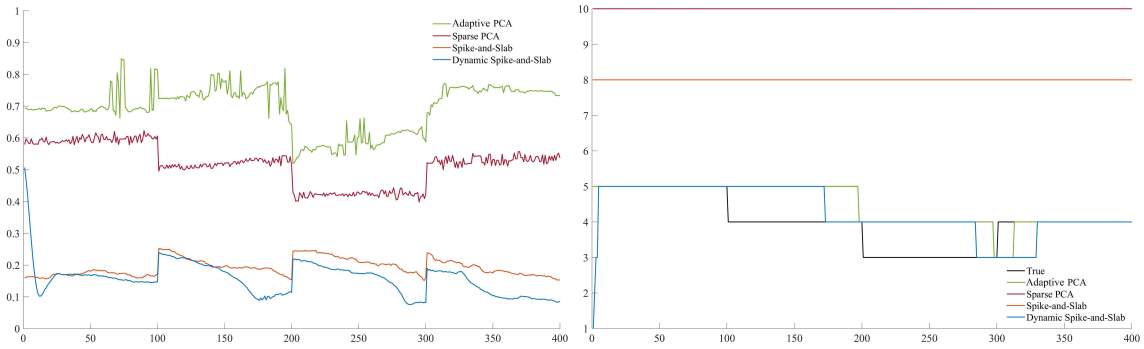
(PCA) with estimated number of factors, and with "Sparse PCA" using $K = 10$, which is a rolling-window LASSO-based regularization method with cross-validation for selecting the level of shrinkage [Witten et al., 2009]. All these methods are estimated using a rolling window of size 100, where we generate extra 100 training data points using the sparsity pattern $\boldsymbol{B}_1^0$. We choose $\boldsymbol{\Phi} = \widetilde{\phi}\mathbb{I}_K$ with $\widetilde{\phi} = 0.95$ and $K = 10$. Choosing $\widetilde{\phi}$ close to 1 ensures that the latent factor processes are stationary and their means do not deviate too far away from past values. To deploy the dynamic spike-and-slab priors, we set $\phi_0 = 0$, $\phi_1 = 0.98$, $\lambda_0 = 0.9$, $\lambda_1 = 10(1 - \phi_1^2)$, and $\Theta = 0.9$. To improve the performance of our EM method, we initialize the procedure using the output from the rolling window static spike-and-slab factor model of Rockova and George [2016].

Focusing on the reconstruction of factor loadings, we take snapshots at times $t = \{100, 200, 300, 400\}$ and visually compare the output to the truth (Figure 5.2). We see that both spike-and-slab methods achieve good recovery. However, the static spike-and-slab cannot fully contain the dynamic loadings, where we see a lot of spillover to other factors. Dynamic spike-and-slab shrinkage, on the other hand, smooths out the sparsity over time, clearly improving on the recovery. "Adaptive PCA" performs well, correctly specifying the number of factors. However, the factor loadings are non-sparse and rotated. "Sparse PCA" with $K = 10$ is fairly successful, recovering the blocking structure correctly, but splitting the signal among multiple factors [an observation made also by Rockova and George, 2016]. For the spike-and-slab methods, these patterns can be alternatively obtained by thresholding conditional inclusion probabilities rather than just looking at nonzero entries in $\widehat{\boldsymbol{B}}_{1:T}$.

We further explore how the root mean squared errors (RMSE) change over time for one of the simulations (Figure 5.3). This is calculated for each $t = 1 : T$ by

$$RMSE(\widehat{\boldsymbol{B}}_t) = \sqrt{\frac{tr(\boldsymbol{B}_t^0 - \widehat{\boldsymbol{B}}_t)'(\boldsymbol{B}_t^0 - \widehat{\boldsymbol{B}}_t)}{P \times K}}, \tag{5.11}$$

where $\widehat{\boldsymbol{B}}_t$ are the estimated factor loadings at time t. Since this comparison is not entirely



(a) RMSE (b) $\widehat{K}$

Figure 5.3: Simulation Study: (Left) The root mean squared error (5.11) and (Right) the estimated number of factors for "Adaptive PCA," "Sparse PCA," static spike-and-slab, and dynamic spike-and-slab, calculated for each $t = 1{:}400$.

meaningful due to the rotational invariance, we compute (5.11) for the left-ordered variants of these matrices. By looking at the speed of decrease in RMSE after a structural change, it is clear that dynamic spike-and-slab adapts faster compared to its rolling window counterpart. The drop of RMSE for "Adaptive PCA" in periods 101:200 and 201:300 can be attributed to the fact that the number of factors was estimated correctly, resulting in many true zero discoveries. On the other hand, the large estimation error of "Sparse PCA" is due to the lack of sparsity and scattered structure of the factors.

Additionally, we plot the estimated number of factors for each method and compare it to the true number of factors. "Sparse PCA" overestimates the number of factors (where we regard a factor as active if it has at least one nonzero loading). This indicates that unstructured sparsity is not enough. Looking at "Adaptive PCA" and our dynamic spike-and-slab factor model, we find that both perform similarly well in terms of estimating the number of factors. Furthermore, we note that dynamic spike-and-slab adapts faster to factors disappearing, while "Adaptive PCA" adapts faster to factors reappearing.

We repeat the experiment 10 times and report the average RMSE over each of the four stationary interim time periods in Table 5.2. Dynamic spike-and-slab achieves good recovery, improving upon the rolling window spike-and-slab by as much as 8% to 34% (except for

	t=1:100			t=101:200			t=201:300			t=301:400		
	RMSE	%	$\widehat{K}$	RMSE	%	$\widehat{K}$	RMSE	%	$\widehat{K}$	RMSE	%	$\widehat{K}$
Adaptive PCA	1.0660	-266.07	5	1.0590	-400.24	4.97	0.9730	-250.38	3.97	1.033	-430.01	3.88
Sparse PCA	0.7862	-169.99	10	0.7260	-242.94	10	0.6377	-129.64	10	0.7383	-278.81	10
Spike-and-Slab	0.1919	34.10	8	0.2843	-34.29	8	0.2988	-7.60	8	0.2447	-25.60	8
Dynamic Spike-and-Slab	0.2912	-	4.89	0.2117	-	4.72	0.2777	-	3.84	0.1949	-	3.71

Table 5.2: Simulation Study: Performance evaluation of the latent factor methods compared to the true coefficients for $t = 1{:}400$. Performance is evaluated based on RMSE within each evaluation period. % is the performance gain compared to dynamic spike-and-slab. $\widehat{K}$ is the average number of factors estimated during that period.

the first period). Large recovery errors of the "Sparse PCA" method can be explained by factor splitting. While "Adaptive PCA" does recover the correct number of factors at each snapshot, the loadings are non-sparse, rotated and non-smooth over time.

5.5 Empirical Study

The empirical application concerns a large-scale monthly U.S. macroeconomic database, (colloquially known as the FRED-MD dataset [McCracken and Ng, 2016] in the Macroeconomics literature) comprising a balanced panel of $P = 127$ monthly macroeconomic and financial variables tracked over the period of 2001/01 to 2015/12 ($T = 180$). These variables are classified into eight main categories, depending on their economic meaning: *Output and Income, Labor Market, Consumption and Orders, Orders and Inventories, Money and Credit, Interest Rate and Exchange Rates, Prices*, and *Stock Market*. A detailed description of how variables were collected and constructed is provided in McCracken and Ng [2016]. A quick table of names and groups of each variable is in the Appendix (Table B3). The variables were centered to have mean zero and standardized following the procedures in McCracken and Ng [2016].

This data, and its various subsets, have been widely studied in the literature, either as a standalone dataset (for macroeconomic forecasting and an impulse/response analysis) or as an essential part of broader data contexts. We review these analyses briefly below. For example, Stock and Watson [2018] deployed this dataset for estimation of dynamic

causal effects in Macroeconomics. In other analyses, Miranda-Agrippino and Ricco [2018] extracted a set of lagged macro-financial dynamic factors to project monetary policy shocks and Gargano et al. [2019] computed the Ludvigson-Ng (LN) macro factors for predicting bond values. Using a quarterly aggregated version of this data, Huber and Feldkircher [2019] fitted a Bayesian vector autoregressive model to forecast a subset of 21 variables. A larger forecasting exercise was conducted by Koop et al. [2019], who used 129 variables spanning over years 1960 to 2014 to predict GDP growth, inflation and short-term interest rates. A subset of this data, in conjunction with additional economic variables has been analyzed in Daniele and Schnaitmann [2019] who study the effects of a monetary policy shock through a regularized factor-augmented vector autoregressive (FAVAR) model. While the central theme of these works has been forecasting and/or impulse response analysis, the primary focus of our analysis in this section is discovering latent interpretable structures and glean insights into the interconnectivity between different sectors of the US macroeconomy, with a particular focus on the 2008 financial crisis. Forecasting will be discussed later in Section 5.6.

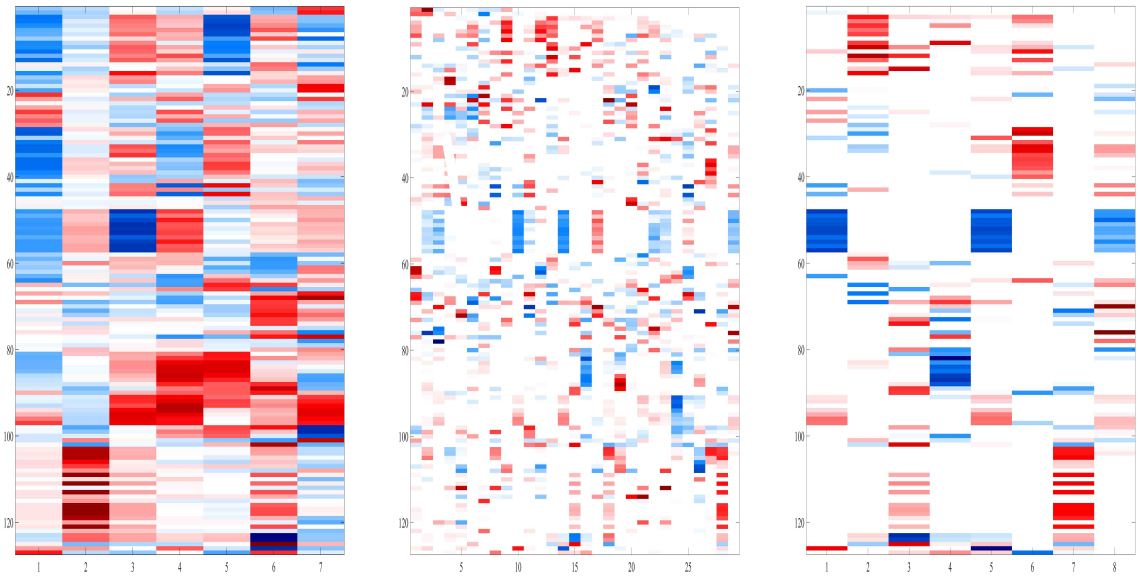
Stock and Watson [2005] analyzed a similar macroeconomic dataset (often referred to as the "Stock and Watson" dataset in Econometrics literature), containing 132 series over the sample 1959:1 to 2003:12. After performing variance decompositions, they found six factors that explain most of the variation in the data. With the same dataset, the IC1 and IC2 criteria developed in Bai and Ng [2002] find seven static factors explaining over 40 percent of the variation in the data. Bai and Ng [2013] used the same data extended to 2007:12 and showed first 7 factors still explain 45 percent of the variation in the data, though the IC2 criterion found the optimal number of factors to be 8.

The purpose of conducting a sparse latent factor analysis on a large-scale economic dataset, such as this one, is at least twofold. Due to the group structure of the data, it is natural to assume that the measured indicators are tied via a few latent factors, the basic premise of latent factor modeling. Moreover, we expect the sparse latent structure to detect

clusters of dependence structures that capture the interconnectivity of indicators spanning many *different* aspects of the economy. Sparsity will help extract such interpretable structures. Second, given the dynamic nature of the economy, there is a substantial interest in understanding how these dependencies change over time and– in particular– how they are affected by shocks. We anticipate non-negligible shifts in the economy, as the data spans over the housing bubble deflation after 2006 and the great financial crisis in late 2008, which led to the Great Recession. Understanding the interplay between contributing factors to the financial crisis has been a subject of rigorous research [see for example, Commission, 2011, Reinhart and Rogoff, 2008, Mian and Sufi, 2009, 2011, Mian et al., 2013, Chodorow-Reich, 2014, Benmelech et al., 2017]. Our analysis is purely data-driven and thereby descriptive rather than causally conclusive. We attempt to characterize patterns of shock proliferation and permanence of structural changes of the economy using our dynamic factor model.

As the dataset is considerably richer than our simulated example, we expand the model (5.1) by incorporating a dynamic intercept to capture location shifts that could not be easily standardized away. The intercepts c_{jt} follow independent random walk evolutions with an initial condition $c_0 \sim N(0, 1)$. The initial condition for the SV variances is $1/\sigma_{j0}^2 \overset{ind}{\sim} G(n_0/2, d_0/2)$ for $1 \leq j \leq P$ with $n_0 = 20$ and $d_0 = 0.002$. The discount factor is set to 0.95.

First, we examine one snapshot of the output from "Adaptive PCA" and "Sparse PCA" (described in Section 5.4) at time 2015/12 (Figures 5.4). Both methods do pick up certain groupings, but do not yield interpretable enough representations. This is likely due to overestimation of the number of factors (Figure 5.4 (b)), factor rotation and lack of sparsity (Figure 5.4 (a)) and/or factor splitting (Figure 5.4 (c)). Next, we deploy the rolling window spike-and-slab factor method with a training period of 10 years to obtain starting values for our dynamic factor model. Priors and their hyper-parameters were chosen as in the simulation study. We choose a generous upper bound $K = 126$ on the number of factors, letting the sparsity rule out factors that are irrelevant.



(a) "Adaptive PCA" (b) "Sparse PCA" (c) "Sparse PCA"

Figure 5.4: Macroeconomic Study: Estimated factor loadings using "Adaptive PCA" (Left), "Sparse PCA" with number of factors set as 30 (Middle), and "Sparse PCA" with number of factors set to 8 from the results of "Adaptive PCA" (Right) at $t = 2015/12$, with the number of series on the y-axis and the number of factors in the x-axis. The factor loading are estimated using a 10 year rolling window.

We now examine the output of our procedure at three time points: 2003/12, 2008/10, and 2015/12. These three snapshots are of particular interest as they represent three distinct states of the economy: relative stability (2003), sharp economic crisis (2008), and recovery (2015). 2008/10 is at the onset of the great financial crisis, where deflation of the housing bubble after 2006 lead to mortgage delinquencies and financial fragility [Commission, 2011]. This distress permeated throughout the rest of the economy, including the labor market, leading to the deepest recession in the post-war history.

The heatmap of estimated factor loadings at time 2003/12 is in Figure 5.5 (left). The output has been left-ordered based on the results at 2015/12, where the more active factors are on the left, in the order of data series, and some of the less active right-most factors (with small or zero loadings) are omitted. There are 24 active factors in total (i.e. factors with at least two non-negligible non-zero factor loadings), with only 5 factors that cluster

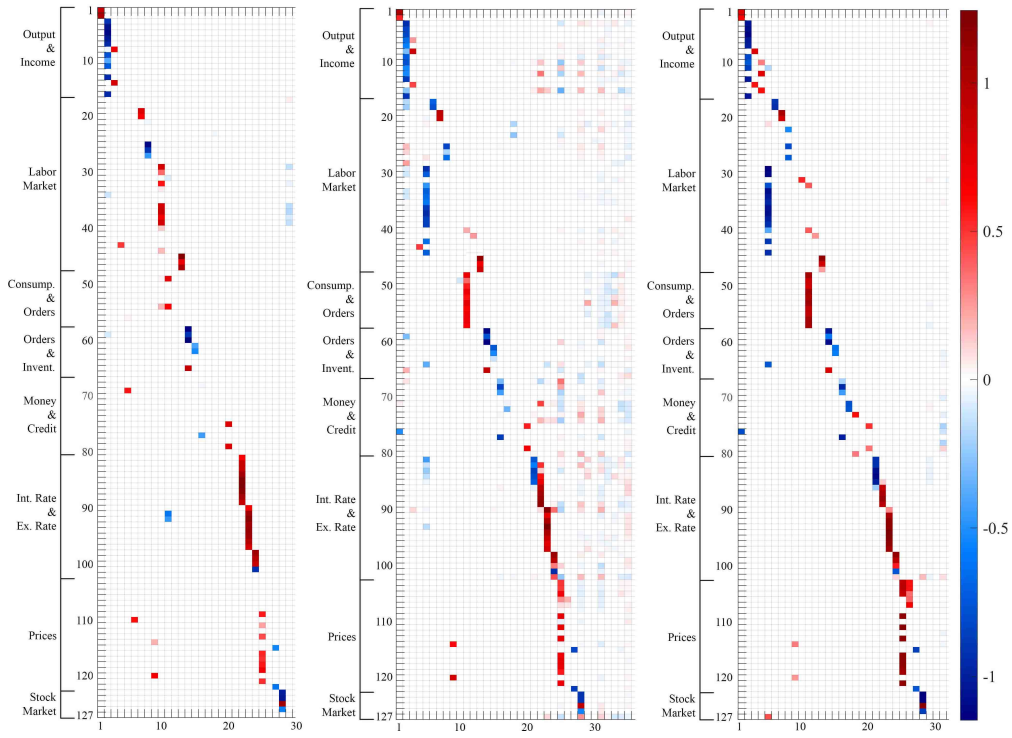


Figure 5.5: Macroeconomic Study: Estimated factor loadings using dynamic sparse factor analysis at $t = 2003/12$ (left), $t = 2008/10$ (center), $t = 2015/12$ (right), with the original series on the y-axis and the factors in the x-axis. The factor loading are estimated dynamically over the period 2001/1:2015/12.

eight or more series (Factors 2, 10, 22, 23, and 25). Since the variables are grouped by their economic meaning, this type of clustering is not entirely unexpected. For example, Factor 2 includes CMRMTSPLx (real manufacturing and trade industry sales), all industrial production indices except nondurable materials, residential utilities, and fuels, CUMFNS (capacity utilization), DMANEMP (durable goods employment), and ISRATIOx (manufacturing and trade inventories to sales ratio). This factor could be interpreted as a factor for *durable goods*, which include industries that are more susceptible to economic trends, where sales, inventories, industrial production, capacity utilization, and employment are all connected. Conversely, we expect nondurable goods, such as utilities and fuels, to have a different dynamic than durable goods, which is reflected in the exclusion of those indices in Factor 2. Similarly, Factor 10 includes employment data (except for mining and logging, manu-

facturing, durable goods, nondurable goods, and government), Factor 22 includes interests rates (fed funds rate, treasury bills, and bond yields), Factor 23 includes the spread between interest rates minus fed funds rate, and Factor 25 includes consumer price indices except apparel, medical care, durables, and services, as well as personal consumptions expenditures on nondurable goods. All of these factors produce meaningful and mostly separated clusters that largely conform with economic intuition.

During the crisis (Figures 5.5; center), radical changes occur in the factor structure. Concerning Factor 2, the dependence structure expands, now spanning over nondurables and fuels, as well as HWI (the help wanted index), UNEMP15OV (unemployment for 15 weeks and over), CLAIMSx (unemployment insurance claims), and PAYEMS (employment, total non-farm, goods-producing, manufacturing, and durable goods). This indicates that the shock might have affected relatively stable industries and unemployment, with the co-movement across industries being largely synchronized under distress (with the exception of residential utilities). Another interesting observation is the emergence of new factors. In particular, Factor 11, which includes housing starts and new housing permits in different regions in the U.S., was *not* present pre-crisis and now surfaces as a connecting thread between housing markets across regions. While in 2003/12 the latent factors were largely separated (loadings had little overlap), we now see at least two factors (namely Factor 25 and 28), whose loadings are non-sparse and far-reaching. In particular, Factor 28 emerges as a non-sparse link between many different sectors of the economy, including retail sales, industrial production, employment (in particular financial services), real M2 money stock, loans, BAA bond yields (but not AAA), exchange rates, consumer sentiment, investment and, most importantly, the stock market indices, including the S&P 500 and the VIX (i.e. the fear index), a popular measure of the stock market's expectation of volatility. This factor loads heavily on stock market indices, which were isolated pre-crisis, but are now connected to the various corners of the economy. Factor 25, on the other hand, is driven mainly by prices (e.g. CPI). Both of these factors could potentially be interpreted as crisis factors as

they permeate through various sectors of the economy, that had little interconnectivity in the pre-crisis era. The only sectors not influenced by these factors are Consumption and Orders and, more interestingly, the housing market.

There is an ongoing discussion on what were the main catalysts of the Great Recession. One line of reasoning focuses on the financial market, where the devaluation of securities, including mortgage backed securities, led to curtailed lending and thereby consumption [Chodorow-Reich, 2014, Benmelech et al., 2017]. The second one focuses directly on the downturn of the housing market [Mian and Sufi, 2009, 2011, Mian et al., 2013]. The "orthogonality" between the housing market factor (Factor 11) and the "crisis factors" (Factor 25 and 28) may suggest that, while the crisis was triggered by the housing market, the main catalyst of the recession was probably the financial market. While our analysis does not necessarily prove this hypothesis, it aligns with the previous lines of reasoning.

Finally, Figure 5.5 (right) shows the end of the analysis at 2015/12, where the economy has mostly recovered from the Great Recession, but has fundamentally changed from what it was before. Although most of the factor overlap has dissipated, we see a notably different structure compared to 2003. In particular, Factor 5 (employment) and Factor 11 (housing) persevere from the crisis. Moreover, the "crisis factors" Factor 25 and 28, representing the prices and the stock market, are no longer strongly tied to other parts of the economy (labor, output, interest and exchange rates, etc.). In addition, the VIX indicator for market sentiment, is no longer connected to many of the key factors, even the stock market, and is only connected to Factor 5, implying that the market's anticipation of volatility is no longer severely intertwined with the rest of the economy. Factor 2 is one of the few factors that have returned back to its original structure, except for CMRMTSPLx and industrial production of nondurable consumer goods. Its dependence with the labor market (e.g. unemployment) has disappeared, suggesting that industry production is no longer in co-movement with the labor market.

We also obtain insights into the effects and duration of the crisis by looking at the

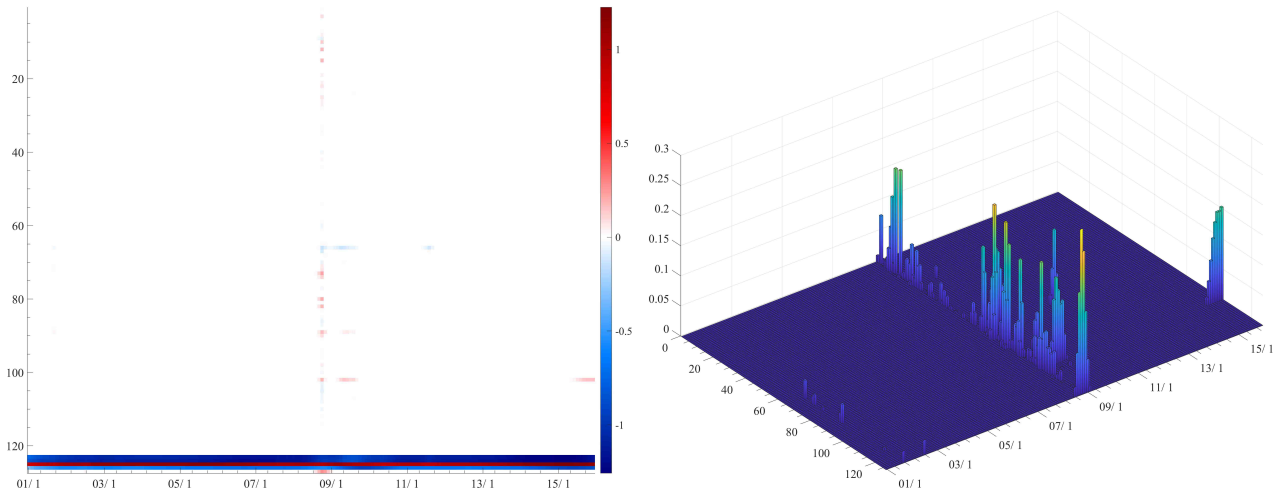


Figure 5.6: Macroeconomic Study: Estimated factor loadings for Factor 28 using dynamic spike-and-slab from $t = 200/12{:}2015/12$, with a heatmap of the entire factor loadings (Left) and a 3-D plot of the factor loadings with the loadings on 123-126 (S&P related indices) set to zero to increase visibility.

evolution of the factor loadings for one of the "crisis" factors, Factor 28. Figure 5.6 shows a dynamic heatmap and a 3-D plot of β_{jk}^t for $1 \leq j \leq 127$ (y-axis) and $1 \leq t \leq 180$ (x-axis) with $k = 28$. For the 3-D plot, the loadings on the S&P indices are suppressed to zero in order to improve visibility. The figure reveals a spur of activity around the sharp financial crisis (late 2008 and early 2009), where the contagion battered multiple corners of the economy. The duration of the active loadings provide additional insights. For example, the loadings on VIX (series 127) emerges and disappears in a eight month span from 06/2008 to 02/2009, while the loadings on the exchange rate between U.S. and Canada lasts for 17 months. However, most factor loadings seem to only emerge for about 4-6 months.

To understand the degree of connectivity/overlap between factors, we plot the average number of active factors (with absolute loadings truncated to above 0.1)[3] per series over time (Figure 5.7). More overlap indicates a more intertwined economy. We observe an increase in late 2008, reflecting the emergence of pervasive crisis factor(s), as well as its build up

3. We use the 0.1 as cutoff, because $(-0.1, 0.1)$ is approximately the shortest 10-percent confidence interval of the spike distribution (Laplace distribution centered at 0 and with variance being equal to 0.9) used in the dynamic spike and slab prior.

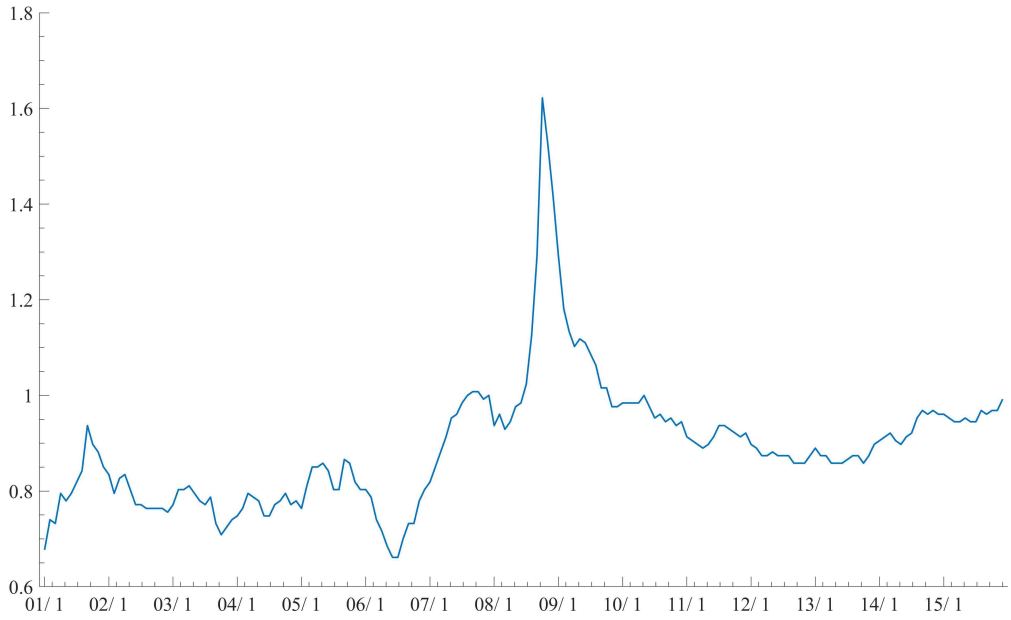


Figure 5.7: Macroeconomic Study: The average number of estimated active factors (with absolute loadings above 0.1)per series over the period 2001/1:2015/12.

from mid-2006. Another point to note is that the level pre-crisis is comparatively lower than post-crisis, indicating a structural shift is the economy brought on by the crisis.

We further our analysis with a few insights into the idiosyncratic variances for variables related to the housing market: HOUST (total housing starts) and its regional variants (North East, Mid-West, South, and West). We choose the housing market for deeper analysis, because the housing market has been subjected to intense scrutiny, following the great recession of 2009, as a suspected trigger of the crisis [Mian and Sufi, 2009, 2011, Mian et al., 2013]. Housing starts is the seasonally adjusted number of new residential construction projects that have begun during any particular month and, as such, is a key part of the U.S. economy, which relates to employment and many industry sectors including banking (the mortgage sector), raw materials production, construction, manufacturing, and real estate. In our earlier analysis (Figure 5.5) we found that, while regional indicators were not clustered pre-crisis, persistent clustering occurs post-crisis. Figure 5.8 portrays the series of residual uncertainties $\{\sigma^2_{jt} : 1 \leq t \leq T\}$ for each regional housing starts indicator. We find several interesting patterns. Figure 5.8 indicates that increased uncertainty in housing starts

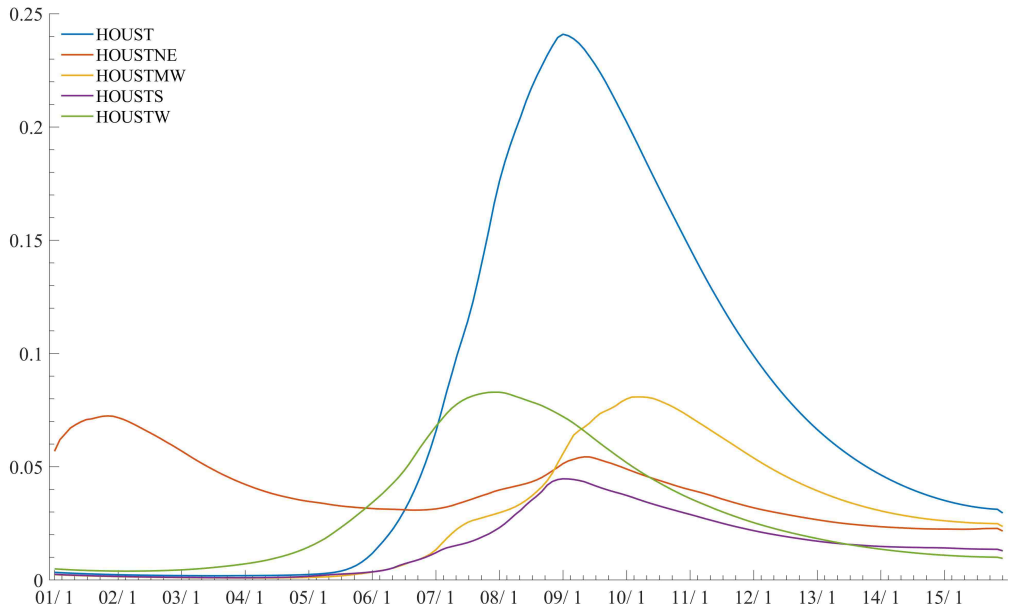


Figure 5.8: Macroeconomic Study: The idiosyncratic variance, $\mathbf{\Sigma}_t$, of U.S. housing starts, over the period 2001/1:2015/12.

is a global phenomenon but that there is heterogeneity across regions as to the magnitude and timing. For example, we find that the West region to react the earliest, followed by Mid-West and South. North-East is somewhat of an exception, as the idiosyncratic variance starts out greater than the other series, falling off pre-crisis, increasing during the crisis, and tapering off to a level similar to the other regions. The speed of mounting uncertainty could be associated with the deflation of the housing bubble after 2006 [Commission, 2011]. As the economy recovers from the Great Recession, we find a steady decrease in uncertainty in all regions, except for the South region, which is persistently high throughout the analysis long after the crisis. Interestingly, the South region was one of the hardest hit regions during the great recession, with the increase in unemployment being the highest of all the regions. Possibly due to this characteristic, we find that the South region does not return to the pre-crisis state. This is an important insight for the decision/policy maker, as this indicates some unique circumstances in the South region that requires further investigation, where the housing bubble from 2004-2006 bursts after mid-2006.

5.6 Forecasting Evaluations

In this section, we compare the forecasting performance of our method (point-wise predictions as well as forecast distributions) with three alternatives. The first one is a static version of our model that assumes that factor loading matrices are constant over time, keeping all the other model assumptions the same. The second method is the sparse Bayesian latent factor stochastic volatility model implemented in the R package "factorstochvol" [Kastner, G., 2017]. The third method is the hierarchical Bayesian vector autoregressive model (BVAR) of Kuschnig and Vashold [2019]. It implements a hierarchical modeling approach to prior selection in the fashion of Giannone et al. [2015]. For both "factorstochvol" and "BVAR" methods, we draw $15,000$ MCMC samples, of which $5,000$ samples are discarded as a burn-in. Both these methods are implemented through their corresponding R packages with default parameter settings. Forecasting comparison is conducted for four examples: (i) a smaller simulated data (described later in this section) with $p = 10$ and $T = 100$, (ii) the same simulated data as in (i) but extended to $T = 400$, (iii) the simulated data discussed in Section 5.4 and (iv) the macroeconomic data discussed in Section 5.5.

For all three factor analysis models: our "Dynamic FA", "Static FA" and "factorstochvol", we use the same upper bound on the number of latent factors (K). For the lower dimensional simulated datasets (i) and (ii) we assign $K = 6$ and for the higher dimensional simulated dataset (iii) we fix $K = 10$. In Section 5.5, we discovered that for the macroeconomic data, the number of active factors never exceeds $K = 28$. Therefore, for evaluating forecasting performance, we used $K = 30$ for all three factor analysis models to facilitate faster computation and higher accuracy.

Forecasting can be performed using both the EM (Section 5.3.1) and the MCMC (Section 5.3.2) implementations. Even though the EM algorithm is more practically feasible for larger datasets, a common drawback of this method is that we get only point estimates of the variables of interest over the forecast period, as opposed to the MCMC which provides the entire predictive distribution. Point forecasts for the time period $T + 1$ can be obtained by

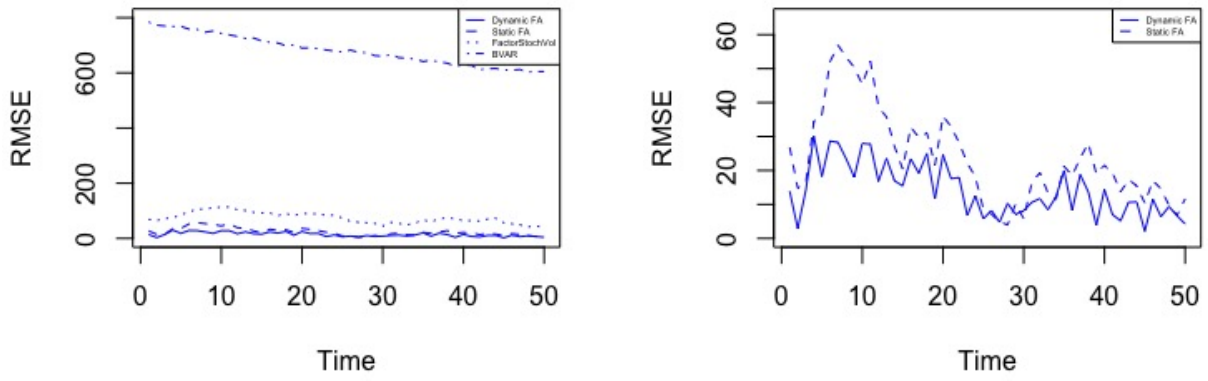


Figure 5.9: Root mean squared error (RMSE) computed over 50 one-step-ahead forecasts from a simulated data with $p = 10$ and $T = 100$. The dynamic sparse factor analysis model (Dynamic FA) is compared against (i) the Static spike and slab factor analysis (Static FA), (ii) FactorStochvol and (iii) Bayesian VAR (BVAR). The plot on the left shows forecasting RMSE over time for all four methods and the plot on the right zooms in on the Dynamic FA and Static FA models.

training the EM algorithm with data $\boldsymbol{Y}_{1:T}$ and then computing $\widehat{\boldsymbol{Y}}_{T+1} = \widehat{\boldsymbol{B}}_{T+1}\widehat{\boldsymbol{\omega}}_{T+1}$, where $\widehat{\boldsymbol{B}}_{T+1}$ and $\widehat{\boldsymbol{\omega}}_{T+1}$ are expectations of the future value conditional on the estimates obtained from the EM algorithm. Forecasting comparisons of predictive distributions obtained from MCMC are described towards the end of this section.

For the smallest simulated data (with $p = 10$ and $T = 100$), we conduct a sequential one-step ahead forecast for 50 consecutive time points into the future with the EM implementation of our model. The same forecasting exercise is also performed with the three competing methods under consideration. Then we plot the root mean squared prediction error (RMSE) over time for all the four methods in Figures 5.9(a) and 5.9(b). We simulate the $p = 10$-dimensional data $\boldsymbol{Y}_{1:150}$ for $T = 150$ time points as follows: The elements of latent factors $\boldsymbol{\Omega}_t$ and idiosyncratic errors $\boldsymbol{\epsilon}_t$ are generated from a standard Gaussian distribution. At time $t = 1$, the active latent factor loadings form a block diagonal structure with 5, 6 and 4 active loadings for factor 1, factor 2 and factor 3 respectively, of which factor 1

and factor 2 overlap for 2 series, while factor 2 and factor 3 overlap for 3 series. Factor 1 and factor 3 never overlap. The sparsity pattern changes structurally over time where (a) at time $t = 35$ the loadings of the third factor become inactive, (b) at $t = 45$ the loadings of the third factor are re-introduced and remain active until $T = 110$, (c) at time $t = 111$ the loadings of the third factor become inactive again and finally, (d) at $t = 120$ the loadings of the third factor are re-introduced and they remain active until $T = 150$. The true nonzero loadings are smooth and arrive from an autoregressive process, i.e. $\beta_{jk}^{0t} = \phi\beta_{jk}^{0t-1} + v_{jk}^{t}$ with $v_{jk}^{t} \overset{iid}{\sim} \mathcal{N}(0, 0.0025)$ for $\phi = 0.99$, initiated at $\beta_{jk}^{01} \overset{iid}{\sim} \mathcal{N}(2, 1)$ for all $1 \leq j \leq P$ and $1 \leq k \leq 5$. When loadings β_{jk}^{0t} become inactive, they are thresholded to zero. The true factor loadings are thereby smooth until they suddenly drop out and can re-emerge. Figure B2 in the Appendix shows a heatmap of the (absolute values of) true factor loadings for times $t = 20$, $t = 40$ and $t = 80$ respectively.

We start the forecasting exercise for the above data by first estimating the model with $\boldsymbol{Y}_{1:100}$ and then forecasting for $T = 101$. Then we incorporate the next observation $\boldsymbol{Y}_{101}$ into the training data, estimate the model again and forecast for $T = 102$. This process is repeated through $T = 101 : 150$. For each time point we compute the RMSE over the $p = 10$ simulated time series. Figure 5.9(a) shows how the RMSE varies over the 50 forecast points for all the four methods under consideration and Figure 5.9(b) zooms in onto our dynamic sparse factor analysis model and the static spike and slab factor model. We see that our dynamic factor model maintains superior forecasting performance, compared to the three competing methods throughout the forecast period. An interesting point to note here is that the forecast accuracy of our dynamic model persists even when the time series structurally changes at time points $T = 111$ and $T = 120$, where the number of factors changes from 3 to 2 and then again from 2 back to 3 respectively. In contrast, the factor models with static loadings (static FA and "factorstochvol") cannot adapt to these structural changes and their forecasting performance decline resulting in higher RMSE.

Next, we compare the five-step ahead forecasting performance of our dynamic model

Data	Dimension	Dynamic FA	Static FA	FactorStochVol	BVAR
Simulation	p=10, T=100	**20.427**	34.359	63.287	752.157
Simulation	p=10, T=400	**1.158**	13.435	2.613	230.75
Simulation	p=100, T=400	**1.407**	15.319	2.389	392.186
Macroeconomy	p=127, T=180	**0.623**	1.018	1.384	1.3453

Table 5.3: Root mean squared forecasting error over five time points into the future

(referred to as "Dynamic FA") with the static model described above (referred to as "Static FA"), the "factorstochvol" method and the BVAR method in Table 5.3. Comparison is done with respect to the cumulative root mean squared forecasting errors (RMSE), computed over five time points into the future. This is a five-step-ahead forecasting exercise, as opposed to the one-step-ahead forecasting over 50 time points demonstrated in Figures 5.9(a) and 5.9(b). The five-step-ahead forecast is conducted by sequential point forecasts from the EM algorithm over five time points into the future. Specifically, we fit the model based on the first T observations $\boldsymbol{Y}_{1:T}$, and compute the one-step ahead forecast $\boldsymbol{Y}_{T+1}$. Then, we add this forecast for $\boldsymbol{Y}_{T+1}$ into the training data $\boldsymbol{Y}_{1:T}$ and compute a forecast $\boldsymbol{Y}_{T+2}$ based on $\boldsymbol{Y}_{1:(T+1)}$. We repeat these sequential updates for five time points to predict $\boldsymbol{Y}_{(T+1):(T+5)}$. These predicted values are compared with the observed/simulated data to get a cumulative root mean squared prediction error. Similarly, for the macroeconomic data we use the first 175 months (2001/01 to 2015/07) to get sequential one-step-ahead forecast for the last 5 months (2015/08 to 2015/12). Table 5.3 shows that, for all the simulated datasets, the three factor analysis models (Dynamic FA, static FA and FactorStochVol) perform significantly better than the Bayesian VAR model. For the macroeconomic data, our Dynamic factor analysis model appears to perform considerably better than the alternatives. This reiterates the merits of using dynamic factor loadings as opposed to constant loading matrices.

The above observations are confirmed after computing the one-step-ahead log predictive density scores (LPDS) measuring the quality of the entire forecast distributions. For this forecasting comparison, we use our MCMC implementation (Section 5.3.2). As described by Kastner [2019], the one step ahead LPDS for the dynamic factor model can be computed

by first drawing M MCMC samples from the distribution of $\mathbf{Y_{1:T}}$ and then averaging over $m = 1, \ldots, M$ densities of

$$\mathcal{N}_p\left(\mathbf{0}, \boldsymbol{B}^{(m)}_{(T+1):[1:T]}\boldsymbol{B}^{(m)'}_{(T+1):[1:T]} + \boldsymbol{\Sigma}^{(m)}_{(T+1):[1:T]}\right)$$

evaluated at $\boldsymbol{Y}_{T+1}$, where $\boldsymbol{B}^{(m)}_{T+1:[T+1]}$ and $\boldsymbol{\Sigma}^{(m)}_{(T+1):[1:T]}$ denote the m-th draw of $\boldsymbol{B}_{T+1}$ and $\boldsymbol{\Sigma}_{T+1}$ respectively, from the posterior distribution up to time T. Next, we compute the one-step ahead log predictive Bayes factor between our dynamic factor model and the static spike-and-slab factor model. Such Bayes factor between any two models $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined as $\log BF(\mathcal{M}_1, \mathcal{M}_2) = \log PL_{T+1}(\mathcal{M}_1) - \log PL_{T+1}(\mathcal{M}_2)$, where $PL_t(\mathcal{M})$ denotes the predictive likelihood of model $\mathcal{M}$ at time $T+1$. When the log predictive Bayes factor is greater than zero at a given point in time, there is evidence in favor of model $\mathcal{M}_1$ as opposed to model $\mathcal{M}_2$, and vice versa. For the simulated examples with $p = 10$ and $T = 100$, the Bayes factor computed between our dynamic factor model and the static factor model turn out to be equal to

$$\log BF(\text{Dynamic FA, Static FA}) = 2.161$$

implying (strong) evidence in favor of the dynamic sparse factor model.

5.7 Further Comments

Motivated by a topical macroeconomic dataset, we developed a Bayesian method for dynamic sparse factor analysis for large-scale time series data. Our proposed methodology aims to tackle three challenges of dynamic factor analysis: time-varying patterns of sparsity, unknown number of factors, and identifiability constraints. By deploying dynamic sparsity, we successfully recover interpretable latent structures that automatically select the number of factors and that incorporate time-varying loadings/factors. We successfully applied our

methodology on a nontrivial simulated example as well as a real dataset comprising of 127 U.S. macroeconomic indices tracked over the period of the Great Recession (and beyond) and obtained several interpretable findings.

Our methodology can be enriched/extended in many ways. One possible extension would be to develop a latent variable method that can capture within, as well as between, connectivity of several high-dimensional time series. This could be achieved with a dynamic extension of sparse canonical correlation analysis [Witten et al., 2009]. By examining two large sets of data using this approach, one would be able to uncover the latent structure among and across multiple groups, a topic that has garnered increased interest after the Great Recession with questions regarding the dynamic change in inter-connectivity between the stock market and the macroeconomy. Our method can also be embedded within FAVAR models [Bernanke et al., 2005] that include both observed and unobserved predictors. Additionally, throughout our analysis we have assumed the covariance of the latent factors to be fixed over time and equal to an identity matrix, one could in principle incorporate dynamic variances with stochastic volatility modeling, along the lines of Zhou et al. [2014].

One possible shortcoming of our EM-based estimation strategy, is the lack of uncertainty assessment, which is essential for forecasting. The EM algorithm, however, was the key to obtaining interpretable latent structures for very high dimensional data. To achieve uncertainty quantification along with interpretability, one could impose structural identification constraints, such as Nakajima and West [2013,], and perform MCMC for DSS priors, as demonstrated in Appendix 5.9. Another approach would be to apply our method simply as a means of obtaining identifiability constraints (i.e. the sparsity pattern) and then reestimate the nonzero loadings with an MCMC strategy. While this would not quantify any sparsity-selection uncertainty, it would be an effective way to balance interpretability and forecasting/decision making. Another unavoidable feature of our method is its sensitivity to starting values. We strongly recommend using the output from the rolling window spike-and-slab factor model.

5.8 Appendix A

5.8.1 Derivation of the E-step

In this section we outline the steps of the parameter expanded EM algorithm. In the E-step, we compute the conditional expectation of the augmented and expanded log-posterior with respect to the missing data $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$, given observed data $\boldsymbol{Y}$ and the parameter values $\boldsymbol{\Delta}^{(m)}$ obtained at the previous M-step setting $\boldsymbol{A}_t = \boldsymbol{I}_K$. The conditional expected log likelihood $\mathcal{E} = \mathbb{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega}\,|\,\boldsymbol{Y},\Delta^{(m)}}[\log \pi(\boldsymbol{B}^{\star}_{0:T}, \boldsymbol{\Sigma}_{1:T}, \boldsymbol{A}_{1:T}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}\,|\,\boldsymbol{Y})]$ can be split into

$$\mathcal{E} = Q_1(\boldsymbol{B}^{\star}_{0:T}\,|\,\boldsymbol{\Sigma}_{1:T}) + Q_2(\boldsymbol{\Sigma}_{1:T}) + Q_3(\boldsymbol{A}_{1:T}) + C, \tag{5.12}$$

where C is a constant and $\boldsymbol{B}^{\star}_{0:T}$ denotes the factor loading matrices in the expanded space.

To simplify the expression for (5.12), we first define the retrospective and prospective penalty functions (as introduced in [Rockova et al., 2020]).

Definition 5.8.1. *For a given set of parameters $(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1)$, we define a prospective penalty function implied by* (5.3) *and* (5.6) *as follows:*

$$pen(\beta\,|\,\beta_{t-1}) = \log\left\{[1-\theta_t(\beta_{t-1})]\,\psi_0(\beta\,|\,\lambda_0) + \theta_t(\beta_{t-1})\,\psi_1(\beta\,|\,\mu_t(\beta_{t-1},\lambda_1)\right\}. \tag{5.13}$$

Similarly, we define a retrospective penalty $pen(\beta_{t+1}\,|\beta)$ as a function of the second argument β in (5.13)*. The Dynamic Spike-and-Slab (DSS) penalty is then defined as*

$$Pen(\beta\,|\,\beta_{t-1},\beta_{t+1}) = pen(\beta\,|\,\beta_{t-1}) + pen(\beta_{t+1}\,|\,\beta) + C, \tag{5.14}$$

where $C \equiv -Pen(0\,|\,\beta_{t-1},\beta_{t+1})$ is a norming constant. For $t = T$, the DSS penalty is defined simply as $Pen(\beta_T|\beta_{T-1}) \equiv pen(\beta_T|\beta_{T-1})$.

The individual penalty terms are defined as follows:

Definition 5.8.2. *For all $j = 1, \ldots, P$, prospective penalty for β_j^0 is defined as*

$$pen(\beta_j^0|\beta_j^1) = \sum_{k=1}^{K} \Big[(1 - \langle\gamma_{jk}^0\rangle)|\beta_{jk}^0|\lambda_0 + \langle\gamma_{jk}^0\rangle(\beta_{jk}^0)^2(1-\phi^2)/2\lambda_1) + \langle\gamma_{jk}^1\rangle(\beta_{jk}^1 - \phi\beta_{jk}^0)^2/2\lambda_1\Big]$$

Similarly, prospective penalty for β_j^t is defined as

$$pen(\beta_j^t|\beta_j^{t+1}) = \sum_{k=1}^{K}[\langle\gamma_{jk}^{t+1}\rangle(\beta_{jk}^{t+1} - \phi\beta_{jk}^t)^2/2\lambda_1]$$

For all $t \geq 1$ and $j = 1, \ldots, P$, retrospective penalty for β_j^t is defined as

$$pen(\beta_j^t|\beta_j^{t-1}) = \sum_{k=1}^{K}[(1 - \langle\gamma_{jk}^t\rangle)|\beta_{jk}^t|\lambda_0 + \langle\gamma_{jk}^t\rangle(\beta_{jk}^t - \phi\beta_{jk}^{t-1})^2/2\lambda_1]$$

There is no retrospective penalty for the initial factor loadings β_{jk}^0.

Define $\boldsymbol{\omega}_{t|T} = \mathbb{E}_{\boldsymbol{\Omega}}[\boldsymbol{\omega}_t \mid \boldsymbol{Y}, \boldsymbol{\Delta}^{(m)}]$, $\boldsymbol{V}_{t|T} = \text{cov}[\boldsymbol{\omega}_t \mid \boldsymbol{Y}, \boldsymbol{\Delta}^{(m)}]$. The terms $\boldsymbol{\omega}_{t|T}$ and $\boldsymbol{V}_{t|T}$ represent the best linear estimator for $\boldsymbol{\omega}_t$ using all observations and the corresponding covariance matrix, respectively. With $\boldsymbol{V}_{t,t-1|T} = \text{cov}[\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t-1} \,|\boldsymbol{Y}, \boldsymbol{\Delta}^{(m)}]$ we denote the covariance matrix of $\boldsymbol{\omega}_t$ and $\boldsymbol{\omega}_{t-1}$ given the data $\boldsymbol{Y}$ and $\boldsymbol{\Delta}^{(m)}$. These quantities can be obtained from the Kalman Filter and Smoother Algorithm (Table A1).

For estimating the observational variances σ_{jt}^2, we use forward filtering backward smoothing (FFBS) algorithm A2 (Section 4.3.7 in [Prado and West, 2010]). For notational simplification, throughout this section we will denote the estimate of $\boldsymbol{\Sigma}_{1:T}$ obtained from FFBS, also by $\boldsymbol{\Sigma}_{1:T}$. Now Denote

$$\langle\gamma_{jk}^0\rangle = \frac{\Theta\psi_1(\beta_{jk}^0|0, \frac{\lambda_1}{1-\phi^2})}{\Theta\psi_1(\beta_{jk}^0|0, \frac{\lambda_1}{1-\phi^2}) + (1-\Theta)\psi_0(\beta_{jk}^0|0, \lambda_0)},$$

$$\langle\gamma_{jk}^t\rangle = \frac{\theta_{jk}^t\psi_1(\beta_{jk}^t|\phi\beta_{jk}^{t-1}, \lambda_1)}{\theta_{jk}^t\psi_1(\beta_{jk}^t|\phi\beta_{jk}^{t-1}, \lambda_1) + (1-\theta_{jk}^t)\psi_0(\beta_{jk}^t|0, \lambda_0)},$$

Algorithm: *Kalman Filter and Smoother*	
Initialize $\boldsymbol{\omega}_{0\,|\,0} = \mathbf{0}$ and $\boldsymbol{V}_{0\,|\,0} = 1/(1-\widehat{\phi}^2)\boldsymbol{I}_K$ Repeat the Prediction Step and Correction Step for $t = 1, \ldots, T$	
Prediction Step	$\boldsymbol{\omega}_{t\,|\,t-1} = \boldsymbol{\omega}_{t-1\,|\,t-1}$ $\boldsymbol{V}_{t\,|\,t-1} = \boldsymbol{V}_{t-1\,|\,t-1} + \boldsymbol{I}_K$
Correction Step	$\boldsymbol{K}_t = \boldsymbol{V}_{t\,|\,t-1}\boldsymbol{B}_t'(\boldsymbol{B}_t\boldsymbol{V}_{t\,|\,t-1}\boldsymbol{B}_t' + \boldsymbol{\Sigma}_t)^{-1}$ $\boldsymbol{\omega}_{t\,|\,t} = \boldsymbol{\omega}_{t\,|\,t-1} + \boldsymbol{K}_t(\boldsymbol{Y}_t - \boldsymbol{B}_t\boldsymbol{\omega}_{t\,|\,t-1})$ $\boldsymbol{V}_{t\,|\,t} = \boldsymbol{V}_{t\,|\,t-1} - \boldsymbol{K}_t\boldsymbol{B}_t\boldsymbol{V}_{t\,|\,t-1}$
Initialize $\boldsymbol{V}_{T,T-1\,|\,T} = (I - \boldsymbol{K}_T\boldsymbol{B}_T)\boldsymbol{V}_{T-1\,|\,T-1}$ Repeat the smoothing step for $t = T, \ldots, 1$	
Smoothing Step	$\boldsymbol{\omega}_{t-1\,|\,T} = \boldsymbol{\omega}_{t-1\,|\,t-1} + \boldsymbol{Z}_{t-1}(\boldsymbol{\omega}_{t\,|\,T} - \boldsymbol{\omega}_{t\,|\,t-1})$ $\boldsymbol{V}_{t-1\,|\,T} = \boldsymbol{V}_{t-1\,|\,t-1} + \boldsymbol{Z}_{t-1}(\boldsymbol{V}_{t\,|\,T} - \boldsymbol{V}_{t\,|\,t-1})\boldsymbol{Z}_{t-1}'$ $\boldsymbol{V}_{t,t-1\,|\,T} = \boldsymbol{V}_{t-1\,|\,t-1}\boldsymbol{Z}_{t-2}' + \boldsymbol{Z}_{t-1}(\boldsymbol{V}_{t,t-1\,|\,T} - \boldsymbol{V}_{t-1\,|\,t-1})\boldsymbol{Z}_{t-2}'$ where $\boldsymbol{Z}_{t-1} = \boldsymbol{V}_{t-1\,|\,t-1}\boldsymbol{V}_{t\,|\,t-1}^{-1}$

Table A1: Kalman Filter and Smoother Algorithm for Parameter Expanded EM using rotated loading matrices $\boldsymbol{B}_{1:T}$

The functions $Q_1(\cdot)$, $Q_2(\cdot)$ and $Q_3(\cdot)$ in (5.12) can be written as follows:

$$
\begin{aligned}
-Q_1(\boldsymbol{B}_{0:T}^{\star} \,|\, \boldsymbol{\Sigma}_{1:T}) =& C + \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{P}\log\sigma_{jt}^2 \\
&+ tr\left\{\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\Sigma}_t^{-1}\left[(\boldsymbol{Y}_t - \boldsymbol{B}_t^{\star}\boldsymbol{\omega}_{t|T})(\boldsymbol{Y}_t - \boldsymbol{B}_t^{\star}\boldsymbol{\omega}_{t|T})' + \boldsymbol{B}_t^{\star}\boldsymbol{V}_{t|T}\boldsymbol{B}_t^{\star\prime}\right]\right\} \\
&+ \sum_{j=1}^{P}\sum_{k=1}^{K}\left[\frac{\langle\gamma_{jk}^0\rangle(\beta_{jk}^{0*})^2}{2\lambda_1/(1-\phi^2)} + (1 - \langle\gamma_{jk}^0\rangle)|\beta_{jk}^{0*}|\lambda_0\right] \\
&+ \sum_{t=1}^{T}\sum_{j=1}^{P}\sum_{k=1}^{K}\left[\frac{\langle\gamma_{jk}^t\rangle(\beta_{jk}^{t*} - \phi\beta_{jk}^{t-1*})^2}{2\lambda_1} + (1 - \langle\gamma_{jk}^t\rangle)|\beta_{jk}^{t*}|\lambda_0\right],
\end{aligned}
$$

where the terms $\boldsymbol{\omega}_{t|T}$ and $\boldsymbol{V}_{t|T}$ are obtained from the Kalman filter and smoother algorithm given in Table A1. Next, write $Q_2(\boldsymbol{\Sigma}_{1:T})$ as

$$
-Q_2(\boldsymbol{\Sigma}_{1:T}) = \sum_{t=1}^{T-1}\sum_{j=1}^{P}\left[pen(\sigma_{jt}^2 \,|\, \sigma_{j(t-1)}^2) + pen(\sigma_{jt}^2 \,|\, \sigma_{j(t+1)}^2)\right] + \sum_{j=1}^{P} pen(\sigma_{jT}^2 \,|\, \sigma_{j(T-1)}^2)
$$

The terms $pen(\sigma_{jt}^2 \,|\, \sigma_{j(t+1)}^2))$ and $pen(\sigma_{jT}^2 \,|\, \sigma_{j(T-1)}^2)$ in the above expression represent the prospective and retrospective penalties corresponding to the idiosyncratic variance. These

terms are defined as

$$pen(\sigma^2_{jt} \,|\, \sigma^2_{j(t-1)}) = \left(\frac{\delta n_{t-1}}{2} - 1\right) \log \sigma^2_{jt} - \left(\frac{(1-\delta) n_{t-1}}{2} - 1\right) \log \left(1 - \frac{\delta \sigma^2_{j(t-1)}}{\sigma^2_{jt}}\right),$$

$$pen(\sigma^2_{jt} \,|\, \sigma^2_{j(t+1)}) = -\left(\frac{\delta n_{t}}{2} - 1\right) \log \sigma^2_{jt} + \left(\frac{(1-\delta) n_{t}}{2} - 1\right) \log \left(1 - \frac{\delta \sigma^2_{jt}}{\sigma^2_{j(t+1)}}\right),$$

Finally, the term $Q_3(\boldsymbol{A}_{1:T})$ can be written as

$$-Q_3(\boldsymbol{A}_{1:T}) = \frac{1}{2} \sum_{t=1}^{T} \log |\boldsymbol{A}_t| + \frac{1}{2} tr\{\boldsymbol{A}_t^{-1} (\boldsymbol{M}_{1t} - \boldsymbol{M}_{12t} - \boldsymbol{M}'_{12t} + \boldsymbol{M}_{2t})\},$$

where

$$\begin{aligned}
\boldsymbol{M}_{1t} &= (\boldsymbol{\omega}_{t-1\,|\,T} \boldsymbol{\omega}'_{t-1\,|\,T} + \boldsymbol{V}_{t-1\,|\,T}), \\
\boldsymbol{M}_{12t} &= (\boldsymbol{\omega}_{t-1\,|\,T} \boldsymbol{\omega}'_{t\,|\,T} + \boldsymbol{V}_{t,t-1\,|\,T}), \\
\boldsymbol{M}_{2t} &= (\boldsymbol{\omega}_{t\,|\,T} \boldsymbol{\omega}'_{t\,|\,T} + \boldsymbol{V}_{t\,|\,T}).
\end{aligned}$$

5.8.2 Derivation of the M-step

In the M-step, we optimize the function $Q_1(\cdot)$ with respect to $\boldsymbol{B}^{\star}_{0:T}$, given values of $\boldsymbol{\Sigma}_{1:T}$ from the previous M-step. Given the new values $\boldsymbol{B}^{\star(m+1)}_{0:T}$ and the posterior moment estimates of the latent factors obtained from the Kalman filter, we optimize $Q_1(\cdot) + Q_2(\cdot)$, with respect to $\boldsymbol{\Sigma}_{1:T}$. Finally, we optimize the function $Q_3(\cdot)$ with respect to $\boldsymbol{A}_{1:T}$.

Optimizing $Q_1(\cdot)$ with respect to $\boldsymbol{B}^{\star}_{0:T}$ boils down to solving a series of independent dynamic spike and slab LASSO regressions. This is justified by the following lemma.

Lemma 5.8.1. *Let* $\boldsymbol{Y}^t = (Y^t_1, \ldots, Y^t_P)' \in \mathbb{R}^P$ *denote the snapshot of the series at time* t *and for* $1 \leq j \leq P$ *define a zero-augmented response vector at time* t *with* $\widetilde{\boldsymbol{Y}}^t_j =$

$(Y_j^t, \underbrace{0, \ldots, 0}_{K})' \in \mathbb{R}^{K+1}$. *For the SVD decomposition* $\boldsymbol{V}_{t\,|\,T} = \sum_{k=1}^{K} s_k \boldsymbol{U}_k^t (\boldsymbol{U}_k^t)'$, *we denote with* $\widetilde{\boldsymbol{U}}_k^t = \sqrt{s_k}\boldsymbol{U}_k^t$ *and with* $\boldsymbol{\Omega}^t = [\boldsymbol{\omega}_{t|T}, \widetilde{\boldsymbol{U}}_1^t, \ldots, \widetilde{\boldsymbol{U}}_K^t]' \in \mathbb{R}^{(1+K)\times K}$ *and we let* $\boldsymbol{\beta}_j^{t\star\prime} \in \mathbb{R}^K$ *be the* j^{th} *row of* $\boldsymbol{B}_t^\star$. *Then we can decompose*

$$Q_1(\boldsymbol{B}_{0:T}^\star \,|\, \boldsymbol{\Sigma}_{1:T}) = C + \sum_{j=1}^{P} \left[Q_j(\boldsymbol{\beta}_j^{t\star}) + Q^0(\boldsymbol{\beta}_j^{0\star}) + \widetilde{Q}(\boldsymbol{\beta}_j^{1\star}, \ldots, \boldsymbol{\beta}_j^{T\star}) \right],$$

where

$$\begin{aligned}
Q^0(\boldsymbol{\beta}_j^{0\star}) &= \sum_{k=1}^{K} \left[\frac{\langle \gamma_{jk}^0 \rangle (\beta_{jk}^{0*})^2}{2\lambda_1/(1-\phi^2)} + (1 - \langle \gamma_{jk}^0 \rangle) |\beta_{jk}^{0*}| \lambda_0 \right] \\
Q_j(\boldsymbol{\beta}_j^{t\star}) &= \sum_{t=1}^{T} \left[\frac{1}{2} \log \sigma_{jt}^2 + \frac{1}{2\sigma_{jt}^2} ||\widetilde{\boldsymbol{Y}}_j^t - \boldsymbol{\Omega}^t \boldsymbol{\beta}_j^{t\star}||_2^2 \right] \\
\widetilde{Q}(\boldsymbol{\beta}_j^{1\star}, \ldots, \boldsymbol{\beta}_j^{T\star}) &= \sum_{t=1}^{T} \sum_{k=1}^{K} \left[\frac{\langle \gamma_{jk}^t \rangle (\beta_{jk}^{t*} - \phi \beta_{jk}^{t-1*})^2}{2\lambda_1} + (1 - \langle \gamma_{jk}^t \rangle) |\beta_{jk}^{t*}| \lambda_0 \right].
\end{aligned}$$

Proof. Denote with

$$L \equiv tr \left\{ \frac{1}{2} \sum_{t=1}^{T} \boldsymbol{\Sigma}_t^{-1} \left[(\boldsymbol{Y}_t - \boldsymbol{B}_t^\star \boldsymbol{\omega}_{t|T})(\boldsymbol{Y}_t - \boldsymbol{B}_t^\star \boldsymbol{\omega}_{t|T})' + \boldsymbol{B}_t^\star \boldsymbol{V}_{t|T} \boldsymbol{B}_t^{\star\prime} \right] \right\}.$$

where the terms $\boldsymbol{\omega}_{t|T}$ and $\boldsymbol{V}_{t|T}$ are obtained from the Kalman filter and smoother algorithm given in Table A1.

Since by singular value decomposition, $\boldsymbol{B}_t^\star \boldsymbol{V}_{t\,|\,T} \boldsymbol{B}_t^{\star\prime} = \boldsymbol{B}_t^\star \sum_{k=1}^{K} s_k \boldsymbol{U}_k^t \boldsymbol{U}_k^{t\prime} (\boldsymbol{B}_t^\star)' = \sum_{k=1}^{K} (\boldsymbol{0} - \boldsymbol{B}_t^\star \widetilde{\boldsymbol{U}}_k^t)(\boldsymbol{0} - \boldsymbol{B}_t^\star \widetilde{\boldsymbol{U}}_k^t)'$, we can write

$$tr \left\{ \boldsymbol{\Sigma}_t^{-1} \boldsymbol{B}_t^\star \boldsymbol{V}_{t\,|\,T} \boldsymbol{B}_t^{\star\prime} \right\} = \sum_{k=1}^{K} (\boldsymbol{0} - \boldsymbol{B}_t^\star \widetilde{\boldsymbol{U}}_k^t)' \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{0} - \boldsymbol{B}_t^\star \widetilde{\boldsymbol{U}}_k^t).$$

Since $\boldsymbol{\Sigma}_t = \text{diag}(\sigma_{1t}^2, \ldots, \sigma_{Pt}^2)$, the RHS can be written as a sum of $P \times T$ quantities. Thus

we can write

$$
\begin{aligned}
L &= \frac{1}{2}\sum_{j=1}^{P}\sum_{t=1}^{T}\left[\frac{(Y_j^t - \boldsymbol{\omega}'_{t\,|\,T}\boldsymbol{\beta}_j^{t\star})^2}{\sigma_{jt}^2} + \sum_{k=1}^{K}\frac{(0 - \widetilde{\boldsymbol{U}}_k^{t\prime}\boldsymbol{\beta}_j^{t\star})^2}{\sigma_{jt}^2}\right] \\
&= \sum_{t=1}^{T}\sum_{j=1}^{P}\frac{1}{\sigma_{jt}^2}||(Y_j^t, 0, \ldots, 0)' - (\boldsymbol{\omega}'_{t\,|\,T}\boldsymbol{\beta}_j^{t\star}, \widetilde{\boldsymbol{U}}_1^{t\prime}\boldsymbol{\beta}_j^{t\star}, \ldots, \widetilde{\boldsymbol{U}}_K^{t\prime}\boldsymbol{\beta}_j^{t\star})'||^2 \\
&= \sum_{j=1}^{P}\sum_{t=1}^{T}\frac{1}{2\sigma_{jt}^2}||\widetilde{\boldsymbol{Y}}_j^t - \boldsymbol{\Omega}^t\boldsymbol{\beta}_j^{t\star}||_2^2. \qquad \square
\end{aligned}
$$

Each summand corresponds to a penalized dynamic regression with $K+1$ observations at each time t. Given $\boldsymbol{\Sigma}_t$, finding $\boldsymbol{B}^{\star(m+1)}$ thereby reduces to solving these J individual regressions. As shown in Rockova et al. [2020], each regression can be decomposed into a sequence of univariate optimization problems. We use the one-step late EM variant in Rockova et al. [2020] to obtain closed form one-site updates for each $\beta_{jk}^{\star t}$ for (j,k,t). Note that this corresponds to a generalized EM, which is aimed at improving the objective relative to the last iteration (not necessarily maximizing it).

These univariate updates are slightly different from Rockova et al. [2020], because we now have $K+1$ observations at time t, not just one. Denote with $\widehat{\beta}_{jl}^{*t}$ the most recent update of the coefficient $\beta_{jl}^{\star t}$. Let

$$
z_{jk}^t = \frac{1}{\sigma_{jt}^2}\sum_{r=1}^{K+1}(\widetilde{Y}_{jr}^t - \sum_{l\neq k}\widetilde{\omega}_{rl}^t\widehat{\beta}_{jl}^{t*})\widetilde{\omega}_{rk}^t
$$

and denote

$$
Z_{jk}^t = z_{jk}^t + \frac{\langle\gamma_{jk}^t\rangle\phi_1}{\lambda_1}\widehat{\beta}_{jk}^{t-1} + \frac{\langle\gamma_{jk}^{t+1}\rangle\phi_1}{\lambda_1}\widehat{\beta}_{jk}^{t+1}
$$

and

$$
W_{jk}^t = \frac{1}{\sigma_{jt}^2}\sum_{r=1}^{K+1}(\widetilde{\omega}_{rk}^t)^2 + \frac{\langle\gamma_{jk}^t\rangle}{\lambda_1} + \frac{\langle\gamma_{jk}^{t+1}\rangle\phi_1^2}{\lambda_1}.
$$

Then from the calculations in Section 5 of Rockova et al. [2020] (equations (5.16)-(5.18))

we obtain the following update for $\widehat{\beta}^{*t}_{jk}$:

$$\beta^{t\star(m+1)}_{jk} = \begin{cases} \frac{1}{W^t_{jk}+(1-\phi_1^2)/\lambda_1 M^t_{jk}}[Z^t_{jk} - \Lambda^t_{jk}]_+ \text{sign}(Z^t_{jk}) & \text{for} \quad 1 < t < T \\ \frac{1}{\langle\gamma^1_{jk}\rangle\phi_1^2+\langle\gamma^0_{jk}\rangle(1-\phi_1^2)}[\langle\gamma^0_{jk}\rangle\widehat{\beta}^1_{jk}\phi_1 - (1-\langle\gamma^0_{jk}\rangle)\lambda_0\lambda_1]_+ \text{sign}(\widehat{\beta}^1_{jk}) & \text{for} \quad t = 0 \end{cases} \tag{5.15}$$

where $M^t_{jk} = \langle\gamma^{t+1}_{jk}\rangle(1-\theta^{t+1}_{jk}) - (1-\langle\gamma^{t+1}_{jk}\rangle)\theta^{t+1}_{jk}$ and $\Lambda^t_{jk} = \lambda_0[(1-\langle\gamma^t_{jk}\rangle) - M^t_{jk}]$.

Given $\boldsymbol{B}^{\star(m+1)}$, optimizing $Q_1(\cdot)+Q_2(\cdot)$ with respect to $\boldsymbol{\Sigma}_{1:T}$ is done using the Forward Filtering Backward Smoothing algorithm [Ch. 4.3.7 Prado and West, 2010]. In order to maximize the posterior log likelihood with respect to $\boldsymbol{\Sigma}_{1:T}$, we first estimate the parameters of the posterior distribution $\pi(\boldsymbol{\Sigma}_{1:T} \mid \Omega, \boldsymbol{Y})$, given the updated factor loading matrices $B_{1:T}$, and then calculate the mode of the posterior. Although the exact analytical posterior is unattainable, a fast Gamma approximation exists [Ch. 10.8 West and Harrison, 1997]. Appropriate Gamma approximations to the posterior have the form

$$\pi(1/\sigma^2_{j,T-k} \mid \Omega, \boldsymbol{Y}) = \boldsymbol{G}[\eta_{jT}(-k)/2, d_{jT}(-k)/2],$$

where $d_{jT}(-k) = \eta_{jT}(-k)s_{jT}(-k)$, with

$$s_{jT}(-k)^{-1} = (1-\delta)s^{-1}_{j,T-k} + \delta s_{jT}(-k+1)^{-1}$$

, and filtered degrees of freedom defined by

$$\eta_{jT}(-k) = (1-\delta)\eta_{j,T-k} + \delta\eta_{j,T-k+1},$$

initialized at $\eta_{jT}(0) = \eta_{jT}$. Here $s_{j,T-k}$ denotes $\mathbb{E}(\sigma^2_{j,T-k} \mid \Omega_{T-k}, \boldsymbol{Y}_{T-k})$. The details of the algorithm is given in Algorithm A2. In the algorithm we denote the diagonal matrices with diagonal entries $\eta_{j,T-k}$ by $\boldsymbol{\eta}_{T-k}$ and analogously define matrices $\boldsymbol{D}_T(-k)$, $\boldsymbol{S}_{T-k}$ and $S_T(-k)$ for $k = 0, 1, \ldots, T-1$ so that we can update the parameters of the posterior

Algorithm: *Forward Filtering Backward Smoothing*	
Input: $\boldsymbol{B}_{1:T}$ and $\boldsymbol{\Sigma}_{1:T}$ from previous iteration Initialize $\boldsymbol{\eta}_0$, $\boldsymbol{D}_0$, $\boldsymbol{S}_0 = \boldsymbol{D}_0\boldsymbol{\eta}_0^{-1}$ Repeat the Forward Step for $t = 1, \ldots, T$	
Forward Step	$\boldsymbol{\eta}_t = \delta\boldsymbol{\eta}_{t-1} + \mathrm{I}$
	$\boldsymbol{D}_t = \delta\boldsymbol{D}_{t-1} + \boldsymbol{S}_{t-1}\boldsymbol{E}_t\boldsymbol{E}_t'\boldsymbol{Q}_t^{-1}$
	$\boldsymbol{S}_t = \boldsymbol{D}_t\boldsymbol{\eta}_t^{-1}$
where	$\boldsymbol{E}_t = \boldsymbol{Y}_t - \boldsymbol{B}_t\boldsymbol{\omega}_{t\,|\,t-1}$
	$\boldsymbol{Q}_t = \boldsymbol{B}_t'\boldsymbol{V}_{t\,|\,t-1}\boldsymbol{B}_t + \boldsymbol{\Sigma}_t$
Initialize $\boldsymbol{S}_T(0) = S_T$ Repeat the Backward Step for $k = 1, \ldots, T-1$	
Backward Step	$\boldsymbol{\eta}_T(-k) = (1-\delta)\boldsymbol{\eta}_{T-k} + \delta\boldsymbol{\eta}_{T-k+1}$
	$\boldsymbol{S}_T(-k)^{-1} = (1-\delta)\boldsymbol{S}_{T-k}^{-1} + \delta\boldsymbol{S}_T(-k+1)^{-1}$
	$\boldsymbol{D}_T(-k) = \boldsymbol{\eta}_T(-k)\boldsymbol{S}_T(-k)$
	$\boldsymbol{\Upsilon}_{T-k} = (\boldsymbol{\eta}_T(-k) - \mathrm{I})\boldsymbol{D}_T(-k)^{-1}$
Compute Mode	$\boldsymbol{\Sigma}_{T-k} = \boldsymbol{\Upsilon}_{T-k}^{-1}$

Table A2: Forward Filtering Backward Smoothing algorithm for estimating idiosyncratic variances.

distribution simultaneously for all j and fixed t. In our study, we set the prior degrees of freedom η_0 to its limit $\eta_0 = (1-\delta)^{-1}$ in order to achieve stability and efficiency. Given the parameters of the posterior distribution (the expectation and degrees of freedom), computing the posterior mode is straight forward.

Finally, the updates for the covariance matrices $\boldsymbol{A}_{1:T}$, obtained by maximizing $Q_3(\cdot)$, have the following closed form

$$\boldsymbol{A}_t^{(m+1)} = \boldsymbol{M}_{1t} - \boldsymbol{M}_{12t} - \boldsymbol{M}_{12t}' + \boldsymbol{M}_{2t} \quad for \quad t = 1, \ldots, T.$$

After completing the expanded M-step in the $(m+1)^{st}$ iteration, we perform a rotation step towards the reduced parameter space to obtain

$$\boldsymbol{B}_t{}^{(m+1)} = \boldsymbol{B}_t^{\star(m+1)}\boldsymbol{A}_{tL}{}^{(m+1)},$$

where $\boldsymbol{A}_t{}^{(m+1)} = \boldsymbol{A}_{tL}^{(m+1)}\boldsymbol{A}_{tL}{}^{(m+1)'}$ is the Cholesky decomposition. These rotated factor loading matrices are carried forward to the next E-step, where we again use the reduced parameter form by keeping $\boldsymbol{A}_t = \boldsymbol{I}_K$.

Algorithm: *MCMC algorithm for DSS with a Gaussian spike*	
	Initialize $(\boldsymbol{B}_{0:T}, \boldsymbol{\Sigma}_{0:T})$ and choose n_0, d_0.
	Sampling Latent Factors
Kalman Filter and Smoother	Compute $\boldsymbol{\omega}_{t\,|\,T}, \boldsymbol{V}_{t\,|\,T}$ and $\boldsymbol{V}_{t,t-1\,|\,T}$ for $1 \leq t \leq T$
	Sample $\boldsymbol{\omega}_t \sim \mathcal{N}_K(\boldsymbol{\omega}_{t\,|\,T}, \boldsymbol{V}_{t\,|\,T})$
	Sampling Factor Loadings
Forward filtering	For $t = 1, \ldots, T$ and $j = 1, \ldots, P$,
	Compute $\boldsymbol{a}_j^t = \boldsymbol{H}_j^t + \boldsymbol{\Gamma}_{j.}^t(\boldsymbol{m}_j^{t-1} - \boldsymbol{H}_j^t)$.
	Compute $\boldsymbol{R}_j^t = \boldsymbol{\Gamma}_j^t \boldsymbol{C}_j^{t-1} \boldsymbol{\Gamma}_{j.}^{t\,\prime} + \boldsymbol{W}_j^t$.
	Compute $f_j^t = \boldsymbol{\omega}'_{t\,|\,T} \boldsymbol{a}_j^t$.
	Compute $q_j^t = \boldsymbol{\omega}'_{t\,|\,T} \boldsymbol{R}_j^t \boldsymbol{\omega}_{t\,|\,T} + \sigma_{jt}^2$ and $e_j^t = y_j^t - f_j^t$.
	Compute $\boldsymbol{m}_j^t = \boldsymbol{a}_j^t + \boldsymbol{A}_j^t e_j^t$ and $\boldsymbol{C}_j^t = \boldsymbol{R}_j^t - \boldsymbol{A}_j^t \boldsymbol{A}_j^{t\prime} q_j^t$ with $\boldsymbol{A}_j^t = \boldsymbol{R}_j^t \boldsymbol{\omega}_{t\,|\,T}/q_j^t$.
Backward sampling	Simulate $\boldsymbol{B}_{j.}^T \sim \mathcal{N}(\boldsymbol{m}_j^T, \boldsymbol{C}_j^T)$.
	For $t = T-1, \ldots, 0$ and $j = 1, \ldots, P$
	Compute $\boldsymbol{a}_j^T(t-T) = \boldsymbol{m}_j^t + \boldsymbol{L}_j^t[\boldsymbol{B}_{j.}^{(t+1)\star} - \boldsymbol{a}_j^{t+1}]$.
	Compute $\boldsymbol{R}_j^T(t-T) = \boldsymbol{C}_j^t - \boldsymbol{L}_j^t \boldsymbol{R}_j^{t+1} \boldsymbol{L}_j^{t\prime}$, where $\boldsymbol{L}_j^t = \boldsymbol{C}_j^t \boldsymbol{\Gamma}_{j.}^{t+1\prime} \boldsymbol{R}_j^{t+1\,-1}$.
	Simulate $\boldsymbol{B}_{j.}^t \sim \mathcal{N}(\boldsymbol{a}_j^T(t-T), R_j^T(t-T))$.
	Sampling Indicators
	For $j = 1, \ldots, p$ and $k = 1, \ldots, K$
	Compute $\theta_{jk}^t = \theta(\beta_{jk}^{t-1})$ for $1 \leq t \leq T$ from (5.6).
	Compute ${p_{jk}^\star}^t = {p_{jk}^\star}^t(\beta_{jk}^t)$ for $1 \leq t \leq T$ from (5.8).
	Compute ${p_{jk}^\star}^0 = \theta(\beta_{jk}^0)$ from (5.6).
	Sample $\gamma_{jk}^t \sim$ Bernoulli$[{p_{jk}^\star}^t(\beta_{jk}^t)]$ for $0 \leq t \leq T$.
	Sampling Precisions $\nu_j^t = 1/(\sigma_j^t)^2$
	For $t = 1, \ldots, T$ and $j = 1, \ldots, P$
Forward filtering	Compute $n_j^t = \delta n_j^{t-1} + 1$ and $d_j^t = \delta d_j^{t-1} + (r_j^t)^2$, where $r_j^t = y_j^t - \boldsymbol{\omega}'_{t\,|\,T} \boldsymbol{\beta}_j^t$.
Backward sampling	Sample $\nu_j^T \sim G(n_j^T/2, d_j^T/2)$.
	For $t = 1, \ldots, T$
	Sample $\eta_j^{T-t} \sim G[(1-\delta) n_j^{T-t}/2, d_j^{T-t}/2]$.
	Set $1/(\sigma_j^{T-t})^2 = \eta_j^{T-t} + \delta/(\sigma_j^{T-t+1})^2$.

Table B1: An MCMC algorithm with *DSS* priors and a Gaussian spike. Note that $G(a, b)$ denotes a gamma distribution with a mean a/b.
$\boldsymbol{H}_j^t = \phi_0 \boldsymbol{\Gamma}_j^{t\prime}, \quad \boldsymbol{m}_j^0 = \phi_0 \gamma_j^0, \quad \boldsymbol{W}_j^t = diag\{\boldsymbol{\Gamma}_j^t \lambda_1 + (1 - \boldsymbol{\Gamma}_j^t)\lambda_0\}$

5.9 Appendix: B

5.9.1 MCMC on Simulated Examples

In this section, we apply our MCMC estimation procedure on a simulated example (using the lower-triangular identifiability constraint). First, we generate a single dataset with $P = 10$ responses, $K = 3$ candidate latent factors, and $T = 100$ time series observations. We choose the dimensionality of this example to be much smaller than our in EM implementation in Section 5.4 because the computational times of our MCMC sampling procedure are less favorable compared to our EM approach. Table B2 shows a comparison between computation

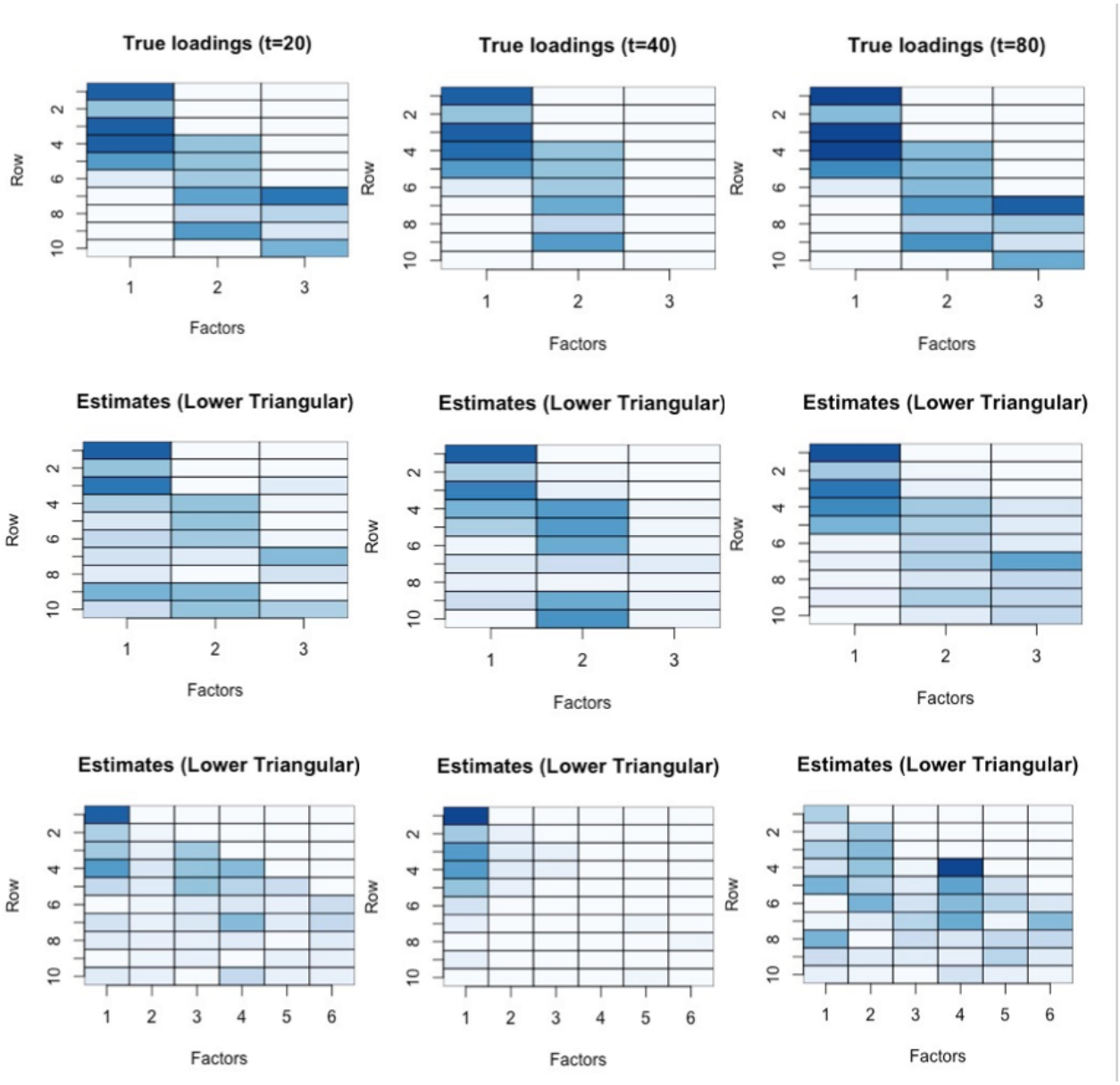


Figure B1: Simulated Example: Absolute Value of factor loading matrices estimated by MCMC (with 2000 draws after discarding 8000 burn-in samples) for a simulated data with $P = 10$ and $T = 100$. True number of nonzero factors are 3, 2 and 3 for $t = 20$, $t = 40$ and $t = 80$ respectively. (a) First row shows true factor loading matrices. (b) Second row shows estimated loading matrices with $K = 3$. (c) Third row shows estimated factor loading matrices with $K = 6$.

Method	K_{max}	Time (sec)
MCMC	3	2439.3
MCMC	4	2526.9
MCMC	6	2852.2
EM	6	302.5

Table B2: Comparison between elapsed computation time (in seconds) for MCMC (10000 draws) and EM algorithm. K_{max} denotes an upper bound on the number of factors, used in the implementation of MCMC and EM.

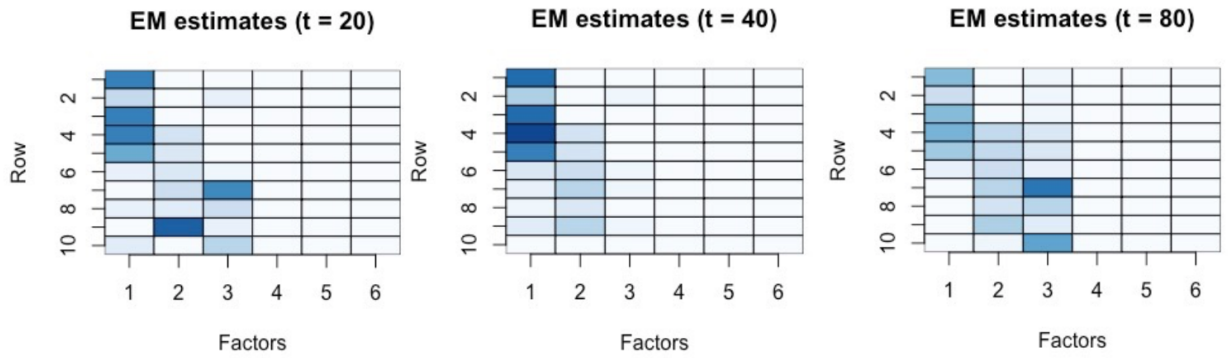


Figure B2: Simulated Example: Absolute Value of factor loading matrices estimated by EM algorithm with maximum number of factors $K = 6$.

times for MCMC and EM for this example.

The elements of latent factors $\boldsymbol{\Omega}_t$ and idiosyncratic errors $\boldsymbol{\epsilon}_t$ are generated from a standard Gaussian distribution. We now describe the true loading matrices $\boldsymbol{B}^0 = [\boldsymbol{B}_1^0, \ldots, \boldsymbol{B}_T^0]$, which were used to generate the data, where $\boldsymbol{B}_t^0 = \{\beta_{jk}^{0t}\} \in \mathbb{R}^{P \times K}$. At time $t = 1$, the active latent factor loadings form a block diagonal structure with 5, 6 and 4 active loadings for factor 1, factor 2 and factor 3 respectively, of which factor 1 and factor 2 overlap for 2 series, while factor 2 and factor 3 overlap for 3 series. Factor 1 and factor 3 never overlap.

The sparsity pattern changes structurally over time where (a) at time $t = 35$ the loadings of the third factor become inactive, (b) at $t = 45$ the loadings of the third factor are re-introduced and active until $T = 100$. The true nonzero loadings are smooth and arrive from an autoregressive process, i.e. $\beta_{jk}^{0t} = \phi\beta_{jk}^{0t-1} + v_{jk}^t$ with $v_{jk}^t \overset{iid}{\sim} \mathcal{N}(0, 0.0025)$ for $\phi = 0.99$, initiated at $\beta_{jk}^{01} \overset{iid}{\sim} \mathcal{N}(2, 1)$ for all $1 \leq j \leq P$ and $1 \leq k \leq 5$. When loadings β_{jk}^{0t} become

inactive, they are thresholded to zero. The true factor loadings are thereby smooth until they suddenly drop out and can re-emerge.

Figure B2 shows a heatmap of the (absolute values of) true and estimated factor loadings. Row 1 shows the true factor loading matrices for times $t = 20$, $t = 40$ and $t = 80$ respectively. Row 2 and 3 show estimated factor loadings when upper limit of latent factors are set to $K = 3$ and $K = 6$ respectively. Posterior mean estimates of factor loadings are obtained from an MCMC sample of size 2000, after discarding 8000 burn-in samples. We see that for $K = 6$, MCMC is not able to identify actual number of factors due to factor splitting. However estimated loading structures are close to the true loadings for $K = 3$.

Software

All empirical results presented in this chapter have been implemented using the opensource software `R`. All the relevant code are available in the following GitHub repository:

`https://github.com/Enakshi-Saha/Dynamic-Sparse-Factor-Analysis`

	Output and Income		Labor Market		Consumption and Orders		Orders and Inventories		Money and Credit		Interest rate and Exchange Rates		Prices		Stock Market
1	RPI	17	HWI	48	HOUST	58	DPCERA3M086SBEA	67	M1SL	81	FEDFUNDS	103	WPSFD49207	123	S&P 500
2	W875RX1	18	HWIURATIO	49	HOUSTNE	59	CMRMTSPLx	68	M2SL	82	CP3Mx	104	WPSFD49502	124	S&P: indust
3	INDPRO	19	CLF16OV	50	HOUSTMW	60	RETAILx	69	M2REAL	83	TB3MS	105	WPSID61	125	S&P div yield
4	IPFPNSS	20	CE16OV	51	HOUSTS	61	AMDMNOx	70	AMBSL	84	TB6MS	106	WPSID62	126	S&P PE ratio
5	IPFINAL	21	UNRATE	52	HOUSTW	62	ANDENOx	71	TOTRESNS	85	GS1	107	OILPRICEx	127	VXOCLSx
6	IPCONGD	22	UEMPMEAN	53	PERMIT	63	AMDMUOx	72	NONBORRES	86	GS5	108	PPICMM		
7	IPDCONGD	23	UEMPLT5	54	PERMITNE	64	BUSINVx	73	BUSLOANS	87	GS10	109	CPIAUCSL		
8	IPNCONGD	24	UEMP5TO14	55	PERMITMW	65	ISRATIOx	74	REALLN	88	AAA	110	CPIAPPSL		
9	IPBUSEQ	25	UEMP15OV	56	PERMITS	66	UMCSENTx	75	NONREVSL	89	BAA	111	CPITRNSL		
10	IPMAT	26	UEMP15T26	57	PERMITW			76	CONSPI	90	COMPAPFFx	112	CPIMEDSL		
11	IPDMAT	27	UEMP27OV					77	MZMSL	91	TB3SMFFM	113	CUSR0000SAC		
12	IPNMAT	28	CLAIMSx					78	DTCOLNVHFNM	92	TB6SMFFM	114	CUSR0000SAD		
13	IPMANSICS	29	PAYEMS					79	DTCTHFNM	93	T1YFFM	115	CUSR0000SAS		
14	IPB51222S	30	USGOOD					80	INVEST	94	T5YFFM	116	CPIULFSL		
15	IPFUELS	31	CES1021000001							95	T10YFFM	117	CUSR0000SA0L2		
16	CUMFNS	32	USCONS							96	AAAFFM	118	CUSR0000SA0L5		
		33	MANEMP							97	BAAFFM	119	PCEPI		
		34	DMANEMP							98	TWEXMMTH	120	DDURRG3M086SBEA		
		35	NDMANEMP							99	EXSZUSx	121	DNDGRG3M086SBEA		
		36	SRVPRD							100	EXJPUSx	122	DSERRG3M086SBEA		
		37	USTPU							101	EXUSUKx				
		38	USWTRADE							102	EXCAUSx				
		39	USTRADE												
		40	USFIRE												
		41	USGOVT												
		42	CES0600000007												
		43	AWOTMAN												
		44	AWHMAN												
		45	CES0600000008												
		46	CES2000000008												
		47	CES3000000008												

Table B3: Macroeconomic Study: The list of economic variables used in the study.

CHAPTER 6

DISCUSSION ON FUTURE WORKS

Our research focuses on developing flexible Bayesian tools for the analysis of large datasets and on providing theoretical justifications for their effectiveness. In the initial parts of this dissertation we have examined the posterior concentration properties of Bayesian Additive Regression Trees (BART) and its various incarnations, adapted to a wide range of high dimensional learning tasks including regression, classification, causal inference and survival analysis, thus demonstrating the adaptability of Bayesian trees and their ensembles to a wide range of practical problems. From a theoretical perspective, these results build the foundation for deriving uncertainty quantification statements, thus providing a very promising future research direction. From a practical perspective, these results also emphasize the flexibility of BART models, thus motivating us to explore the adaptability of Bayesian trees in novel empirical applications. For instance, we have demonstrated that BART exhibits superior predictive performance compared to existing models for discrete consumer choice data. BART provides a semi-parametric nonlinear tool for discrete choice modelling which are the centerpiece of numerous marketing applications. Till date such models have mostly relied on strong parametric assumptions, often built upon the assumption of linearity. BART has the potential of relaxing several stringent assumptions that plague most conventional discrete choice models used in Econometrics as well as various social and behavioral sciences. However for most applications involving choice decisions made by an active agent, besides prediction, another very important objective is counterfactual inference which tries to evaluate the potential impact of an unobserved intervection. Machine learning models such as BART are notorious for their lack of interpretability, which might curtail the scope of these models for practical applications. An important future research direction would be to develop strategies for inference using BART.

Another strand of our work has focused on dynamic factor analysis (DFA) for high-dimensional time series. We have developed a comprehensive Bayesian framework for sparse

factor analysis endowed with the following features: (1) time-varying factor dimensionality, (2) residual stochastic volatility and (3) dynamic sparsity priors on the factor loadings. Our new scalable EM implementation has extended the reach of Bayesian DFA to larger datasets. To highlight the efficacy and usefulness of our proposed method, we applied our model to a large-scale balanced panel of macroeconomic variables covering multiple facets of the US economy. Our model captures dynamic changes in the factor structure and dimensionality, in particular around the Great Recession of 2008. Future research in this line of work will focus on the development of factor models on dynamic networks. We are particularly interested in examining how spike and slab prior based dynamic factor analysis tools, as discussed in Chapter 4 can be adapted to community detection tasks [Gopalan and Blei, 2013, Abbe, 2017] in time varying social and biological networks. Another interesting avenue of research would be adapting our sparse factor analysis model to multiple response categories, both in static and dynamic setup. Traditionally these kinds of responses are transformed to unconstrained spaces before subjecting them to factor analysis. Another popular technique is to use an unsupervised analogue of a generalized linear model [Wedel and Kamakura, 2001, Collins et al., 2001]. Our goal would be to devise flexible factor analysis methods for multiple response categories, that require less calibration and/or exhibit robust performance, with minimal identifiability constraints, while allowing for non-trivial dependencies between data variables with different distributions.

REFERENCES

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Alan Agresti. On the extinction times of varying and random environment branching processes. *Journal of Applied Probability*, 12(1):39–46, 1975.

[3] Omar Aguilar and Mike West. *Bayesian dynamic factor models and variance matrix discounting for portfolio allocation.* Institute of Statistics and Decision Sciences, Duke University, 1998.

[4] Omar Aguilar and Mike West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

[5] Omar Aguilar, Gabriel Huerta, Raquel Prado, and Mike West. Bayesian inference on latent structure in time series. *Bayesian Statistics*, 6(1):1–16, 1998.

[6] Isabel P Albaladejo and M Teresa Díaz-Delfa. The effects of motivations to go to the country on rural accommodation choice: A hybrid discrete choice model. *Tourism Economics*, page 1354816620912062, 2020.

[7] Susan Athey, David Blei, Robert Donnelly, Francisco Ruiz, and Tobias Schmidt. Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. In *AEA Papers and Proceedings*, volume 108, pages 64–67, 2018.

[8] Nattapol Aunsri and Paponpat Taveeapiradeecharoen. A time-varying bayesian compressed vector autoregression for macroeconomic forecasting. *IEEE Access*, 8:192777–192786, 2020.

[9] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[10] Jushan Bai and Serena Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

[11] Marta Bańbura, Domenico Giannone, and Lucrezia Reichlin. Large bayesian vector auto regressions. *Journal of applied Econometrics*, 25(1):71–92, 2010.

[12] John J Bates. Measuring travel time values with a discrete choice model: a note. *The Economic Journal*, 97(386):493–498, 1987.

[13] Christiane Baumeister, Philip Liu, and Haroon Mumtaz. Changes in the transmission of monetary policy: Evidence from a time-varying factor-augmented VAR. *Working Paper No. 401, Bank of England.*

[14] Efraim Benmelech, Ralf R Meisenzahl, and Rodney Ramcharan. The real effects of liquidity during the financial crisis: Evidence from automobiles. *The Quarterly Journal of Economics*, 132(1):317–365, 2017.

[15] Ben S Bernanke, Jean Boivin, and Piotr Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of Economics*, 120(1):387–422, 2005.

[16] Wim Bernasco and Richard Block. Where offenders choose to attack: A discrete choice model of robberies in chicago. *Criminology*, 47(1):93–130, 2009.

[17] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[18] Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

[19] Gérard Biau and Luc Devroye. On the risk of estimates for block decreasing densities. *Journal of multivariate analysis*, 86(1):143–165, 2003.

[20] Gérard Biau, Luc Devroye, and Gäbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

[21] Angela Bitto and Sylvia Frühwirth-Schnatter. Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75–97, 2019.

[22] Justin Bleich, Adam Kapelner, Edward I George, and Shane T Jensen. Variable selection for bart: an application to gene regulation. *The Annals of Applied Statistics*, pages 1750–1781, 2014.

[23] L Brieman, J Friedman, R Olshen, and C Stone. Classification and regression trees. belmont (ca): Wadsworth. *Google Scholar*, 1984.

[24] Arthur F Burns and Wesley C Mitchell. *Measuring Business Cycles*. The National Bureau of Economic Research, 1947.

[25] Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

[26] Carlos M Carvalho, Hedibert F Lopes, and Omar Aguilar. Dynamic stock selection strategies: A structured factor model framework. *Bayesian Statistics*, 9:1–21, 2011.

[27] Ismaël Castillo et al. Lower bounds for posterior rates with gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

[28] Ismaël Castillo et al. Pólya tree posterior distributions on densities. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 2074–2102. Institut Henri Poincaré, 2017.

[29] Xu Cheng, Zhipeng Liao, and Frank Schorfheide. Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies*, 83(4):1511–1543, 2016.

[30] Pradeep K Chintagunta and Harikesh S Nair. Structural workshop paper—discrete-choice models of consumer demand in marketing. *Marketing Science*, 30(6):977–996, 2011.

[31] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

[32] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[33] Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. High-dimensional nonparametric monotone function estimation using bart. *arXiv preprint arXiv:1612.01619*, 2016.

[34] Gabriel Chodorow-Reich. Effects of unconventional monetary policy on financial institutions. Technical report, National Bureau of Economic Research, 2014.

[35] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.

[36] Financial Crisis Inquiry Commission. *The financial crisis inquiry report, authorized edition: Final report of the National Commission on the Causes of the Financial and Economic Crisis in the United States*. Public Affairs, 2011.

[37] Marc Coram, Steven P Lalley, et al. Consistency of bayes estimators of a binary regression function. *The Annals of Statistics*, 34(3):1233–1269, 2006.

[38] Dennis D Cox. An analysis of bayesian inference for nonparametric regression. *The Annals of Statistics*, pages 903–923, 1993.

[39] Yves Croissant and Maintainer Yves Croissant. Package 'mlogit', 2020.

[40] Yves Croissant and Maintainer Spencer Graves. Package 'ecdat'. 2020.

[41] Maurizio Daniele and Julie Schnaitmann. A regularized factor-augmented vector autoregressive model. *arXiv preprint arXiv:1912.06049*, 2019.

[42] Angus Deaton and John Muellbauer. *Economics and consumer behavior*. Cambridge university press, 1980.

[43] Marco Del Negro and Christopher Otrok. Dynamic factor models with time-varying parameters: measuring changes in international business cycles. *FRB of New York Staff Report No.326*, 2008.

[44] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[45] David GT Denison, Bani K Mallick, and Adrian FM Smith. A bayesian cart algorithm. *Biometrika*, 85(2):363–377, 1998.

[46] Sameer K Deshpande, Ray Bai, Cecilia Balocchi, and Jennifer E Starling. VC-BART: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*, 2020.

[47] Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

[48] Francis X Diebold and Marc Nerlove. The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *Journal of Applied Econometrics*, 4(1):1–21, 1989.

[49] Thomas A Domencich and Daniel McFadden. Urban travel demand-a behavioral analysis. Technical report, 1975.

[50] Rob Donnelly, Francisco R Ruiz, David Blei, and Susan Athey. Counterfactual inference for consumer choice across many product categories. *arXiv preprint arXiv:1906.02635*, 2019.

[51] David L Donoho et al. Cart and best-ortho-basis: a connection. *Annals of statistics*, 25(5):1870–1911, 1997.

[52] Junliang Du and Antonio R Linero. Interaction detection with Bayesian decision tree ensembles. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 108–117. PMLR, 2019.

[53] Meyer Dwass. The total progeny in a branching process and a related random walk. *Journal of Applied Probability*, 6(3):682–686, 1969.

[54] Anastasios Evgenidis, Dionisis Philippas, and Costas Siriopoulos. Heterogeneous effects in the international transmission of the us monetary policy: a factor-augmented var perspective. *Empirical Economics*, 56(5):1549–1579, 2019.

[55] Leland Farmer, Lawrence Schmidt, and Allan Timmermann. Pockets of predictability. *CEPR Discussion Paper No.DP12885*, 2018.

[56] Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202, 1994.

[57] Sylvia Frühwirth-Schnatter and H Lopes. Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical report, University of Chicago Booth School of Business, 2009.

[58] Sylvia Frühwirth-Schnatter and Hedibert Freitas Lopes. Sparse Bayesian Factor Analysis when the number of factors is unknown. *arXiv:1804.04231*, 2018.

[59] Sylvia Frühwirth-Schnatter and Helga Wagner. Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154(1):85–100, 2010.

[60] Chao Gao, Fang Han, and Cun-Hui Zhang. Minimax risk bounds for piecewise constant models. *arXiv preprint arXiv:1705.06386*, 2017.

[61] Antonio Gargano, Davide Pettenuzzo, and Allan Timmermann. Bond return predictability: Economic value and links to the macroeconomy. *Management Science*, 65 (2):508–540, 2019.

[62] Edward I George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81(394):437–445, 1986.

[63] Edward I George. Minimax multiple shrinkage estimation. *The Annals of Statistics*, pages 188–205, 1986.

[64] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[65] J. Geweke. The Dynamic Factor Analysis of Economic Time Series. *In: Aigner, D.J. and Goldberger, A.S., Eds., Latent Variables in Socio-Economic Models 1, North-Holland, Amsterdam.*, 1977.

[66] John Geweke and Guofu Zhou. Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587, 1996.

[67] Subhashis Ghosal, Aad Van Der Vaart, et al. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

[68] Joyee Ghosh and David B Dunson. Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009.

[69] Domenico Giannone, Michele Lenza, Daphne Momferatou, and Luca Onorante. Short-term inflation projections: A bayesian vector autoregressive approach. *International journal of forecasting*, 30(3):635–644, 2014.

[70] Domenico Giannone, Michele Lenza, and Giorgio E Primiceri. Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451, 2015.

[71] Thomas F Golob and Amelia C Regan. Trucking industry adoption of information technology: a multivariate discrete choice model. *Transportation Research Part C: Emerging Technologies*, 10(3):205–228, 2002.

[72] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[73] Louis Gordon and Richard A Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10(4):611–627, 1980.

[74] Dilan Görür, Frank Jäkel, and Carl Edward Rasmussen. A choice model with infinitely many latent features. In *Proceedings of the 23rd international conference on Machine learning*, pages 361–368, 2006.

[75] Robert B Gramacy and Herbert K H Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[76] Rajarshi Guhaniyogi and David B Dunson. Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015.

[77] André Hackbarth and Reinhard Madlener. Consumer preferences for alternative fuel vehicles: A discrete choice analysis. *Transportation Research Part D: Transport and Environment*, 25:5–17, 2013.

[78] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. 2017.

[79] Marc Hallin and Roman Liska. Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617, 2007.

[80] James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989.

[81] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

[82] W Michael Hanemann. Discrete/continuous models of consumer demand. *Econometrica: Journal of the Econometric Society*, pages 541–561, 1984.

[83] Jerry A Hausman and David A Wise. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica: Journal of the econometric society*, pages 403–426, 1978.

[84] Jingyu He, Saar Yalov, and P Richard Hahn. XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138, 2019.

[85] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[86] Florian Huber and Martin Feldkircher. Adaptive shrinkage in bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, 37(1):27–39, 2019.

[87] Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[88] Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani, Maintainer Trevor Hastie, and MASS Suggests. Package 'islr'. *R-Project. October*, 19, 2017.

[89] Seonghyun Jeong and Veronika Rockova. The art of BART: On flexibility of Bayesian forests. *arXiv preprint arXiv:2008.06620*, 2020.

[90] Leena Kalliovirta, Mika Meitz, and Pentti Saikkonen. A Gaussian mixture autoregressive model for univariate time series. *Journal of Time Series Analysis*, 36(2):247–266, 2015.

[91] Ayush Kanodia and Sakshi Ganeriwal. Deep consumer choice models.

[92] Gregor Kastner. Sparse bayesian time-varying covariance estimation in many dimensions. *Journal of Econometrics*, 210(1):98–115, 2019.

[93] Gregor Kastner and Florian Huber. Sparse bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, 2020.

[94] Gregor Kastner, Sylvia Frühwirth-Schnatter, and Hedibert Freitas Lopes. Efficient bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, 26(4):905–917, 2017.

[95] Kastner, G. Bayesian estimation of (sparse) latent factor stochastic volatility models. R package version 0.8.4, 2017.

[96] Sylvia Kaufmann and Simon Beyeler. Factor augmented var revisited-a sparse dynamic factor model approach. 2018.

[97] Sylvia Kaufmann and Christian Schumacher. Identifying relevant and irrelevant variables in sparse factor models. *Journal of Applied Econometrics*, 32(6):1123–1144, 2017.

[98] Sylvia Kaufmann and Christian Schumacher. Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification. *Journal of Econometrics*, 210(1):116–134, 2019.

[99] Michael P Keane. Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327, 1997.

[100] Michael P Keane et al. *Panel data discrete choice models of consumer demand*. Nuffield College, 2013.

[101] Junghun Kim, Hyunjoo Lee, and Jongsu Lee. Smartphone preferences and brand loyalty: A discrete choice model reflecting the reference point and peer effect. *Journal of Retailing and Consumer Services*, 52:101907, 2020.

[102] Gary Koop and Dimitris Korobilis. Large time-varying parameter vars. *Journal of Econometrics*, 177(2):185–198, 2013.

[103] Gary Koop, Dimitris Korobilis, et al. Bayesian multivariate time series methods for empirical Macroeconomics. *Foundations and Trends® in Econometrics*, 3(4):267–358, 2010.

[104] Gary Koop, Dimitris Korobilis, and Davide Pettenuzzo. Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1):135–154, 2019.

[105] Tom Kornstad and Thor O Thoresen. A discrete choice model for labor supply and childcare. *Journal of Population Economics*, 20(4):781–803, 2007.

[106] Dimitris Korobilis. Var forecasting using bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230, 2013.

[107] Rico Krueger, Michel Bierlaire, Ricardo A Daziano, Taha H Rashidi, and Prateek Bansal. On the usefulness of mixed logit models with unobserved inter-and intra-individual heterogeneity. 2020.

[108] Nikolas Kuschnig and Lukas Vashold. BVAR: bayesian vector autoregressions with hierarchical prior selection in R. 2019.

[109] Balaji Lakshminarayanan, Daniel Roy, and Yee Whye Teh. Top-down particle filtering for bayesian decision trees. In *International Conference on Machine Learning*, pages 280–288, 2013.

[110] Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. In *Advances in neural information processing systems*, pages 3140–3148, 2014.

[111] Antonio R Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.

[112] Antonio Ricardo Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *arXiv preprint arXiv:1707.09461*, 2017.

[113] C Liu, D B Rubin, and Y N Wu. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998.

[114] Chuanhai Liu, Donald B Rubin, and Ying Nian Wu. Parameter expansion to accelerate em: the px-em algorithm. *Biometrika*, 85(4):755–770, 1998.

[115] Jun S Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

[116] Yi Liu and Veronika Rockova. Variable selection via thompson sampling. *arXiv preprint arXiv:2007.00187*, 2020.

[117] Yi Liu, Veronika Rockova, and Yuexi Wang. Abc variable selection with bayesian forests. *arXiv preprint arXiv:1806.02304*, 2018.

[118] Larry Lockshin, Wade Jarvis, Fran**c**cois d'Hauteville, and Jean-Philippe Perrouty. Using simulations from discrete choice experiments to measure consumer sensitivity to brand, region, price, and awards in wine choice. *Food quality and preference*, 17(3-4): 166–178, 2006.

[119] Hedibert F Lopes, RE McCulloch, and Ruey Tsay. Cholesky stochastic volatility. *Unpublished Technical Report, University of Chicago, Booth Business School*, 2, 2010.

[120] Hedibert Freitas Lopes and Carlos Marinho Carvalho. Factor stochastic volatility with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference*, 137(10):3082–3091, 2007.

[121] Hedibert Freitas Lopes and Helio S Migon. Comovements and contagion in emergent markets: stock indexes volatilities. In *Case studies in Bayesian statistics*, pages 285–300. Springer, 2002.

[122] Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67, 2004.

[123] Michael W McCracken and Serena Ng. FRED-MD: a monthly database for Macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[124] Daniel McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974.

[125] Daniel McFadden. Economic choices. *American economic review*, 91(3):351–378, 2001.

[126] X-L Meng and David Van Dyk. Fast em-type implementations for mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):559–578, 1998.

[127] Xiao-Li Meng and David A Van Dyk. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320, 1999.

[128] Atif Mian and Amir Sufi. The consequences of mortgage credit expansion: Evidence from the US mortgage default crisis. *The Quarterly Journal of Economics*, 124(4): 1449–1496, 2009.

[129] Atif Mian and Amir Sufi. House prices, home equity-based borrowing, and the US household leverage crisis. *American Economic Review*, 101(5):2132–56, 2011.

[130] Atif Mian, Kamalesh Rao, and Amir Sufi. Household balance sheets, consumption, and the economic slump. *The Quarterly Journal of Economics*, 128(4):1687–1726, 2013.

[131] Silvia Miranda-Agrippino and Giovanni Ricco. The transmission of monetary policy shocks. 2018.

[132] Gemma Elyse Moran. Bayesian approaches for modeling variation. 2019.

[133] Edward R Morey, W Douglass Shaw, and Robert D Rowe. A discrete-choice model of recreational participation, site choice, and activity valuation when complete trip data are not available. *Journal of Environmental Economics and Management*, 20(2): 181–201, 1991.

[134] Taka Morikawa, Moshe Ben-Akiva, and Daniel McFadden. Discrete choice models incorporating revealed preferences and psychometric data. *Advances in Econometrics*, 16:29–56, 2002.

[135] Evan Munro. Deep learning models for restaurant choice. 2018.

[136] Jared S Murray. Log-linear bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, (just-accepted):1–35, 2020.

[137] Jouchi Nakajima and Mike West. Dynamic factor volatility modeling: A Bayesian latent threshold approach. *Journal of Financial Econometrics*, 11(1):116–153, 2013.

[138] Jouchi Nakajima and Mike West. Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164, 2013.

[139] Jouchi Nakajima, Mike West, et al. Dynamics & sparsity in latent threshold factor models: A study in multivariate EEG signal processing. *Brazilian Journal of Probability and Statistics*, 31(4):701–731, 2017.

[140] Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

[141] Alexei Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.

[142] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[143] Amil Petrin. Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy*, 110(4):705–729, 2002.

[144] Mark Pitt and Neil Shephard. Time varying covariances: a factor stochastic volatility approach. *Bayesian Statistics*, 6:547–570, 1999.

[145] Nicholas G Polson and Rockova. Posterior concentration for sparse deep learning.

[146] Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384, 1985.

[147] Galina Potjagailo. Spillover effects from euro area monetary policy across europe: A factor-augmented var approach. *Journal of International Money and Finance*, 72: 127–147, 2017.

[148] R. Prado and M. West. *Time Series: Modelling, Computation & Inference*. Chapman & Hall/CRC Press, 2010.

[149] Matthew Pratola, Hugh Chipman, Edward George, and Robert McCulloch. Heteroscedastic bart using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*, 2017.

[150] Matthew T Pratola, Hugh A Chipman, James R Gattiker, David M Higdon, Robert McCulloch, and William N Rust. Parallel Bayesian additive regression trees. *Journal of Computational and Graphical Statistics*, 23(3):830–852, 2014.

[151] Carmen M Reinhart and Kenneth S Rogoff. Is the 2007 US sub-prime financial crisis so different? An international historical comparison. *American Economic Review*, 98 (2):339–44, 2008.

[152] Brian Ripley, William Venables, and Maintainer Brian Ripley. Package 'nnet'. *R package version*, 7:3–12, 2016.

[153] Veronika Rockova. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437, 2018.

[154] Veronika Rockova and Edward I George. Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622, 2016.

[155] Veronika Rockova and Edward I George. The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113:431–444, 2018.

[156] Veronika Rockova and Enakshi Saha. On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848, 2019.

[157] Veronika Rockova, Kenichiro McAlinn, et al. Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis*, 2020.

[158] Veronika Rockova, Stéphanie van der Pas, et al. Posterior concentration for Bayesian regression trees and forests. *Annals of Statistics*, 48(4):2108–2131, 2020.

[159] Peter E Rossi, Greg M Allenby, and Rob McCulloch. *Bayesian statistics and marketing*. John Wiley & Sons, 2012.

[160] Daniel M Roy, Yee Whye Teh, et al. The mondrian process. In *NIPS*, pages 1377–1384, 2008.

[161] Donald B Rubin and Dorothy T Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[162] Francisco JR Ruiz, Susan Athey, David M Blei, et al. Shopper: A probabilistic model of consumer choice with substitutes and complements. *Annals of Applied Statistics*, 14 (1):1–27, 2020.

[163] Enakshi Saha. A theoretical analysis of Generalized BART models. *under review at the Journal of the American Statistical Association*, 2021.

[164] Enakshi Saha and Veronika Rockova. Dynamic sparse factor analysis. *Journal of Applied Econometrics (Revision submitted).*, 2021. The authors would like to acknowledge Dr. Kenichiro McAllin for his help in preparing an earlier version of the manuscript.

[165] Katharina Sammer and Rolf Wüstenhagen. The influence of eco-labelling on consumer behaviour–results of a discrete choice analysis for washing machines. *Business Strategy and the Environment*, 15(3):185–199, 2006.

[166] Thomas J Sargent, Christopher A Sims, et al. Business cycle modeling without pretending to have too much a priori economic theory. *New methods in Business cycle research*, 1:145–168, 1977.

[167] Takuya Satomura, Jaehwan Kim, and Greg M Allenby. Multiple-constraint choice models with corner and interior solutions. *Marketing Science*, 30(3):481–490, 2011.

[168] Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

[169] Brian Sifringer, Virginie Lurkin, and Alexandre Alahi. Enhancing discrete choice models with neural networks. In *Proceedings of the 18th Swiss Transport Research Conference (STRC), Monte Verità/Ascona, Switzerland*, pages 16–18, 2018.

[170] Kenneth A Small. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society*, pages 409–424, 1987.

[171] Adriaan R Soetevent and Peter Kooreman. A discrete-choice model with social interactions: with an application to high school teen behavior. *Journal of Applied Econometrics*, 22(3):599–624, 2007.

[172] Rodney Sparapani, Charles Spanbauer, and Robert McCulloch. The BART R package, 2019.

[173] Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in medicine*, 35(16):2741–2753, 2016.

[174] Jennifer E Starling, Jared S Murray, Carlos M Carvalho, Radek K Bukowski, James G Scott, et al. Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *Annals of Applied Statistics*, 14(1):28–50, 2020.

[175] Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

[176] James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460): 1167–1179, 2002.

[177] James H Stock and Mark W Watson. Implications of dynamic factor models for var analysis. Technical report, National Bureau of Economic Research, 2005.

[178] James H Stock and Mark W Watson. Modeling inflation after the crisis. Technical report, National Bureau of Economic Research, 2010.

[179] James H Stock and Mark W Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier, 2016.

[180] James H Stock and Mark W Watson. Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal*, 128 (610):917–948, 2018.

[181] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[182] Matthew A Taddy, Robert B Gramacy, and Nicholas G Polson. Dynamic trees for learning and design. *Journal of the American Statistical Association*, 106(493):109–123, 2011.

[183] Yaoyuan Vincent Tan, Carol AC Flannagan, and Michael R Elliott. Predicting human-driving behavior to help driverless vehicles drive: random intercept bayesian additive regression trees. *arXiv preprint arXiv:1609.07464*, 2016.

[184] Kenneth Train and Yves Croissant. Kenneth train's exercises using the mlogit package for r. *R*, 25:0–2, 2012.

[185] Truong P Truong and David A Hensher. Measurement of travel time values and opportunity cost from a discrete-choice model. *The Economic Journal*, pages 438–451, 1985.

[186] Stéphanie van der Pas and Veronika Rockova. Bayesian dyadic trees and histograms for regression. In *Advances in neural information processing systems*, pages 2089–2099, 2017.

[187] Aad W van der Vaart, J Harry van Zanten, et al. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[188] David A Van Dyk and Ruoxi Tang. The one-step-late pxem algorithm. *Statistics and Computing*, 13(2):137–152, 2003.

[189] Zulfiqar Ali Wagan, Zhang Chen, Hakimzadi Wagan, et al. A factor-augmented vector autoregressive approach to analyze the transmission of monetary policy. *Prague Economic Papers*, 28(6):709–728, 2019.

[190] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[191] Mengting Wan, Di Wang, Matt Goldman, Matt Taddy, Justin Rao, Jie Liu, Dimitrios Lymberopoulos, and Julian McAuley. Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1103–1112, 2017.

[192] Mark W Watson and Robert F Engle. Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3):385–400, 1983.

[193] Michel Wedel and Wagner A Kamakura. Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66(4):515–530, 2001.

[194] M. West. Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003.

[195] M. West and P. J. Harrison. *Bayesian Forecasting & Dynamic Models*. Springer Verlag, 2nd edition, 1997.

[196] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[197] Chun S Wong and Wai K Li. On a logistic mixture autoregressive model. *Biometrika*, 88(3):833–846, 2001.

[198] Chun Shan Wong and Wai Keung Li. On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):95–115, 2000.

[199] Sally Wood, Ori Rosen, and Robert Kohn. Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics*, 20(1):174–195, 2011.

[200] Ryo Yoshida and Mike West. Bayesian learning in sparse graphical factor models via variational mean-field annealing. *Journal of Machine Learning Research*, 11(May): 1771–1798, 2010.

[201] Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.

[202] Xiaocong Zhou, Jouchi Nakajima, and Mike West. Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting*, 30(4):963–980, 2014.

[203] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[204] Alain F Zuur, RJ Fryer, IT Jolliffe, R Dekker, and JJ Beukema. Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, 14 (7):665–685, 2003.