Large-Scale Quasi-Bayesian Inference With Spike-and-Slab Priors

by

Anwesha Bhattacharyya

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Yves Atchadé, Chair
Assistant Professor Yang Chen
Professor Xuming He
Associate Professor Jian Kang

Anwesha Bhattacharyya

anwebha@umich.edu

ORCID iD: 0000-0002-5960-1631

To Diptavo and Arkaprabha

ACKNOWLEDGEMENTS

This dissertation has been partially supported by NSF grant DMS 1513040.

I would first and foremost like to acknowledge Dr. Yves Atchadé, my advisor and mentor. Without his constant support, motivation and help this dissertation would not have been possible. I am immensely grateful for the patience he showed and encouragement he provided over these five years. Next, I would also like to thank Dr. Xuming He, my co-advisor, for his guidance, support and direction. I would also like to express my heartfelt thanks to my committee members Dr. Yang Chen and Dr. Jian Kang.

I take this opportunity to thank the faculty and the students of the Statistics department for the positivity, encouragement and academic discussions without which these five years would not have been as intellectually stimulating and fruitful as it has been. A special note of thanks to Dr. Shyamala Nagraj and Dr. Yang Chen with whom the experience as a graduate student instructor was memorable.

On a personal note I would like to thank the Indian community in the Statistics and Biostatistics department for making Ann Arbor a home away from home with special mention to Aritra Guha, Debarghya Mukherjee, Dr. Avijit Shea, Debraj Bose, Dr. Rounak Dey, Dr. Shrijita Bhattacharya and Dr. Bhramar Mukherjee.

I would also like to thank my friend Nandini Kundu and aunt Paulomi Bhattacharya who have acted as my support system for as long as I remember.

I would like to sincerely thank my father Dr. Gautam Bhattacharyya. It was his constant encouragement and support that instilled in me the spirit and courage to pursue my ambition. I am also extremely grateful to my mother Dr. Moushumi Bhattacharyya, and my mother-in-law Mrs Mukuta Dutta, both of whom have been extremely encouraging and supportive of my career and education. A special note of thanks to Gulchu, for whom my affection will remain boundless forever. This acknowledgment would be incomplete without mentioning my husband, my best friend Dr. Diptavo Dutta. I would like to take this opportunity to thank him for being my rock, my source of joy and for having confidence in me when I was at my worst.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation studies a general framework using spike-and-slab prior distributions to facilitate the development of high-dimensional Bayesian inference. Our framework allows inference with a general quasi-likelihood function to address scenarios where likelihood based inference are infeasible or the underlying optimization problems are not the same as the data generating mechanisms. We show that highly efficient and scalable Markov Chain Monte Carlo (MCMC) algorithms can be easily constructed to sample from the resulting quasi-posterior distributions. We study the large scale behavior of the resulting quasi-posterior distributions as the dimension of the parameter space grows, and we establish several convergence results. In large-scale applications where computational speed is important, variational approximation methods are often used to approximate posterior distributions. We show that the contraction behaviors of the quasi-posterior distributions can be exploited to provide theoretical guarantees for their variational approximations. We illustrate the theory with several examples. Finally we develop a quasi-likelihood based algorithm for estimation of Ising/Potts models that incorporates inbuilt mechanism for parallel computation. We illustrate the usability of the method by analyzing 16 Personality Factors data under the setup of Five-level Potts Model. The data analysis recovers known clusters of personality traits and also indicates plausible novel clusters.

CHAPTER I

Introduction

The current age of big data has ushered in unprecedented opportunities with regards to the topics that might come under the purview of data-driven decision making. Despite that, we are still faced with scenarios where the number of available observations are relatively limited in comparison to the number of variables under consideration, i.e. the dimensionality of the problem. These problems commonly arise in experimental setups where each response to be recorded comes at a certain price. For example, in clinical trials, sociological studies or psychometric studies, data on a large number of variables is collected for a relatively small number of subjects enrolled in the study. In such scenarios, incorporation of field knowledge or knowledge based on past studies can bolster the accuracy of the results of these studies especially when the sample size is relatively small. In Bayesian Statistics, the past knowledge is easily incorporated in the form of prior distributions. The prior, coupled with the likelihood of the data, gives us a posterior distribution on the parameters of interest. Bayesian inference has two built-in features that are in growing demand in the applications: a) the ability to incorporate existing knowledge in the form of a prior distribution (Greenfield et al. [2013], Studham et al. [2014], Peng et al. [2013]), and b) a simple mechanism for uncertainty quantification in the inference through the posterior distribution. The theoretical analysis of these posterior distributions in the high-dimensional setting has only recently begun (Martin et al. [2017], Castillo et al. [2015], Bhattacharya et al. [2015], Moreno et al. [2015], Rockova and George [2014], Narisetty and He [2014], Polson and Scott [2010]). This dissertation contributes to that literature.

However, with increasing complexity of problems at hand, we often face situations where inference is not based on the likelihood but some other non-likelihood (or quasi-likelihood) functions. This may be because the optimization criterion of interest is different from the data-generating likelihood function or simply because likelihood based inference is computationally intractable. Non-likelihood functions (also known as quasi-likelihood, pseudo-likelihood or composite likelihood functions) are routine in frequentist statistics, particularly to deal with large scale problems (Meinshausen and Buhlmann [2006], Zou et al. [2006], Shen and Huang [2008], Ravikumar et al. [2010], Varin et al. [2011], Lei and Vu [2015]). In semi/non-parametric statistics and econometrics, the idea is closely related to moments restrictions inference (Ichimura [1993], Chernozhukov et al. [2007]).

A Bayesian analog involves defining a quasi-likelihood function that replaces the likelihood in an otherwise standard Bayesian inference procedure. There is indeed an increasing Bayesian literature where non-likelihood functions are combined with prior distributions (Chernozhukov and Hong [2003], Jiang and Tanner [2008], Liao and Jiang [2011], Yang and He [2012], Kato [2013], Li and Jiang [2014], Atchade [2017], Atchadé [2019]).

A suitable example where the quasi-likelihood approach may be used, would be inference on model based Pairwise Markov Random Fields (PMRFs), such as Gaussian graphical models or in case of binary or finite ordinal data, Ising models and

Potts Models respectively (Atchade [2017], Atchadé [2019]). Most existing Bayesian methods for fitting graphical models do not scale well as the number of nodes in the graph grows, despite the recent progress with Gaussian graphical models (Dobra et al. [2011], Khondker et al. [2013], Peterson et al. [2015], Banerjee and Ghosal [2013]). The computational challenge only intensifies when dealing with discrete graphical models (Ising and Potts models). Indeed, a full Bayesian treatment of most discrete graphical models leads to the so-called doubly-intractable posterior distributions for which specialized MCMC algorithms are needed (Zhou and Schmidler [2009], Murray et al. [2006], Lyne et al. [2015]). However, these algorithms also do not scale well when dealing with large graphs. In the frequentist literature, there is a long history of fitting discrete graphical models using quasi/pseudo-likelihood methods instead of the full likelihood (the idea dates back at least to Besag [1974]; see also Guyon [1995]). In fact, quasi-likelihood methods have become the *de facto* approach in the frequentist literature when dealing with large graphical models (Meinshausen and Buhlmann [2006], Höfling and Tibshirani [2009], Ravikumar et al. [2010], Guo et al. [2015], Roy et al. [2017]). As shown for instance in (Atchadé [2019]) these quasi-likelihood functions can be used to fit Gaussian graphical models in the Bayesian framework at a scale unmatched by fully Bayesian alternatives. The crux of the method is the use of a product-form pseudo-likelihood function (as used in the frequentist literature) that makes it possible to split the resulting quasi-posterior distribution into a product of linear regression Bayesian posterior distributions. Significant reduction in computational costs can then be achieved by deploying this approach on a multi-core computer system.

Motivated by the increasingly widespread use of quasi-likelihood function in Bayesian literature, in this dissertation, we aim to develop a general Bayesian framework for inference based on quasi-likelihood function in high dimension under a spike and slab prior distribution. We focus on settings where the parameter of interest $\theta \in \mathbb{R}^p$ is sparse and the problem of variable selection is addressed by introducing an auxiliary selection parameter $\delta \in \{0,1\}^p$ that acts as the support of θ. Here $\delta_j = 1$ implies θ_j is active or included in the model. We then follow a well-established practice in the Bayesian literature that imposes a spike-and-slab prior distribution jointly on (θ, δ) with a Gaussian spike and Gaussian slab distribution (Mitchell and Beauchamp [1988], George and McCulloch [1997], Narisetty and He [2014]). More precisely, we actually follow here a computationally efficient version of the standard spike and slab prior of (George and McCulloch [1997]). We use an atypical sparsification trick on the quasi-likelihood which ensures that the resulting quasi-posterior distribution can be used to construct standard efficient MCMC algorithms that are scalable with increasing dimensionality. Without getting into details, we would like to note here briefly that by virtue of sparsification, the spike prior on the inactive components does not influence the marginal quasi-likelihood of the active components but it does affect the mixing time of the MCMC chains constructed and in that sense works similar to the pseudo-priors of (Carlin and Chib [1995]).

In the first chapter, we explore the theoretical properties of a general quasi-posterior distribution under growing dimensionality. To this end, we consider a general log-quasi-likelihood function ℓ and a random sample Z such that $\ell(\cdot; Z)$ is (locally) strongly concave with maximizer located near some parameter value of interest $\theta_\star \in \mathbb{R}^p$. The parameter value $\theta_\star$ is typically (but not necessarily) defined as the maximizer of the population version of the log-quasi-likelihood function:

$$\theta_\star = \underset{\theta \in \mathbb{R}^p}{\mathsf{Argmax}}\ \mathbb{E}_\star \left[\ell(\theta; Z)\right].$$

We proceed to study the contraction properties of the quasi-posterior for $p \to \infty$ and sample size n, growing with p. Under certain regularity conditions on the quasi-likelihood, we can show that with optimally chosen prior parameters, the quasi-posterior distribution is sparse in δ [Theorem 1]. We also show that the quasi-posterior puts most of its probability mass around $(\delta_\star, \theta_\star)$, where $\delta_\star$ is the support of $\theta_\star$ [Theorem 2]. We can also show model selection consistency where we show that given sufficient signal strength, the true model is always selected [Theorem 3].

For sufficiently strong signal $\theta_\star$, we also show that the quasi-posterior actually behaves like a product of a point mass at $\delta_\star$ and the Gaussian approximation of the conditional quasi-posterior distribution of θ given $\delta = \delta_\star$ (Bernstein-von Mises approximation [Theorem 4]). The results have implications for variational approximation methods, and as an application of the main results, we derive some sufficient conditions under which variational approximations of the quasi-posterior are consistent [Theorem 5]. We illustrate the theory with examples from linear regression (Section 2.5.1), Gaussian graphical models (Section 2.5.2), logistic regression (Section 2.5.3) and sparse principal component analysis(PCA) (Section 2.5.4).

In case of linear regression and Gaussian graphical models, we establish sparsity in the quasi-posterior distributions and derive contraction rates of the same. We have also established the Bernstein-Von-Mises phenomenon and bounds on variational approximation for the linear models and consequently on the Gaussian graphical models. We further studied the logistic regression as they form the building blocks of estimation of Ising models under a quasi-likelihood approach. For the logistic regression, the construction of the proofs poses difficulty in establishing posterior sparsity even though simulation results exhibit sparsity in the posterior. We have again established contraction rates, model selection consistency and the

Bernstein-Von-Mises phenomenon under certain assumptions that rely on the restricted eigenvalue conditions of the double derivative of the quasi-likelihood. It is worthwhile to note here that while the eigenvalue conditions are easy to verify for Gaussian graphical models, they are not as easily verifiable for logistic regression or sparse PCA problems, thus posing some limitations in checking the applicability of these results. We have also shown that by virtue of the results obtained for a general quasi-likelihood approach, we can provide a contraction rate for the quasi-posterior distribution obtained in the context of a sparse PCA example. However, any other results such as model selection consistency would require a better understanding about the distribution of singular vectors than we currently have. It should be noted that the the interpretation of the resulting quasi-posterior distribution is debatable particularly in the context of frequentist testing and coverage. However, by virtue of the established results, we can claim that the estimates will be close to the truth and any prediction or clustering based on these estimates will have good accuracy. The variance of the asymptotic distribution of the quasi-posterior distribution established in the dissertation, can be adjusted to provide valid frequentist coverage under specific models. This has been previously addressed in literature, specially in low dimension (Chernozhukov and Hong [2003], Yang and He [2012]) and our results indicate that similar techniques can be utilized to provide frequentist validity in high dimension as well. In the second chapter of this dissertation we have developed a scalable Bayesian algorithm for estimating Ising and Potts Models using a quasi-likelihood based approach. As mentioned before, Ising and Potts models are special cases of parametric Pairwise Markov Random Fields (PMRFs). PMRFs are characterized by an undirected graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. The nodes are denoted by a set of random variables and the absence of edge between a pair of nodes implies conditional independence between the corresponding pair of variables given all other variables (Guyon [1995], Murphy [2012]). Using the quasi-likelihood approach, we divide the estimation of a graph composed of p nodes into p separate sub-problems involving the conditional distributions. The method is scalable as it immediately gets rid of the intractable normalizing constant in these graphical models and incorporates in-built parallel computing mechanism that significantly reduces computational time on a multi-core system. The method simultaneously estimates the model parameters and the underlying structure of the graph. The MCMC algorithms constructed for sampling from the quasi-posterior distributions, are shown to be computationally scalable using simulations and we establish the applicability of the method by using it to estimate the underlying graph for the Sixteen Primary Factors Personality Data. We analyzed a dataset comprising of approximately 4000 observations to model a network of 163 questions on the 16 Personality Traits using a five colored Potts Model. Analysis of the estimated network revealed well known clusters consistent with existing literature and also shed light on other plausibly novel associations between personality traits.

CHAPTER II

A large scale quasi-bayesian inference with spike and slab priors

2.1 Introduction

We consider the problem of estimating a p-dimensional parameter using a dataset $z \in \mathcal{Z}$, and a likelihood or quasi-likelihood function $\ell : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z}$ denote a sample space equipped with a reference sigma-finite measure $\mathrm{d}z$. We assume that the quasi-likelihood function $(\theta, z) \mapsto \ell(\theta, z)$ is a jointly measurable function on $\mathbb{R}^p \times \mathcal{Z}$, and thrice differentiable in the parameter θ for any $z \in \mathcal{Z}$. We take a Bayesian approach with a spike-and-slab prior for θ. The prior requires the introduction of a new parameter $\delta \in \Delta \overset{\text{def}}{=} \{0,1\}^p$ with prior distribution $\{\omega(\delta),\ \delta \in \Delta\}$ which can be used for variable selection. The components of θ are then assumed to be conditionally independent given δ, and $\theta_j|\delta$ has a mean zero Gaussian distribution with precision parameter $\rho_1 > 0$ if $\delta_j = 1$ (slab prior), or a mean zero Gaussian distribution with precision parameter $\rho_0 > 0$ if $\delta_j = 0$ (spike prior). Spike-and-slab priors have been popularized by the seminal works Mitchell and Beauchamp [1988], George and McCulloch [1997] among others. Versions with a point-mass at the origin are known to have several optimality properties in high-dimensional problems (Johnstone and Silverman [2004], Castillo and van der Vaart [2012], Castillo et al. [2015], Atchade [2017]), but are computationally difficult to work with. In this work we follow George and McCulloch [1997], Narisetty and He [2014] and others, and replace the point-mass at the origin by a small-variance Gaussian distribution. The precision parameters ρ_1 and ρ_0 are constants that are chosen based on the size and dimensionality (n and p) of the problem, keeping the optimality conditions in mind. We then propose to study the following quasi-posterior distribution on $\Delta \times \mathbb{R}^p$,

$$\Pi(\delta, \mathrm{d}\theta|z) \propto e^{\ell(\theta_\delta; z)} \omega(\delta) \left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \left(\frac{\rho_0}{2\pi}\right)^{\frac{p-\|\delta\|_0}{2}} e^{-\frac{\rho_1}{2}\|\theta_\delta\|_2^2} e^{-\frac{\rho_0}{2}\|\theta-\theta_\delta\|_2^2} \mathrm{d}\theta, \qquad (2.1.1)$$

assuming that it is well-defined, where for $\theta \in \mathbb{R}^p$, and $\delta \in \Delta$, θ_δ denote their component-wise product and $\|\delta\|_0 \overset{\text{def}}{=} \sum_{j=1}^p \mathbf{1}_{\{|\delta_j|>0\}}$. A distinctive feature of (2.1.1) is that we have also replaced the quasi-likelihood $\ell(\theta; z)$ by a sparsified version $\ell(\theta_\delta; z)$. In other words, even if ℓ is a standard log-likelihood, (2.1.1) would still be different from the Gaussian-Gaussian spike-and-slab posterior distribution of George and McCulloch [1997], Narisetty and He [2014]. Due to the sparsification trick, $\theta - \theta_\delta$ does not contribute to the quasi-likelihood and the marginal quasi-posterior of (θ_δ, δ) is invariant to the choice of the spike prior on θ_{δ^c}. In this sense the spike prior is similar to the pseudo-prior of Carlin and Chib [1995]. It has the effect of bringing (2.1.1) closer to the point-mass spike-and-slab posterior distribution in terms of statistical performance, while at the same time providing tremendous computational speed as we will see in Theorem 1. The choice of ρ_0 controls the mixing of the quasi-posterior distribution. A low value of ρ_0 shall ensure faster mixing of the MCMC chains but the recovery and contraction rate of the parameters also depend on the fact that the ratio $\frac{\rho_0}{\rho_1}$ is large enough.

In this chapter, we study the sparsity and contraction properties of Π in Section 2.2 and 2.3 respectively. The Bernstein-von Mises theorem and the behavior of their variational approximations are considered in Section 2.4. We illustrate these results by considering examples such as inferring Gaussian graphical models, logistic regression and sparse principal component estimation in Section 2.5. All the proofs are collected in the appendix.

2.1.1 Notation

Throughout we equip the Euclidean space $\mathbb{R}^p$ ($p \geq 1$ integer) with its usual Euclidean inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|_2$, its Borel sigma-algebra, and its Lebesgue measure. All vectors $u \in \mathbb{R}^p$ are column-vectors unless stated otherwise. We also use the following norms on $\mathbb{R}^p$: $\|\theta\|_1 \overset{\text{def}}{=} \sum_{j=1}^p |\theta_j|$, $\|\theta\|_0 \overset{\text{def}}{=} \sum_{j=1}^p \mathbf{1}_{\{|\theta_j|>0\}}$, and $\|\theta\|_\infty \overset{\text{def}}{=} \max_{1\leq j\leq p} |\theta_j|$.

We set $\Delta \overset{\text{def}}{=} \{0,1\}^p$. For $\theta, \theta' \in \mathbb{R}^p$, $\theta \cdot \theta' \in \mathbb{R}^p$ denotes the component-wise product of θ and θ'. For $\delta \in \Delta$, we set $\mathbb{R}^p_\delta \overset{\text{def}}{=} \{\theta \cdot \delta : \theta \in \mathbb{R}^p\}$, and we write θ_δ as a short for $\theta \cdot \delta$. For $\delta, \delta' \in \Delta$, we write $\delta \supseteq \delta'$ to mean that for any $j \in \{1, \ldots, p\}$, whenever $\delta'_j = 1$, we have $\delta_j = 1$. Given $\theta \in \mathbb{R}^p$, and $\delta \in \Delta \setminus \{0\}$, we write $[\theta]_\delta$ to denote the δ-selected components of θ listed in their order of appearance: $[\theta]_\delta = (\theta_j,\ j \in \{1 \leq k \leq p :\ \delta_k = 1\}) \in \mathbb{R}^{\|\delta\|_0}$. Conversely, if $u \in \mathbb{R}^{\|\delta\|_0}$, we write $(u, 0)_\delta$ to denote the element of $\mathbb{R}^p_\delta$ such that $[(u,0)_\delta]_\delta = u$.

If $f(\theta, x)$ is a real-valued function that depends on the parameter θ and some other argument x, the notation $\nabla^{(k)} f(\theta, x)$, where k is an integer, denotes the k-th partial derivative with respect to θ of the map $(\theta, x) \mapsto f(\theta, x)$, evaluated at (θ, x). For $k = 1$, we write $\nabla f(\theta, x)$ instead of $\nabla^{(1)} f(\theta, x)$.

A continuous function $\mathsf{r} :\ [0, +\infty) \to [0, +\infty)$ is called a rate function if $\mathsf{r}(0) = 0$,

r is increasing and $\lim_{x\downarrow 0} \mathsf{r}(x)/x = 0$.

All constructs and other constants in the dissertation (including the sample size n) depend a priori on the dimension p. And we carry the asymptotics by letting p grow to infinity. We say that a term $x \in \mathbb{R}$ is an absolute constant if x does not depend on p. Throughout the dissertation C_0 denotes some generic absolute constant whose actual value may change from one appearance to the next.

2.2 Main assumptions and Posterior sparsity

We introduce here our two main assumptions. We set

$$\mathcal{L}_{\theta_1}(\theta; z) \stackrel{\text{def}}{=} \ell(\theta; z) - \ell(\theta_1; z) - \langle \nabla \ell(\theta_1; z), \theta - \theta_1 \rangle\,, \;\; \theta \in \mathbb{R}^p,$$

and we assume that the following holds.

H1. *We observe a $\mathcal{Z}$-valued random variable $Z \sim f_\star$, for some probability density $f_\star$ on $\mathcal{Z}$. Furthermore there exists $\delta_\star \in \Delta$, $\theta_\star \in \mathbb{R}^p_{\delta_\star}$, $\theta_\star \neq \boldsymbol{0}_p$, finite positive constants $\bar{\rho}, \bar{\kappa}$, such that $\mathbb{P}_\star(Z \in \mathcal{E}_0) > 0$, where*

$$\mathcal{E}_0 \stackrel{\text{def}}{=} \Big\{ z \in \mathcal{Z} :\; \Pi(\cdot|z) \textit{ is well-defined}, \;\; \|\nabla \ell(\theta_\star; z)\|_\infty \leq \frac{\bar{\rho}}{2}, \;\; \textit{and} \\ \mathcal{L}_{\theta_\star}(\theta; z) \geq -\frac{\bar{\kappa}}{2}\|\theta - \theta_\star\|_2^2, \;\; \textit{for all } \theta \in \mathbb{R}^p_{\delta_\star} \Big\}.$$

Furthermore, we assume that the prior parameter ρ_1 satisfies $32\rho_1\|\theta_\star\|_\infty \leq \bar{\rho}$, and we write $\mathbb{P}_\star$ and $\mathbb{E}_\star$ to denote probability and expectation operator under $f_\star$.

Remark II.1. H1 is very mild. Its main purpose is to introduce the data generating process, the true value of the parameter, and their relationship to the quasi-likelihood function. Specifically, since $\nabla \ell(\cdot; z)$ is null at the maximizer of $\ell(\cdot; z)$, having $z \in \mathcal{E}_0$ implies that the maximizer of $\ell(\cdot; z)$ is close to $\theta_\star$ in some sense, and the largest restricted (restricted to $\mathbb{R}^p_{\delta_\star}$) eigenvalue of the second derivative of $-\ell(\cdot; z)$ is bounded from above by $\bar{\kappa}$. The assumption that $\theta_\star \neq \mathbf{0}_p$ is made only out of mathematical convenience. All the results below continue to hold when $\theta_\star = \mathbf{0}_p$ albeit with minor adjustments. The condition $32\rho_1\|\theta_\star\|_\infty \leq \bar{\rho}$ is a loose condition as long as ρ_1 is chosen to grow at a lower rate than $\bar{\rho}$ but it is not exactly verifiable as $\theta_\star$ is unknown.

□

For convenience we will write $s_\star \overset{\text{def}}{=} \|\theta_\star\|_0$ to denote the number of non-zero components of the elements of $\theta_\star$. We assume next that the prior on δ is a product of independent Bernoulli distribution with small probability of success.

H2. *We assume that*

$$\omega(\delta) = \mathsf{q}^{\|\delta\|_0}(1-\mathsf{q})^{p-\|\delta\|_0}, \quad \delta \in \Delta,$$

where $\mathsf{q} \in (0, 1)$ *is such that* $\frac{\mathsf{q}}{1-\mathsf{q}} = \frac{1}{p^{u+1}}$*, for some absolute constant* $u > 0$*. Furthermore we will assume that* $p \geq 9$*,* $p^{u/2} \geq 2e^{2\rho_1}$*.*

Discrete priors as in H2 and generalizations were introduced by Castillo and van der Vaart [2012]. This is a very strong prior distribution that is well-suited for high-dimensional problems with limited sample where the signal is believed to be very sparse. It should be noted that this prior can perform poorly if these conditions are not met. We show next that the resulting posterior distribution is also typically sparse.

Theorem 1. *Assume H1-H2. Suppose that there exists a constant* r_0 *such that for all* $\delta \in \Delta$,

$$\log \mathbb{E}_\star \left[\boldsymbol{1}_{\mathcal{E}}(Z) e^{\mathcal{L}_{\theta_\star}(u;Z)+\left(1-\frac{\rho_1}{\bar{\rho}}\right)\langle \nabla \ell(\theta_\star;Z), u-\theta_\star\rangle} \right] \\ \leq \begin{cases} -\frac{r_0}{2}\|\delta_\star \cdot (u-\theta_\star)\|_2^2 & \textit{if } \; \|\delta_\star^c \cdot (u-\theta_\star)\|_1 \leq 7\|\delta_\star \cdot (u-\theta_\star)\|_1 \\ 0 & \textit{otherwise} \end{cases}, \tag{2.2.1}$$

for some measurable subset $\mathcal{E} \subseteq \mathcal{E}_0$. *If for some absolute constant* c_0 *we have*

$$s_\star\left(\frac{1}{2}+2\rho_1\right) + \frac{s_\star}{2}\log\left(1+\frac{\bar{\kappa}}{\rho_1}\right) + \frac{2\rho_1^2 s_\star}{r_0} + 2\rho_1\|\theta_\star\|_2^2 \leq c_0 s_\star \log(p), \tag{2.2.2}$$

then it holds that for all $j \geq 1$

$$\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{E}}(Z)\Pi\left(\|\delta\|_0 \geq s_\star\left(1+\frac{2(1+c_0)}{u}\right)+j \;|Z\right)\right] \leq \frac{2}{p^{\frac{uj}{2}}}.$$

Proof. See Appendix A.2. □

Theorem 1 is analogous to Theorem 1 of (Castillo et al. [2015]), and Theorem 3 of (Atchade [2017]), and says that the quasi-posterior distribution Π is automatically sparse in δ (of course θ is never sparse). The main contribution here is the fact that this behavior holds with Gaussian slab priors. The condition in (2.2.2) implies that the precision parameter of the slab density (that is ρ_1) should be of order $\log(p)$ or smaller. Simulation results (not reported here) show indeed that the method performs poorly if ρ_1 is taken too large.

Roughly speaking, the condition (2.2.1) is expected to hold if

$$\mathbf{1}_{\mathcal{E}_0}(Z)\mathcal{L}_{\theta_\star}(u;Z) \leq -\log \mathbb{E}_\star\left[e^{\left(1-\frac{\rho_1}{\bar{\rho}}\right)\langle\nabla\ell(\theta_\star;Z),u-\theta_\star\rangle}\right],$$

for all u in the cone $\mathcal{C} = \{u \in \mathbb{R}^p : \|\delta_\star^c \cdot (u - \theta_\star)\|_1 \leq 7\|\delta_\star \cdot (u - \theta_\star)\|_1\}$. If the quasi-score $\nabla \ell(\theta_\star; Z)$ is sub-Gaussian, then the right-hand side of the last display is lower bounded by $-c_0(1 - \rho_1/\bar{\rho})^2\|u - \theta_\star\|_2^2$, for some positive constant c_0. In this case (2.2.1) will hold if

$$\mathbf{1}_{\mathcal{E}_0}(Z)\mathcal{L}_{\theta_\star}(u; Z) \leq -c_0(1 - \rho_1/\bar{\rho})^2\|u - \theta_\star\|_2^2,$$

for all $u \in \mathcal{C}$. Hence (2.2.1) is a form restricted strong concavity of ℓ over $\mathcal{C}$. We refer the reader to (Negahban et al. [2012]) for more details on restricted strong concavity.

2.2.1 Implications for Markov Chain Monte Carlo sampling

Theorem 1 has implications for Markov Chain Monte Carlo (MCMC) sampling. To show this we consider a Metropolized-Gibbs strategy to sample from Π whereby we update θ keeping δ fixed, and then update δ keeping θ fixed – we refer the reader to (Robert and Casella [2004b]) for an introduction to basic MCMC algorithms. Note that given δ, $[\theta]_\delta$ and $[\theta]_{\delta^c}$ are conditionally independent, and $[\theta]_{\delta^c} \overset{i.i.d.}{\sim} \mathbf{N}(0, \rho_0^{-1})$, whereas $[\theta]_\delta$ can be updated using either its full conditional distribution when available, or using an extra MCMC update. For each j, given θ and δ_{-j}, the variable δ_j has a closed-form Bernoulli distribution. However, we choose to update δ_j using an Independent Metropolis-Hastings kernel with a $\mathsf{Ber}(0.5)$ proposal. Putting these steps together yields the following algorithm.

Algorithm 1. Draw $(\delta^{(0)}, \theta^{(0)}) \in \Delta \times \mathbb{R}^p$ from some initial distribution. For $k = 0, \ldots,$ repeat the following. Given $(\delta^{(k)}, \theta^{(k)}) = (\delta, \theta) \in \Delta \times \mathbb{R}^p$:

(STEP 1) For all j such that $\delta_j = 0$, draw $\theta_j^{(k+1)} \sim \mathbf{N}(0, \rho_0^{-1})$. Using $[\theta]_\delta$, draw jointly $[\theta^{(k+1)}]_\delta$ from some appropriate MCMC kernel on $\mathbb{R}^{\|\delta\|_0}$ with invariant distribution proportional to

$$u \mapsto e^{\ell((u,0)_\delta;z)-\frac{\rho_1}{2}\|u\|_2^2}.$$

(STEP 2) Given $\theta^{(k+1)} = \bar{\theta}$, set $\delta^{(k+1)} = \delta^{(k)}$ and do the following for $j = 1, \ldots, p$. Draw $\iota \sim \mathbf{Ber}(0.5)$. If $\delta_j^{(k+1)} = 0$, and $\iota = 1$, with probability $\min(1, A_j)/2$ change $\delta_j^{(k+1)}$ to ι. If $\delta_j^{(k+1)} = 1$, and $\iota = 0$, with probability $\min(1, A_j^{-1})/2$, change $\delta_j^{(k+1)}$ to ι; where

$$A_j \stackrel{\text{def}}{=} \frac{\mathsf{q}}{1-\mathsf{q}}\sqrt{\frac{\rho_1}{\rho_0}}e^{-(\rho_1-\rho_0)\frac{\bar{\theta}_j^2}{2}}e^{\ell(\bar{\theta}_\delta^{(j,1)};z)-\ell(\bar{\theta}_\delta^{(j,0)};z)}, \qquad (2.2.3)$$

where $\bar{\theta}_\delta^{(j,1)}, \bar{\theta}_\delta^{(j,0)} \in \mathbb{R}^p$ are defined as $(\bar{\theta}_\delta^{(j,1)})_k = (\bar{\theta}_\delta^{(j,0)})_k = (\bar{\theta}_\delta)_k$, for all $k \neq j$, and $(\bar{\theta}_\delta^{(j,1)})_j = \bar{\theta}_j$, $(\bar{\theta}_\delta^{(j,0)})_j = 0$.

□

We have left unspecified the MCMC kernel on $\mathbb{R}^{\|\delta\|_0}$ used in STEP 1, since it can be set up in many ways. Step 2 can also be replaced by adaptive procedures that have better mixing and for this purpose we refer the readers to (Ji and Schmidler [2013], Nott and Kohn [2005]) and references therein. Here we aim to show the efficiency gained in sampling from a sparsified quasi-posterior distribution. Let us call $C_1(\delta^{(k)})$ the computational cost of that part of STEP 1, and let $C_2(\delta)$ denote the cost of computing the quasi-likelihood $\ell(\theta_\delta; z)$ which is the dominant term in (2.2.3). Then as p grows, the total per-iteration cost of Algorithm 1 is of order

$$O\left(C_1(\delta^{(k)}) + pC_2(\delta^{(k)})\right).$$

Since Theorem 1 implies that a typical draw $\delta^{(k)}$ from the quasi-posterior distribution is sparse and satisfies $\|\delta^{(k)}\|_0 = O(s_\star)$, we can conclude that the per-iteration cost of the algorithm is accordingly reduced in problems where the sparsity of δ reduces the cost of the MCMC update in STEP 1, and the cost of computing the sparsified pseudo-likelihood $\ell(\theta_\delta; z)$. For instance, in a linear regression model (see Algorithm 2 in Appendix 2.6 for a detailed presentation), if the Gram matrix $X'X$ is pre-computed then $C_1(\delta^{(k)}) = O(\|\delta^{(k)}\|_0^3) = O(s_\star^3)$ (the cost of Cholesky decomposition), and $C_2(\delta^{(k)}) = O(\|\delta^{(k)}\|_0) = O(s_\star)$. As a result the per-iteration cost of Algorithm 2 grows with p as $O(s_\star^3 + s_\star p) = O(s_\star p)$, which is substantially faster than $O(\min(n,p)p^2)$ as needed by most MCMC algorithms for high-dimensional linear regression (Bhattacharya et al. [2016]). We refer the reader to Section 2.5.1 for a numerical illustration.

2.3 Contraction rate and model selection consistency

If in addition to the assumptions above, the restrictions of ℓ to the sparse subsets $\mathbb{R}^p_\delta$ are strongly concave then one can show that a draw θ from Π is typically close to $\theta_\star$. To elaborate on this, let $\bar{s} \geq s_\star$ be some arbitrary integer and set $\Delta_{\bar{s}} \overset{\text{def}}{=} \{\delta \in \Delta : \|\delta\|_0 \leq \bar{s}\}$, and

$$\mathcal{E}_1(\bar{s}) \overset{\text{def}}{=} \mathcal{E}_0 \cap \left\{ z \in \mathcal{Z} : \mathcal{L}_{\theta_\star}(\theta; z) \leq -\frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_2), \quad \text{for all } \delta \in \Delta_{\bar{s}},\ \theta \in \mathbb{R}^p_\delta \right\},$$

for some rate function r. Hence $z \in \mathcal{E}_1(\bar{s})$ implies that the function $u \mapsto \ell(u; z)$ behaves like a strongly concave function when restricted to $\mathbb{R}^p_\delta$, for all $\delta \in \Delta_{\bar{s}}$, but with a general rate function r. Here also, checking that $Z \in \mathcal{E}_1(\bar{s})$ boils down to checking a strong restricted concavity of ℓ, which can be done using similar methods as in Negahban et al. [2012]. The use of a general rate function r allows to handle problems that are not strongly convex in the usual sense (as for instance with logistic regression). Our main result in this section states that when $z \in \mathcal{E}_1(\bar{s})$, we are automatically guaranteed a minimum rate of contraction for Π given by

$$\epsilon \stackrel{\text{def}}{=} \inf\left\{z > 0 :\ \mathsf{r}(x) - 2(s_\star + \bar{s})^{1/2}\bar{\rho}x \geq 0, \ \text{ for all } x \geq z\right\}. \tag{2.3.1}$$

To gain some intuition on ϵ, consider a linear regression model where $\ell(\theta; z) = -\|z - X\theta\|_2^2/(2\sigma^2)$. Then we have

$$\mathcal{L}_{\theta_\star}(\theta; z) = -\frac{n}{2\sigma^2}(\theta - \theta_\star)'\left(\frac{X'X}{n}\right)(\theta - \theta_\star).$$

If $\theta \in \mathbb{R}^p_\delta$ for some $\delta \in \Delta_{\bar{s}}$, then $\mathcal{L}_{\theta_\star}(\theta; z) \leq -n\underline{v}(\bar{s} + s_\star)\|\theta - \theta_\star\|_2^2/(2\sigma^2)$, where $\underline{v}(\bar{s} + s_\star)$ is the restricted smallest eigenvalue of $X'X/n$ over $(\bar{s} + s_\star)$-sparse vectors. Hence, we can take the rate function $\mathsf{r}(x) = n\underline{v}(\bar{s} + s_\star)x^2/\sigma^2$, In that case the contraction rate in (2.3.1) gives $\epsilon = 2\sigma^2(\bar{s} + s_\star)^{1/2}\bar{\rho}/(n\underline{v}(\bar{s} + s_\star))$. The final form of the rate depends on $\bar{\rho}$ (in H1) which is determined by the tail behavior of the quasi-score $\nabla\ell(\theta_\star; Z)$. In the sub-Gaussian case $\bar{\rho} \propto \sqrt{n\log(p)}$, and this gives $\epsilon \propto \sqrt{(\bar{s} + s_\star)\log(p)/n}$. We refer the reader to the proof of Corollary 6 for more details.

We set

$$\mathsf{B} \stackrel{\text{def}}{=} \bigcup_{\delta \in \Delta_{\bar{s}}} \{\delta\} \times \mathsf{B}^{(\delta)}, \tag{2.3.2}$$

where

$$\mathsf{B}^{(\delta)} \stackrel{\text{def}}{=} \left\{\theta \in \mathbb{R}^p :\ \|\theta_\delta - \theta_\star\|_2 \leq C\epsilon,\ \|\theta - \theta_\delta\|_2 \leq \sqrt{(1 + C_1)\rho_0^{-1}p,}\right\}, \tag{2.3.3}$$

for some absolute constants $C, C_1 \geq 3$, where ϵ is as defined in (2.3.1). Our next result says that if $(\delta, \theta) \sim \Pi(\cdot|Z)$ and $Z \in \mathcal{E}_1(\bar{s})$, then with high probability we have $\theta \in \mathsf{B}^{(\delta)}$ for some $\delta \in \Delta_{\bar{s}}$: θ_δ is close to $\theta_\star$, and $\theta - \theta_\delta$ is small.

Theorem 2. *Assume H1-H2. Let $\bar{s} \geq s_\star$ be some arbitrary integer, and take $\mathcal{E} \subseteq \mathcal{E}_1(\bar{s})$. If*

$$C\bar{\rho}(s_\star + \bar{s})^{1/2}\epsilon \geq 32 \max\left[\bar{s}\log(p),\ (1+u)s_\star \log\left(p + \frac{p\bar{\kappa}}{\rho_1}\right)\right], \tag{2.3.4}$$

then for all p large enough,

$$\mathbb{E}_\star\left[\mathit{1}_{\mathcal{E}}(Z)\Pi\left(\mathsf{B}^c|Z\right)\right] \leq \mathbb{E}_\star\left[\mathit{1}_{\mathcal{E}}(Z)\Pi\left(\|\delta\|_0 > \bar{s}\ |Z\right)\right] + 8e^{-\frac{C}{32}\bar{\rho}(s_\star+\bar{s})^{1/2}\epsilon} + 2e^{-p} \tag{2.3.5}$$

where $\mathsf{B}^c \overset{\text{def}}{=} (\Delta \times \mathbb{R}^p) \setminus \mathsf{B}$.

Proof. See Appendix A.3. □

Remark II.2. The result implies that for j such that $\delta_j = 0$, $|\theta_j| = O(\sqrt{\rho_0^{-1}})$ under Π. As a result we recommend scaling ρ_0^{-1} in practice as

$$\rho_0^{-1} = \frac{C_0}{n}, \quad \text{or} \quad \rho_0^{-1} = \frac{C_0}{p}.$$

When the posterior distribution is known to be sparse one can choose $\bar{s}$ appropriately to make the first term on the right hand side of (2.3.5) small. For instance under the assumptions of Theorem 1, we can take

$$\bar{s} = s_\star\left(1 + \frac{2(1+c_0)}{u}\right) + k.$$

If in addition $\mathbb{P}_\star(Z \notin \mathcal{E}_1(\bar{s})) \to 0$ as $p \to \infty$, we can deduce from (2.3.5) that $\mathbb{E}_\star[\Pi(\mathsf{B}^c|Z)] \to 0$, as $p \to \infty$. If Theorem 1 does not apply, one can modify H2 to impose the sparsity constraint $\|\delta\|_0 \le \bar{s}$ directly in the prior distribution. In this case the first term on the right hand side of (2.3.5) automatically vanishes. The main drawback in this approach is that an a priori knowledge of $\bar{s} \ge s_\star$ is needed in order to use the quasi-posterior distribution with a possible risk of misspecification.

□

We now show that when the non-zero components of $\theta_\star$ are sufficiently large, Π achieves perfect model selection. Given $\delta \in \Delta_{\bar{s}}$ we define the function $\ell^{[\delta]}(\cdot; z) : \mathbb{R}^{\|\delta\|_0} \to \mathbb{R}$ by $\ell^{[\delta]}(u; z) \overset{\text{def}}{=} \ell((u, 0)_\delta; z)$. We then introduce the estimators

$$\hat{\theta}_\delta(z) \overset{\text{def}}{=} \underset{u \in \mathbb{R}^{\|\delta\|_0}}{\mathsf{Argmax}}\ \ell^{[\delta]}(u; z), \quad z \in \mathcal{Z}. \tag{2.3.6}$$

When $\delta = \delta_\star$ we write $\hat{\theta}_\star(z)$. At times, to shorten the notation we will omit the data z and write $\hat{\theta}_\delta$ instead of $\hat{\theta}_\delta(z)$. Recall for $z \in \mathcal{E}_1(\bar{s})$ the functions $\ell^{[\delta]}(\cdot; z)$ are strongly concave. Therefore for $z \in \mathcal{E}_1(\bar{s})$, the estimators $\hat{\theta}_\delta$ are well-defined for all $\delta \in \Delta_{\bar{s}}$. Omitting the data z, we will write $\mathcal{I}_\delta \in \mathbb{R}^{\|\delta\|_0 \times \|\delta\|_0}$ to denote the negative of the matrix of second derivatives of $u \mapsto \ell^{[\delta]}(u; z)$ evaluated at $\hat{\theta}_\delta(z)$. That is

$$\mathcal{I}_\delta \overset{\text{def}}{=} -\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta; z) \in \mathbb{R}^{\|\delta\|_0 \times \|\delta\|_0}.$$

Note that $\mathcal{I}_\delta$ is simply the sub-matrix of $\nabla^{(2)}\ell((\hat{\theta}_\delta, 0)_\delta; z)$ obtained by taking the rows and columns for which $\delta_j = 1$. When $\delta = \delta_\star$, we will write $\mathcal{I}$ instead of $\mathcal{I}_{\delta_\star}$.

For $a > 0$, and $\delta \in \Delta \setminus \{0\}$, we define

$$\varpi(\delta, a; z) \stackrel{\text{def}}{=} \sup_{u \in \mathbb{R}^{\|\delta\|_0}:\ \|u - \hat{\theta}_\delta\|_2 \leq a} \ \max_{1 \leq i,j,k \leq \|\delta\|_0} \left| \frac{\partial^3 \ell^{[\delta]}(u; z)}{\partial u_i \partial u_j \partial u_k} \right|.$$

$\varpi(\delta, a; z)$ measures the deviation of the log-quasi-likelihood from its quadratic approximation around $\hat{\theta}_\delta$. With the rate ϵ as in (2.3.1), we will make the assumption that

$$\min_{j:\ \delta_{\star j} = 1} |\theta_{\star j}| > C\epsilon. \tag{2.3.7}$$

Clearly this assumption is unverifiable in practice since $\theta_\star$ is typically not known. However a strong signal assumption such as (2.3.7) is needed in one form or the other for exact model selection (Narisetty and He [2014], Castillo et al. [2015], Yang et al. [2016]). Furthermore as we show in Section 2.5.1, in specific models (2.3.7) translates into a condition on the sample size n, which in some cases can help the user evaluates in practice whether (2.3.7) seems reasonable or not. An understanding of the behavior of Π when (2.3.7) does not hold remains an interesting problem for future research.

One can readily observe that when (2.3.7) holds, then the set $\mathsf{B}^{(\delta)}$ introduced above is necessarily empty when δ does not contain the true model $\delta_\star$. In other words, when (2.3.7) holds, the set B defined in (2.3.2) can be written as

$$\mathsf{B} = \bigcup_{\delta \in \mathcal{A}_{\bar{s}}} \{\delta\} \times \mathsf{B}^{(\delta)},$$

where

$$\mathcal{A}_{\bar{s}} \stackrel{\text{def}}{=} \{\delta \in \Delta :\ \|\delta\|_0 \leq \bar{s}, \ \ \text{and } \delta \supseteq \delta_\star\},$$

and we recall that the notation $\delta \supseteq \delta'$ means that $\delta_j = 1$ whenever $\delta'_j = 1$ for all j.

More generally, for $j \geq 0$, we set

$$\mathcal{A}_{s_\star+j} \stackrel{\text{def}}{=} \{\delta \in \Delta :\ \|\delta\|_0 \leq s_\star + j,\ \delta \supseteq \delta_\star\}, \quad \text{and} \quad \mathsf{B}_j = \bigcup_{\delta \in \mathcal{A}_{s_\star+j}} \{\delta\} \times \mathsf{B}^{(\delta)}.$$

In particular $\mathsf{B}_0 = \{\delta_\star\} \times \mathsf{B}^{(\delta_\star)}$, and $(\delta, \theta) \in \mathsf{B}_j$ implies that δ has at most j false-positive (and no false-negative). We set

$$\mathcal{E}_2(\bar{s}) \stackrel{\text{def}}{=} \mathcal{E}_1(\bar{s}) \cap \bigcap_{j=1}^{\bar{s}-s_\star} \left\{ z \in \mathcal{Z} :\ \max_{\delta \in \mathcal{A}_{\bar{s}}:\ \|\delta\|_0 = s_\star+j} \ell^{[\delta]}(\hat{\theta}_\delta; z) - \ell^{[\delta_\star]}(\hat{\theta}_\star; z) \leq \frac{ju}{2} \log(p) \right\},$$

which imposes a growth condition on the log-quasi-likelihood ratios of sparse sub-models.

Theorem 3. *Assume H1-H2, and (2.3.7). Let $\bar{s} \geq s_\star$ be some arbitrary integer, and take $\mathcal{E} \subseteq \mathcal{E}_2(\bar{s})$. For some constant $\underline{\kappa} > 0$, suppose that for all $z \in \mathcal{E}$,*

$$\min_{\delta \in \mathcal{A}_{\bar{s}}} \inf_{u \in \mathbb{R}^{\|\delta\|_0}:\ \|u - \hat{\theta}_\delta\|_2 \leq 2\epsilon} \inf \left\{ \frac{v' \left(-\nabla^{(2)} \ell^{[\delta]}(u; z) \right) v}{\|v\|_2^2},\ v \in \mathbb{R}^{\|\delta\|_0},\ v \neq 0 \right\} \geq \underline{\kappa}, \quad (2.3.8)$$

and

$$\max_{\delta \in \mathcal{A}_{\bar{s}}} \sup_{u \in \mathbb{R}^{\|\delta\|_0}} \sup \left\{ \frac{v' \left(-\nabla^{(2)} \ell^{[\delta]}(u; z) \right) v}{\|v\|_2^2},\ v \in \mathbb{R}^{\|\delta\|_0},\ v \neq 0 \right\} \leq \bar{\kappa}, \quad (2.3.9)$$

where $\bar{\kappa}$ is as in H1. Then it holds that for any $j \geq 1$

$$\begin{aligned} &\boldsymbol{1}_{\mathcal{E}}(z) \left(1 - \Pi\left(\mathsf{B}_j | z\right)\right) \\ &\qquad \leq 8 e^{C_0(\rho_1 \|\theta_\star\|_\infty \bar{s}^{1/2}\epsilon + \mathsf{a}_2 \bar{s}^{3/2}\epsilon^3)} e^{\frac{2\mathsf{a}_2 \bar{s}^3 \epsilon}{\underline{\kappa}}} \left(\sqrt{\frac{\rho_1}{\underline{\kappa}}} \frac{1}{p^{\frac{u}{2}}} \right)^{j+1} + \boldsymbol{1}_{\mathcal{E}}(z) \Pi(\mathsf{B}^c | z), \quad (2.3.10) \end{aligned}$$

provided that $\underline{\kappa} p^u \geq 4\rho_1$, and $(C-1)\epsilon\underline{\kappa}^{1/2} \geq 2(s_\star^{1/2}+1)$, where $\mathsf{a}_2 \stackrel{\text{def}}{=} \max_{\delta \in \mathcal{A}_{\bar{s}}} \varpi(\delta, (C+$

$1)\epsilon; z)$, and C_0 some absolute constant.

Proof. See Appendix A.4. □

We note that $\mathsf{B}_0 = \{\delta_\star\} \times \mathsf{B}^{(\delta_\star)} \subset \{\delta_\star\} \times \mathbb{R}^p$. Hence by choosing $j = 0$, (2.3.10) provides a lower bound on the probability of perfect model selection $\Pi(\delta_\star|z)$.

Remark II.3. The left hand sides of (2.3.8) and (2.3.9) are restricted eigenvalues. We note that the infimum on u in (2.3.8) is taken over a small neighborhood of $\hat{\theta}_\delta$, which is an important detail that facilitates the application of the result. The main challenge in using this result is bounding the probability of the event $\mathcal{E}_2(\bar{s})$ (which deals with the behavior of the quasi-likelihood ratio statistics). For linear regression problems, this boils down to deviation bounds for projected Gaussian distributions as we show in Section 2.5.1. An extension to generalized linear models via the Hanson-Wright inequality seems plausible although not pursed here.

□

2.4 Posterior approximations

We show here that a Bernstein-von Mises approximation holds in the KL-divergence sense. We consider the distribution

$$\Pi_\star^{(\infty)}(\delta, \mathrm{d}\theta|z) \propto \mathbf{1}_{\delta_\star}(\delta) e^{-\frac{1}{2}([\theta]_{\delta_\star} - \hat{\theta}_\star)' \mathcal{I}([\theta]_{\delta_\star} - \hat{\theta}_\star) - \frac{\rho_0}{2}\|\theta - \theta_{\delta_\star}\|_2^2} \mathrm{d}\theta, \tag{2.4.1}$$

which puts probability one on $\delta_\star$, and draws independently $[\theta]_{\delta_\star} \sim \mathbf{N}(\hat{\theta}_\star, \mathcal{I}^{-1})$, and $[\theta]_{\delta_\star^c} \overset{i.i.d.}{\sim} \mathbf{N}(0, \rho_0^{-1})$. Our version of the Bernstein-von Mises theorem says that Π behaves like $\Pi_\star^{(\infty)}$. If μ, ν are two probability measures on some measurable space we define the Kulback-Leibler divergence (KL-divergence) of μ respect to ν as

$$\mathsf{KL}\left(\mu|\nu\right) \overset{\text{def}}{=} \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu, & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise.} \end{cases}$$

A Bernstein-von Mises approximation in the KL-divergence sense – unlike the analogous result in the total variation metric – requires a control of the tails of the log-quasi-likelihood. To limit the technical details we will focus on the case where those tails are quadratic.

Theorem 4. *Assume H1-H2. For some integer $\bar{s} \geq s_\star$, and some constant $\underline{\kappa} > 0$, let $\mathcal{E}$ be some measurable subset of $\mathcal{Z}$ such that for all $z \in \mathcal{E}$, $\Pi(\delta_\star|z) \geq 1/2$, (2.3.9) holds with $\bar{\kappa}$ as in H1, and*

$$\min_{\delta\in\mathcal{A}_{\bar{s}}} \inf_{u\in\mathbb{R}^{\|\delta\|_0}} \inf\left\{\frac{v'\left(-\nabla^{(2)}\ell^{[\delta]}(u;z)\right)v}{\|v\|_2^2},\ v\in\mathbb{R}^{\|\delta\|_0},\ v\neq 0\right\} \geq \underline{\kappa}. \qquad (2.4.2)$$

Then there exists an absolute constants C_0 such that

$$\boldsymbol{1}_{\mathcal{E}}(z)\mathsf{KL}\left(\Pi_\star^{(\infty)}|\Pi\right) \leq C_0\left(\rho_1\bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}\epsilon^3\right) + \frac{3\rho_1^2(\epsilon+\|\theta_\star\|_2)^2}{2(\rho_1+\bar{\kappa})}$$
$$+ C_0(\rho_1+\bar{\kappa})\epsilon^2\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^{\frac{s_\star}{2}} e^{-\frac{(C-1)^2\epsilon^2\underline{\kappa}}{32}} + C_0(\rho_1+\bar{\kappa})e^{-p} + 2\boldsymbol{1}_{\mathcal{E}}(z)(1-\Pi(\delta_\star|z)), \quad (2.4.3)$$

provided that $\underline{\kappa}(C-1)\epsilon \geq 4\max(\sqrt{s_\star\underline{\kappa}}, \rho_1(\epsilon + s_\star^{1/2}\|\theta_\star\|_\infty))$, where C is as in Theorem 2.

Proof. See Appendix A.5. □

Remark II.4. The upper bound in (2.4.3) implies an upper bound on the total variation distance between Π and $\Pi_\star^{(\infty)}$ via Pinsker's inequality (see e.g. Boucheron et al.

[2013] Theorem 4.19). The leading term in (2.4.3) is typically $C_0(\rho \bar{s}^{1/2}\epsilon + \mathsf{a}_2 \bar{s}^{3/2}\epsilon^3)$ which gives a non-trivial convergence rate in the Bernstein-von Mises approximation. The fact that we define the approximating distribution $\Pi_\star^{(\infty)}$ as restricted on $\{\delta_\star\} \times \mathsf{B}^{(\delta_\star)}$ is not restrictive. Indeed, in most examples one can easily show (by standard Gaussian deviation) that the total variation distance between $\Pi_\star^{(\infty)}$ and its unrestricted version converges to zero as $p \to \infty$. And this can be combined with Theorem 4 and the Pinsker's inequality to guarantee that the total variation distance between Π and the unrestricted version of $\Pi_\star^{(\infty)}$ converges to zero as well, as $p \to \infty$. In fact we could have worked directly with the unrestricted version of $\Pi_\star^{(\infty)}$ to obtain the bound on the KL-divergence in Theorem 4. We have chosen not to proceed that way because the resulting bound is much more involved.

□

2.4.1 Implications for variational approximations

When dealing with very large scale problems, practitioners often turn to variational approximation methods to obtain fast approximations of Π. We explore some implications of Theorem 4 on the behavior of variational approximation methods in the high-dimensional setting. Let $\mathcal{S} \in \{0,1\}^{p\times p}$ be a symmetric matrix, and let $\mathcal{M}_p^+(\mathcal{S})$ be the set of all $p \times p$ symmetric positive definite (spd) matrices with sparsity pattern $\mathcal{S}$ (that is $M \in \mathcal{M}_p^+(\mathcal{S})$ means that $\mathcal{S} \cdot M = M$, where $A \cdot B$ is the component-wise product of A, B). We assume in addition that $\mathcal{S}$ is such that if M is spd then $\mathcal{S} \cdot M$ is also spd. We consider the family $\mathcal{Q} \overset{\text{def}}{=} \{Q_\Psi,\ \Psi\}$ of probability measures on $\Delta \times \mathbb{R}^p$, indexed by $\Psi = (q, \mu, C) \in (0,1)^p \times \mathbb{R}^p \times \mathcal{M}_p^+(\mathcal{S})$, where

$$Q_\Psi(\mathrm{d}\delta, \mathrm{d}\theta) = \prod_{j=1}^{p} \mathbf{Ber}(q_j)(\mathrm{d}\delta_j)\mathbf{N}_p(\mu, C)(\theta)\mathrm{d}\theta, \qquad (2.4.4)$$

In these definitions $\mathbf{Ber}(\alpha)(\mathrm{d}x)$ is the probability measure on $\{0,1\}$ that assigns probability α to 1, and $\mathbf{N}_p(m,V)(\cdot)$ is the density of p-dimensional Gaussian distribution $\mathbf{N}_p(m,V)$. Let Q be the minimizer of the KL-divergence $\mathsf{KL}\left(Q|\Pi\right)$ over the family $\mathcal{Q}$:

$$Q \overset{\text{def}}{=} \underset{Q\in\mathcal{Q}}{\mathsf{Argmin}}\ \mathsf{KL}\left(Q|\Pi\right). \tag{2.4.5}$$

We call Q the variational approximation of Π over the family $\mathcal{Q}$. Although not shown in the notation, Q depends on the data z. We will consider the following examples.

Example 1 (Skinny variational approximation)*.* If $\mathcal{S} = I_p$, then Q corresponds to a mean-field variational approximation of Π. We will refer to this approximation below as the skinny variational approximation (skinny-VA) of Π.

Example 2 (full and midsize variational approximations)*.* If $\mathcal{S}$ is taken as the full matrix with all entries equal to 1, we will refer to Q as the full variational approximation (full-VA) of Π. More generally let $\delta^{(\mathsf{i})}$ be some element of $\{0,1\}^p$ that we call a template. Ideally we want $\delta^{(i)}$ to be sparse and to contain the true model, but this needs not be assumed. We then define $\mathcal{S}$ as follows: $\mathcal{S}_{ij} = 1$ if $i=j$, and $\mathcal{S}_{ij} = \delta^{(\mathsf{i})}_i\delta^{(\mathsf{i})}_j$ if $i\neq j$. If $\delta^{(\mathsf{i})}$ is sparse, matrices $M\in\mathcal{M}_p^+(\mathcal{S})$ are also sparse. In that case we call Q a midsize variational approximation (midsize-VA) of Π. We note that we also recover the skinny-VA by taking $\delta^{(\mathsf{i})} = \mathbf{0}_p$, and we recover the full-VA by taking $\delta^{(\mathsf{i})}$ as the vector with components equal to 1.

The appeal of variational approximation methods is that Q can be approximated using algorithms that are order of magnitude faster than MCMC. We note however that the optimization problem in (2.4.5) is non-convex in general. Hence, convergence guarantees for these algorithms are difficult to establish. We do not address these issues here. Instead we would like to explore the behavior of Q in view of

Theorem 4. Let us rewrite the distribution $\Pi_\star^{(\infty)}$ in (2.4.1) as

$$\Pi_\star^{(\infty)}(\delta, \mathrm{d}\theta|z) \propto \mathbf{1}_{\delta_\star}(\delta) e^{-\frac{1}{2}(\theta-\hat{\theta}_\star)'\bar{\mathcal{I}}_\gamma(\theta-\hat{\theta}_\star)}\mathrm{d}\theta,$$

where we abuse notation to write $(\hat{\theta}_\star, 0)_{\delta_\star}$ as $\hat{\theta}_\star$, and $\bar{\mathcal{I}}_\gamma \in \mathbb{R}^{p\times p}$ is such that $[\bar{\mathcal{I}}_\gamma]_{\delta_\star,\delta_\star} = \mathcal{I}$, $[\bar{\mathcal{I}}_\gamma]_{\delta_\star,\delta_\star^c} = [\bar{\mathcal{I}}_\gamma]'_{\delta_\star^c,\delta_\star} = 0$, and $[\bar{\mathcal{I}}_\gamma]_{\delta_\star^c,\delta_\star^c} = (1/\gamma)I_{p-s_\star}$. Then we set

$$\tilde{\Pi}_\star^{(\infty)}(\delta, \mathrm{d}\theta|z) \propto \mathbf{1}_{\delta_\star}(\delta) e^{-\frac{1}{2}(\theta-\hat{\theta}_\star)'\left(\mathcal{S}\cdot\bar{\mathcal{I}}_\gamma\right)(\theta-\hat{\theta}_\star)}\mathrm{d}\theta. \tag{2.4.6}$$

The total variation metric between two probability measure is defined as

$$\|\mu-\nu\|_{\mathrm{tv}} \overset{\mathrm{def}}{=} \sup_{A \text{ meas.}} \left(\mu(A)-\nu(A)\right).$$

Theorem 5. *Assume H1-H2. For all $z \in \mathcal{Z}$ such that $\Pi(\cdot|z)$ and $\Pi_\star^{(\infty)}(\cdot|z)$ are well-defined we have*

$$\|Q-\tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2 \leq 8\zeta + 16\int_{\delta_\star\times\mathbb{R}^p} \log\left(\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{\Pi}_\star^{(\infty)}, \tag{2.4.7}$$

where

$$\zeta = \log\left(\frac{\det(\bar{\mathcal{I}}_\gamma)}{\det(\mathcal{S}\cdot\bar{\mathcal{I}}_\gamma)}\right) + \mathit{Tr}\left(\bar{\mathcal{I}}_\gamma^{-1}(\mathcal{S}\cdot\bar{\mathcal{I}}_\gamma)\right) - p. \tag{2.4.8}$$

Proof. See Appendix A.6. □

Remark II.5. As we show below in the proof of Theorem 4, the integral on the right size of (2.4.7) behaves like $\mathsf{KL}\left(\Pi_\star^{(\infty)}|\Pi\right)$, which can be shown to vanish using the Bernstein-von Mises theorem (Theorem 4) under appropriate regularity conditions. In this case, whether Q behaves like $\tilde{\Pi}_\star^{(\infty)}$ can be deduced from the behavior of ζ, a term that is easier to analyze. For instance for the full-VA $\zeta = 0$. More generally for any midsize-VA such that $\delta^{(\mathsf{i})} \supseteq \delta_\star$, we have $\zeta = 0$. In the case of the skinny-VA (mean field variational approximation), $\zeta > 0$ in general, but $\zeta = o(1)$ when the off-diagonal elements of the information matrix $\mathcal{I}$ are $o(1)$.

□

Remark II.6. Theorem 5 gives an approximation (in total variation sense) of the variational approximation. To the exception of (Wang and Blei [2018]) most of the theoretical work on variational approximation methods have focused on concentration: whether the variational approximation put most of its probability mass around the true value (see e.g. Alquier and Ridgway [2017], Ray and Szabo [2019] for some recent results, and Wang and Blei [2018] for an overview of the literature), without addressing whether other aspects of the distribution are recovered well. One important limitation of Wang and Blei [2018] which makes the extension of their approach to high-dimension problematic is their reliance on a) local asymptotic normality assumptions, and b) the assumption that the variational family can be viewed as a re-scaled version of some sample-size independent family.

□

2.5 Examples

The theory developed in this thesis while applicable to a general quasi-likelihood setting, also holds when we proceed with the true likelihood. In our effort to illustrate the implications of the results, we first show the application in the settings of simple linear regression followed by logistic regression, gaussian graphical models and finally end with possible applications to sparse principal component analysis.

2.5.1 Linear Regression

In case of linear regression, posterior contraction results with spike and slab priors have been studied extensively by Castillo et al. [2015],Narisetty and He [2014],Rockova and George [2014] and others. In this example we merely aim to illustrate how our theory behaves in the linear regression settings without getting into comparison with the aforementioned works. The theory for linear regression can be extended to a quasi-likelihood setting in case of Gaussian Graphical Models discussed later.

A 1. *For some parameter $\theta_\star \in \mathbb{R}^p \setminus \{0\}$, $Y = X\theta_\star + V$, where $V^{n\times 1} \sim N(0, \sigma^2)$, $\sigma^2 > 0$ and $X \in \mathbb{R}^{n\times p}$ is the covariate matrix with columns $X_j, j = 1, \cdots, p$ such that* $\max_{j=1,\cdots p} \|X_j\|_\infty \leq \tau$.

For $\rho_0 > 0, \rho_1 > 0$, the posterior distribution on $\Delta \times \mathbb{R}^p$ is given by

$$\Pi(\delta, \mathrm{d}\theta|Z) \propto$$
$$e^{-\frac{1}{2\sigma^2}\|Y - X\theta_\delta\|_2^2}\omega(\delta)\left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}}\left(\frac{\rho_0}{2\pi}\right)^{\frac{p-\|\delta\|_0}{2}} e^{-\frac{\rho_1}{2}\|\theta_\delta\|_2^2} e^{-\frac{\rho_0}{2}\|\theta - \theta_\delta\|_2^2}\mathrm{d}\theta, \quad (2.5.1)$$

We further define

$$\underline{\nu} = \inf\{\frac{u'X'Xu}{n\|u\|_2};\ u \in \mathbb{R}^p;\ u \neq 0\ \|\delta_\star^c \cdot (u - \theta_\star)\|_1 \leq 7\|\delta_\star \cdot (u - \theta_\star)\|_1\};$$
$$\text{and}\ \ \underline{\nu}(s) = \inf\{\frac{u'X'Xu}{n\|u\|_2};\ u \in \mathbb{R}^p\ u \neq 0\ \|u\|_0 \leq s\} \quad (2.5.2)$$

We use the theory of Section 2.2-2.4 to describe the behavior of this approach to infer $\theta_\star$. We focus on the case where $n = o(p)$, and we recall that C_0 is an absolute constant whose value may be different from one expression to the other. Let $\Pi_\star^{(\infty)}$

be the corresponding limiting distribution of Π as defined in (2.4.1), and let $\tilde{\Pi}_\star^{(\infty)}$ be the corresponding approximation given in (2.4.6). In this particular case, $\Pi_\star^{(\infty)}$ is the probability measure on $\Delta \times \mathbb{R}^p$ that puts probability one on $\delta_\star$ (the support of $\theta_\star$), draws $[\theta]_{\delta_\star} \sim \mathbf{N}\left(\hat{\theta}_\star, \sigma^2(X'_{\delta_\star}X_{\delta_\star})^{-1}\right)$, and draws independently all other components i.i.d. from $\mathbf{N}(0, \rho_0^{-1})$, where $\hat{\theta}_\star$ is the OLS estimator $(X_{\delta_\star}X_{\delta_\star})^{-1}X'_{\delta_\star}Y$. We set $s_\star \overset{\text{def}}{=} \|\theta_\star\|_0$. Let Q denote the variational approximation of Π based on the family (2.4.4) with sparsity pattern $\mathcal{S}$, and let ζ denote the corresponding term in (2.4.8).

Corollary 6. *Assume H2, A1, and suppose that $s_\star > 0$, $\|\theta_\star\|_\infty = O(1)$, and $s_\star = O(\log(p))$ as $p \to \infty$.Suppose also that $u > 2$ and choose the prior parameter ρ_1 as*

$$\rho_1 = \frac{1}{\sigma}\sqrt{\frac{\log(p)}{n}}.$$

Set

$$\bar{s} \overset{\text{def}}{=} s_\star\left(1 + \frac{6}{u}\right) + \frac{u}{4}, \quad \epsilon \overset{\text{def}}{=} C_0\sigma\sqrt{\frac{(\bar{s} + s_\star)\log(p)}{n}},$$

Assume

$$\frac{1}{\underline{\nu}} + \frac{1}{\underline{\nu}(s)} \sim O(1) \tag{2.5.3}$$

Further suppose that the sample size n satisfies $n = o(p)$, as $p \to \infty$,

$$n \geq C_0\bar{s}\log(p),$$

and the strong signal assumption

$$\min_{k:\, |\theta_{\star,k}|>0} |\theta_{\star,k}| > C_0\epsilon \tag{2.5.4}$$

holds. Then there exists a measurable set $\mathcal{G}$ with $\mathbb{P}_\star(Z \notin \mathcal{G}) \to 0$ as $p \to \infty$ such *that*

$$\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{G}}(Z)\mathsf{KL}\left(\Pi_\star^{(\infty)}|\Pi\right)\right] \leq C_0(\bar{s}+s_\star)\frac{\log(p)}{n} + \frac{C_0}{p^{1\wedge\left(\frac{u}{2}-1\right)}}. \tag{2.5.5}$$

Furthermore the variational approximation Q satisfies

$$\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{G}}(Z)\|Q-\tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2\right] \leq 8\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{G}}(Z)\zeta\right] + C_0(\bar{s}+s_\star)\frac{\log(p)}{n} + \frac{C_0}{p^{1\wedge\left(\frac{u}{2}-1\right)}}. \tag{2.5.6}$$

Proof. See Appendix A.7. □

Remark II.7. The assumption (2.5.3) on the growth of the restricted eigenvalues ensures that we restrict ourselves to problems that do not become intrinsically harder as p increases.

□

2.5.1.1 Numerical illustration

We perform a simulation study to assess the behavior of the posterior distribution and its variational approximations as described in Corollary 6. We set $p = 1000$, $n \in \{100, 500\}$, and we generate $Z = [Y, X] \in \mathbb{R}^{n\times(p+1)}$ as follows. We first generate the matrix X by simulating the rows of X independently from a Gaussian distribution with correlation $\psi^{|j-i|}$

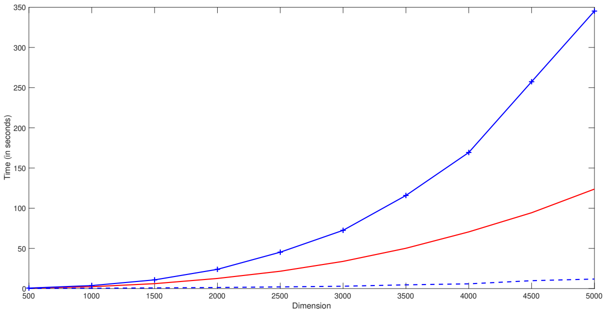


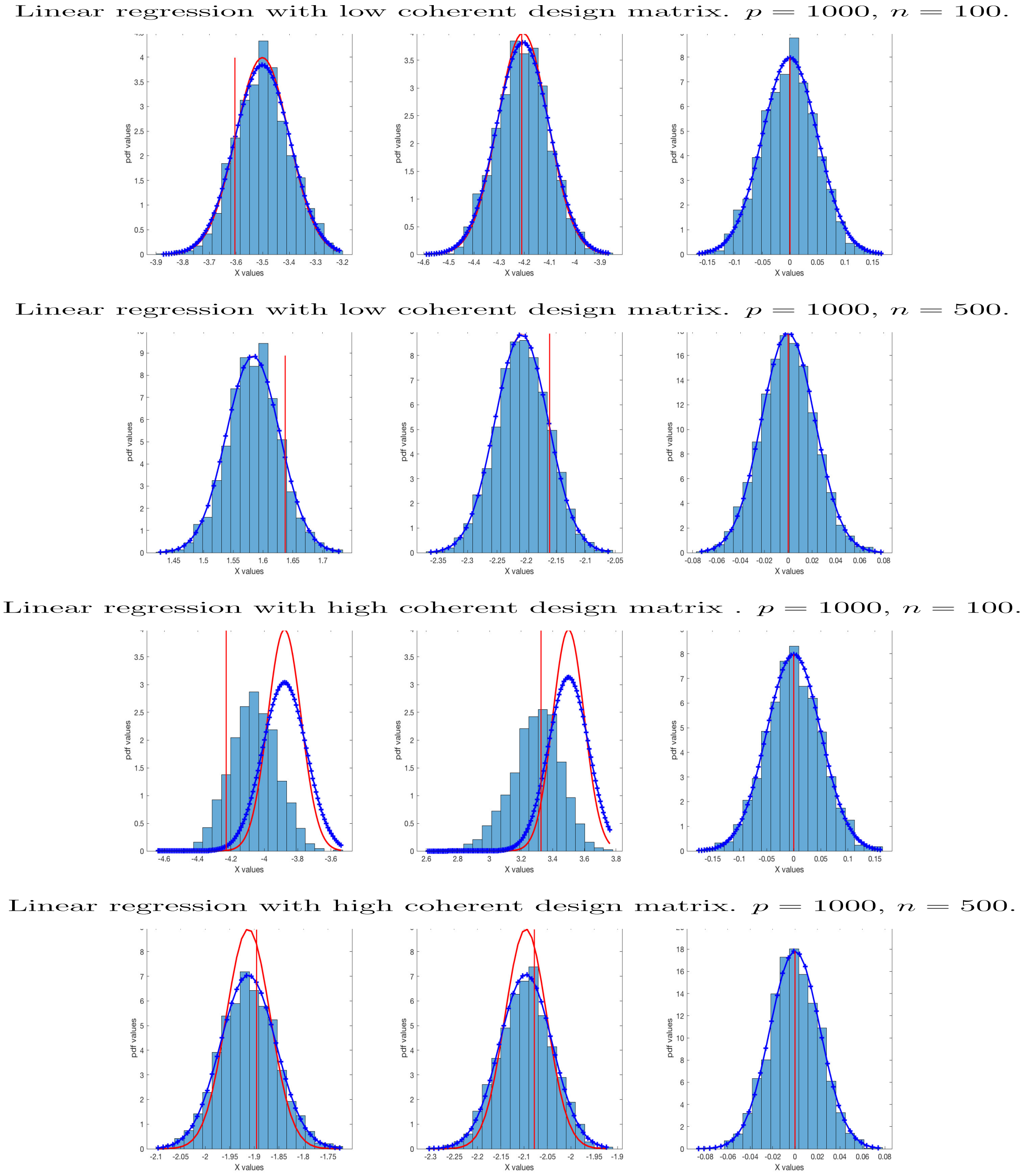
Figure 2.1: Computational cost for MCMC and VA

Costs of: p iterations of Metropolized Gibbs sampler (red solid line); 50 iterations of full-VA (blue+ line); and 50 iterations of midsize-VA with $\|\delta^{(\mathrm{i})}\|_0 = 100$ (blue-dashed line), as functions of the dimension p.

between components i and j,

where $\psi \in \{0, 0.8\}$. When $\psi = 0$, the resulting matrix X has a low coherence, but the coherence increases when $\psi = 0.8$. Using X, we general $Y = X\theta_\star + \epsilon$, with $\epsilon \sim N(0, 1)$. that we assume known. We build $\theta_\star$ with $s_\star = 10$ non-zeros components that we fill with draws from the uniform distribution $\pm\mathbf{U}(a, a + 1)$, where $a = 4\sqrt{s_\star \log(p)/n}$.

We build Π with $\sigma^2 = 1$, $u = 2$, $\rho_1 = \sqrt{\log(p)/n}$, and $\rho_0^{-1} = 1/(4n)$. We sample from Π using Algorithm 2. We consider two variational approximation. The full-VA, and a mid-size VA with template $\delta^{(\mathrm{i})}$ that contains the support of $\theta_\star$, and such that $\|\delta^{(\mathrm{i})}\|_0 = 100$. We approximate the variational approximations by coordinate ascent variational inference (see e.g. Blei et al. [2017]). The details of these algorithms are given in Appendix 2.6. We initialize all three algorithms from the lasso solution. In Figure 2.1 we plot the computational cost of the three algorithms as p increases. It shows that the full-VA is actually more expensive than the MCMC sampler. This is due to the need to form the Cholesky decomposition of a large $p \times p$ matrix at each iteration of the full-VA. In contrast, and as explained in Section 2.2.1 the per-iteration cost of Algorithm 2 is of order $O(s_\star p)$. On the other hand, for $p = 5,000$ the midsize VA is more than 10 times faster than the MCMC sampler. Figure 2.2 shows the (estimated) posterior distributions for the parameters θ_1, θ_2 and θ_3 from one MCMC run of $5,000$ iterations and single CAVI-runs of 50 iterations. Here we are comparing the skinny-VA, and the midsize-VA with $\|\delta^{(\mathrm{i})}\|_0 = 100$, for a template $\delta^{(\mathrm{i})}$ that contains the support of $\theta_\star$. Since we are working in a high signal-to-noise ratio setting the results are fairly consistent across replications. The true signal $\theta_\star$ is such that $\theta_{\star,1} \neq 0$ and $\theta_{\star,2} \neq 0$ while $\theta_{\star,3} = 0$. Figure 2.2 shows that as n increases both VA approximations approximate well the quasi-posterior distribution in the low coherence regime. However in presence of correlation, the skinny-VA systematically underestimates the marginal posterior variances when there is correlation between the relevant variables. However, as suggested by Corollary 6, the midsize-VA approximates the whole distribution well.



Linear regression with low coherent design matrix. $p = 1000$, $n = 100$.

Linear regression with low coherent design matrix. $p = 1000$, $n = 500$.

Linear regression with high coherent design matrix . $p = 1000$, $n = 100$.

Linear regression with high coherent design matrix. $p = 1000$, $n = 500$.

Figure 2.2: Posterior inference: Linear regression

Posterior inference for β_1 (first column), β_2 (second column) and β_3 in the linear regression example based on one MCMC run (histogram), one skinny-VA run (continuous red line), and one midesize-VA run (+ blue line). Vertical lines locate the true values of the parameters.

2.5.2 Gaussian graphical models via Linear regressions

Fitting large sparse graphical models in the Bayesian framework is computationally challenging (Dobra et al. [2011], Lenkoski and Dobra [2011], Khondker et al. [2013], Peterson et al. [2015], Banerjee and Ghosal [2013]). A quasi-Bayesian approach based on the neighborhood selection of Meinshausen and Buhlmann [2006] offers a simple, yet effective alternative. The idea was explored in Atchadé [2019] using point-mass spike and slab priors. The approach proposed in this dissertation yields a highly scalable quasi-posterior distribution with equally strong theoretical backing. We make the following data generating assumption.

B 1. *$Z \in \mathbb{R}^{n\times(p+1)}$ is a random matrix with i.i.d. rows from $\boldsymbol{N}_{p+1}(0, \vartheta_\star^{-1})$ for some positive definite matrix $\vartheta_\star$. We set $\Sigma \overset{\text{def}}{=} \vartheta_\star^{-1}$ and also assume that as $p \to \infty$,*

$$\frac{1}{\lambda_{min}(\Sigma)} + \lambda_{max}(\Sigma) = O(1). \tag{2.5.7}$$

Remark II.8. The assumption in (2.5.7) restricts our focus to problems that in some sense do not become intrinsically harder as p increases. It can be relaxed by tracking more carefully the constants in the proofs.

□

Given the data matrix $Z \in \mathbb{R}^{n\times(p+1)}$, we wish to estimate the precision matrix $\vartheta_\star$. Instead of a full likelihood approach (explored in the references cited above), we consider a pseudo-likelihood approach that estimates each column of $\vartheta_\star$ separately. Given $1 \le j \le p+1$, we partition the data matrix Z as $Z = [Y^{(j)}, X^{(j)}]$, where $Y^{(j)} \in \mathbb{R}^n$ denotes the j-th column of Z, and $X^{(j)} \in \mathbb{R}^{n\times p}$ collects the remaining columns. In that case the conditional distribution of $Y^{(j)}$ given $X^{(j)}$ is

$$\mathbf{N}_n\left(X^{(j)}\theta_\star^{(j)}, \frac{1}{[\vartheta_\star]_{jj}} I_n\right),$$

where $\theta_\star^{(j)} \stackrel{\text{def}}{=} (-1/[\vartheta_\star]_{jj})[\vartheta_\star]_{-j,j} \in \mathbb{R}^p$. Therefore, for some user-defined parameters $\sigma_j > 0$, $\rho_{0,j} > 0$, and $\rho_{1,j}$ the quasi-posterior distribution on $\Delta \times \mathbb{R}^p$ given by

$$\Pi^{(j)}(\delta, \mathrm{d}\theta|Z) \propto \\ e^{-\frac{1}{2\sigma_j^2}\|Y^{(j)}-X^{(j)}\theta_\delta\|_2^2}\omega(\delta)\left(\frac{\rho_{1,j}}{2\pi}\right)^{\frac{\|\delta\|_0}{2}}\left(\frac{\rho_{0,j}}{2\pi}\right)^{\frac{p-\|\delta\|_0}{2}} e^{-\frac{\rho_{1,j}}{2}\|\theta_\delta\|_2^2} e^{-\frac{\rho_{0,j}}{2}\|\theta-\theta_\delta\|_2^2}\mathrm{d}\theta, \quad (2.5.8)$$

can be used to estimate $\theta_\star^{(j)}$, and hence the j-th column of $\vartheta_\star$, if an estimate of $[\vartheta_\star]_{jj}$ is available[1]. This is basically the quasi-Bayesian analog of the neighborhood selection of Meinshausen and Buhlmann [2006]. The same procedure can be repeated – possibly in parallel – to recover the entire matrix $\vartheta_\star$. We use the theory of Section 2.2-2.4 to describe the behavior of this approach to infer $\vartheta_\star$. We focus on the case where $n = o(p)$, and we recall that C_0 is an absolute constant whose value may be different from one expression to the other. Let $\Pi_\star^{(j,\infty)}$ be the corresponding limiting distribution of $\Pi^{(j)}$ as defined in (2.4.1), and let $\tilde{\Pi}_\star^{(j,\infty)}$ be the corresponding approximation given in (2.4.6). In this particular case, $\Pi_\star^{(j,\infty)}$ is the probability measure on $\Delta \times \mathbb{R}^p$ that puts probability one on $\delta_\star^{(j)}$ (the support of $\theta_\star^{(j)}$), draws $[\theta]_{\delta_\star^{(j)}} \sim \mathbf{N}\left(\hat{\theta}_\star^{(j)}, \sigma_j^2(X'_{\delta_\star^{(j)}} X_{\delta_\star^{(j)}})^{-1}\right)$, and draws independently all other components i.i.d. from $\mathbf{N}(0, \rho_0^{-1})$, where $\hat{\theta}_\star^{(j)}$ is the OLS estimator $(X_{\delta_\star^{(j)}} X_{\delta_\star^{(j)}})^{-1} X'_{\delta_\star^{(j)}} Y^{(j)}$. We set $s_\star^{(j)} \stackrel{\text{def}}{=} \|\theta_\star^{(j)}\|_0$. Let $Q^{(j)}$ denote the variational approximation of $\Pi^{(j)}$ based on the family (2.4.4) with sparsity pattern $\mathcal{S}^{(j)}$, and let ζ_j denote the corresponding

[1] A full Bayesian approach can be adopted to estimate both $\theta_\star^{(j)}$ and $[\vartheta_\star]_{jj}$. But for simplicity's sake we will not pursue this here

term in (2.4.8).

Corollary 7. *Assume H2, B1, and suppose that* $s_\star^{(j)} > 0$, $\max_j \|\theta_\star^{(j)}\|_\infty = O(1)$, *and* $\max_j s_\star^{(j)} = O(\log(p))$ *as* $p \to \infty$. *Suppose also that* $u > 2$, *and* $u\sigma_j^2[\vartheta_\star]_{jj} \geq 16$. *Choose the prior parameter* $\rho_{1,j}$ *as*

$$\rho_{1,j} = \sqrt{\frac{\log(p)}{n}}.$$

Set

$$\bar{s}^{(j)} \overset{\text{def}}{=} s_\star^{(j)}\left(1 + \frac{6}{u}\right) + \frac{u}{4}, \quad \epsilon^{(j)} \overset{\text{def}}{=} C_0\sqrt{\frac{(\bar{s}^{(j)} + s_\star^{(j)})\log(p)}{[\vartheta_\star]_{jj}\; n}}, \quad \textit{and} \quad \bar{s} = \max_j \bar{s}^{(j)}.$$

Suppose that the sample size n *satisfies* $n = o(p)$, *as* $p \to \infty$, *and*

$$n \geq C_0 \bar{s}\log(p),$$

and the strong signal assumption

$$\min_{k\colon |\theta_{\star,k}^{(j)}|>0} |\theta_{\star,k}^{(j)}| > C_0\epsilon^{(j)} \tag{2.5.9}$$

holds. Then there exists a measurable set $\mathcal{G}$ *with* $\mathbb{P}_\star(Z \notin \mathcal{G}) \to 0$ *as* $p \to \infty$ *such that*

$$\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{G}}(Z) \max_{1\leq j\leq p+1} \mathsf{KL}\left(\Pi_\star^{(j,\infty)}|\Pi^{(j)}\right)\right] \leq \frac{C_0 \max_j(\bar{s}^{(j)} + s_\star^{(j)})}{\min_j[\vartheta_\star]_{jj}}\frac{\log(p)}{n} + \frac{C_0}{p^{1\wedge\left(\frac{u}{2}-1\right)}}. \tag{2.5.10}$$

Furthermore the variational approximation $Q^{(j)}$ satisfies

$$\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{G}}(Z)\max_{1\leq j\leq p+1}\|Q^{(j)}-\tilde{\Pi}_\star^{(j,\infty)}\|_{\mathrm{tv}}^2\right]\leq 8\mathbb{E}_\star\left[\boldsymbol{1}_{\mathcal{G}}(Z)\max_{1\leq j\leq p+1}\zeta^{(j)}\right] + \frac{C_0\max_j(\bar{s}^{(j)}+s_\star^{(j)})}{\min_j[\vartheta_\star]_{jj}}\frac{\log(p)}{n}+\frac{C_0}{p^{1\wedge\left(\frac{u}{2}-1\right)}}. \quad (2.5.11)$$

Proof. See Appendix A.7. □

Remark II.9. 1. We have focused in the Corollary on the Bernstein-von Mises approximation and the behavior of the VA approximation. Other results, and generally more precise results are given in the proof. In particular we show that the rate of contraction of $\Pi^{(j)}$ is $\epsilon^{(j)}$, and that $\Pi^{(j)}$ achieves perfect model selection.

2. One cannot easily remove the indicator $\mathbf{1}_{\mathcal{G}}$ from (2.5.10). However by Pinsker's inequality we get

$$2\mathbb{E}_\star\left[\max_{1\leq j\leq p+1}\|\Pi_\star^{(j,\infty)}-\Pi^{(j)}\|_{\mathrm{tv}}^2\right]\leq 2\mathbb{P}_\star[Z\notin\mathcal{G}] + \frac{C_0\max_j(\bar{s}^{(j)}+s_\star^{(j)})}{\min_j[\vartheta_\star]_{jj}}\frac{\log(p)}{n}+\frac{C_0}{p^{1\wedge\left(\frac{u}{2}-1\right)}}.$$

3. If the variational approximation $Q^{(j)}$ is constructed from some template $\delta^{(\mathsf{i},j)}$, then the remainder $\zeta^{(j)}$ is zero if $\delta^{(\mathsf{i},j)}\supseteq\delta_\star^{(j)}$. When this is the case we also have $\tilde{\Pi}_\star^{(j,\infty)}=\Pi_\star^{(j,\infty)}$. This holds for instance if $\delta^{(\mathsf{i},j)}$ is the vector with all components equal to 1 (full-VA). However the full-VA is expensive to compute. In fact, as we illustrate below the full-VA is more expensive to compute than direct MCMC sampling from $\Pi^{(j)}$. However if $\delta^{(\mathsf{i},j)}$ is sparse, for instance if $\delta^{(\mathsf{i},j)}$ is the support of the lasso solution – or some equally well-behaved

frequentist estimate – then the scaling of the computational cost of $Q^{(j)}$ can be extremely favorable. Hence Corollary implies that extremely fast variational approximation of $\Pi^{(j)}$ with strong theoretical guarantees can be computed in large scale Gaussian graphical models.

□

2.5.3 Logistic Regression

In this example we shall study the behavior of the posterior distribution in case of logistic regression. In frequentist high dimensional setup, the behavior of the logistic regression has been previously studied in (Abramovich and Grinshtein [2019], Sur and Candés [2019]). The interest stems from the fact that for discrete Pairwise Markov Random Fields (PMRFs) such as the Ising model (Ising [1925]), using a pseudo-likelihood approach results in performing logistic regression on each node given the other nodes in the graph. In this example we will specifically study the bounds obtained from Theorems 2, 3 and 4 for logistic regression.

C 1. *Let $Y \in \{0,1\}^n$ be a vector of independent observations with*

$$\mathbb{P}(Y_i = 1|x_i) = \frac{\exp\left(\langle x_i, \theta_\star \rangle\right)}{1 + \exp\left(\langle x_i, \theta_\star \rangle\right)}. \tag{2.5.12}$$

$\theta_\star \in \mathbb{R}^p$ is the true generating parameter with $\|\theta_\star\|_0 = s_\star$. $X \in \mathbb{R}^{n \times p}$ is the covariate matrix with columns $X_j, j = 1, \cdots, p$ such that $\max\limits_{j=1,\cdots,p} \|X_j\|_\infty \leq \mathbf{b}$, where $\mathbf{b} > 0$ is some absolute constant.

The likelihood and score functions are defined as

$$\ell(\theta; z) = \exp\left(\sum_{i=1}^n y_i \langle x_i, \theta \rangle - g(\langle x_i, \theta \rangle)\right)$$

$$\text{and} \quad \nabla \ell(\theta_\star; z) = X'y - \sum_{i=1}^n x_i' g^{(1)}(\langle x_i, \theta_\star \rangle)$$

respectively where $g(\langle x_i, \theta \rangle) = \log(1 + \exp(\langle x_i, \theta \rangle))$. The information matrix at $\theta_\star$ can then be defined as

$$\nabla^{(2)} \ell(\theta_\star; z) \overset{\text{def}}{=} X' W(\theta_\star) X,$$

where $W^{n\times n}(\theta_\star) = diag\left(g^{(2)}(\langle x_1, \theta_\star\rangle), \cdots g^{(2)}(\langle x_n, \theta_\star\rangle)\right)$. We also define the following restricted eigenvalues

1. $$\bar{v}(s) \overset{\text{def}}{=} \sup\left\{\frac{u'X'Xu}{n\|u\|_2^2},\ u \neq 0,\ u \in \mathbb{R}^p,\ \|u\|_0 \leq s\right\}.$$

2. $$\underline{v}(s) \overset{\text{def}}{=} \inf\left\{\frac{u'X'W(\theta_\star)Xu}{n\|u\|_2^2},\ u \neq 0,\ u \in \mathbb{R}^p,\ \|u\|_0 \leq s\right\}.$$

3. $$\underline{v}_1(\bar{s}) \overset{\text{def}}{=} \min_{\delta\in\mathcal{A}_{\bar{s}}} \inf_{\theta\in\mathbb{R}^{\|\delta\|_0}:\ \|\theta-\hat{\theta}_\delta\|_2\leq 2\epsilon} \inf\left\{\frac{u'X'W(\theta)Xu}{n\|u\|_2^2},\ u \in \mathbb{R}^{\|\delta\|_0},\ u \neq 0\right\}.$$

4. $$\underline{v}_2(\bar{s}) \overset{\text{def}}{=} \min_{\delta\in\mathcal{A}_{\bar{s}}} \inf_{\theta\in\mathbb{R}^{\|\delta\|_0}} \inf\left\{\frac{u'X'W(\theta)Xu}{n\|u\|_2^2},\ u \in \mathbb{R}^{\|\delta\|_0},\ u \neq 0\right\}.$$

The nature of the likelihood in logistic regression does not make the application of Theorem 1 on posterior sparsity obvious when $(1 - \frac{\rho_1}{\bar{\rho}}) \sim O(1)$. This may be a construction of the proof and as a result we cannot assume sparsity in the posterior automatically. To circumvent this, we modify the Bernoulli prior on the δ from H(2) to encode strict sparsity.

C 2. *We assume that*

$$\bar{\omega}_\delta \propto \mathsf{q}^{\|\delta\|_0}(1-\mathsf{q})^{p-\|\delta\|_0}; \quad \delta \in \Delta(\bar{s}),$$

where $\frac{q}{1-q} = \frac{1}{p^{u+1}}$. *Here* $u > 0$ *is an absolute constant and* $\Delta(\bar{s}) = \{\delta \in \Delta;\ \|\delta\|_0 \leq \bar{s}\}$. *We further assume that* $2s_\star \geq \bar{s} \geq s_\star$.

Remark II.10.

1. The prior in Assumption C2 restricts the probability mass function to sparse subsets of Δ where the sparsity is encoded by the pre-specified quantity $\bar{s}$ and ideally $\bar{s} \geq s_\star$. The drawback of such a prior is that if we do not have information on the true sparsity $\bar{s}$ is difficult to choose and is left up to the judgment of the researcher.

2. Simulation results for logistic regression show posterior sparsity even under prior H2. This further indicates that the lack of posterior sparsity results is due to the construction of the proof.

The resultant posterior distribution for (θ, δ) given $Z = (Y, X)$ can then be expressed as

$$\Pi(\delta, d\theta|z) \propto \exp\left(\sum_{i=1}^{n} y_i\langle x_i, \theta\rangle - g(\langle x_i, \theta\rangle)\right) \\ \times \bar{\omega}_\delta \left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \left(\frac{\rho_0}{2\pi}\right)^{\frac{p-\|\delta\|_0}{2}} \left[\prod_{j\colon \delta_j=1} e^{-\frac{\rho_1\theta_j^2}{2}}\right] \left[\prod_{j\colon \delta_j=0} e^{-\frac{\rho_0\theta_j^2}{2}}\right] d\theta. \quad (2.5.13)$$

We make the final following assumption related to the coherence of the design matrix.

C 3. *Define*

$$\mathsf{R} = \frac{\max_{j\in\delta_\star, k\in\delta_\star^c}\langle X_j, W(\theta_\star)X_k\rangle}{n}.$$

Assume

$$\frac{\bar{s}\mathsf{R}}{\underline{v}(\bar{s})} \leq 1/2 \quad \textit{and} \quad \frac{\bar{v}(\bar{s})}{\underline{v}(\bar{s})} + \frac{\bar{v}(\bar{s})}{\underline{v}_2(\bar{s})} \sim O(1).$$

Remark II.11. Note that the quantity R is bounded above by $\frac{1}{n}\|X_{j\cdot}\|_2\|X_{k\cdot}\|_2$. Hence by Assumption C1, R is bounded above by $\mathbf{b}^2$. Thus assumption C3 prevents the problem from becoming intrinsically harder with growing dimensions by imposing bounds on the restricted eigenvalues and is usually difficult to verify.

Corollary 8. *Under assumptions C1-C3, suppose that $s_\star > 0$ and $\|\theta_\star\|_\infty = O(1)$. Choose*

$$u > 2, \;\; \rho_1 \sim \sqrt{\log(p)/n} \;\; \textit{and} \;\; \bar{\rho} = 4\mathbf{b}\sqrt{n\log(p)}.$$

We require $n > \max\left(\frac{1}{\underline{v}_1(\bar{s})^2}\left(\bar{s}^2\log(p)\right)^3, \left((16/3)\mathbf{b}^2\frac{(s_\star+\bar{s})}{\underline{v}(s_\star+\bar{s})}\sqrt{\log(p)}\right)^2\right)$.
We define

$$\epsilon \overset{\text{def}}{=} \frac{16\mathbf{b}}{\underline{v}(s_\star+\bar{s})}\sqrt{\frac{(s_\star+\bar{s})\log(p)}{n}}$$

and assume strong signal given by

$$\min_{k:\,|\theta_{\star k}|>0} |\theta_{\star k}| > C_0\epsilon.$$

Then we can find a set $\mathcal{G}$ with $P(Y \in \mathcal{G}|X) \to 0$ as $p \to \infty$ for which the following bounds hold.

1. *For $C > 3$ and*

$$\mathsf{B} \overset{\text{def}}{=} \bigcup_{\delta\in\Delta_{\bar{s}}} \{\delta\} \times \left\{\theta \in \mathbb{R}^p : \; \|\theta_\delta - \theta_\star\|_2 \le C\epsilon, \; \|\theta - \theta_\delta\|_2 \le 3\sqrt{\rho_0^{-1}p}, \right\}$$

   we can have , such that

$$\mathbb{E}_\star[\mathbf{1}_{\mathcal{G}}\Pi(\mathsf{B}^c|Z)|X] \le 8\exp[-C\frac{s_\star+\bar{s}}{2\underline{v}(s_\star+\bar{s})}\mathbf{b}^2\log(p)] + 2e^{-p}.$$

2.

$$\mathbb{E}_\star[\mathbf{1}_{\mathcal{G}}\Pi(\mathsf{B}_k^c|Z)|X] \le C_0\left(\sqrt{\frac{\rho_1}{n\underline{v}_1(\bar{s})}}\frac{1}{p^{u/2}}\right)^{k+1} + \mathbb{E}_\star[\mathbf{1}_{\mathcal{G}}\Pi(\mathsf{B}^c|Z)|X].$$

3.

$$\mathbb{E}_\star(1_{\mathcal{G}_2}\mathsf{KL}\left((|\Pi)_\star^{(\infty)}|\Pi)|X\right) \leq C_0\frac{(s_\star+\bar{s})\log(p)}{n\underline{v}(s_\star+\bar{s})}+C_1 n\mathbf{b}^3(\bar{s}+s_\star)^3\left(\frac{\log(p)}{n}\right)^{(3/2)} \\ + 2\mathbb{E}_\star[\boldsymbol{1}_{\mathcal{G}}(z)(1-\Pi(\delta_\star|z))].$$

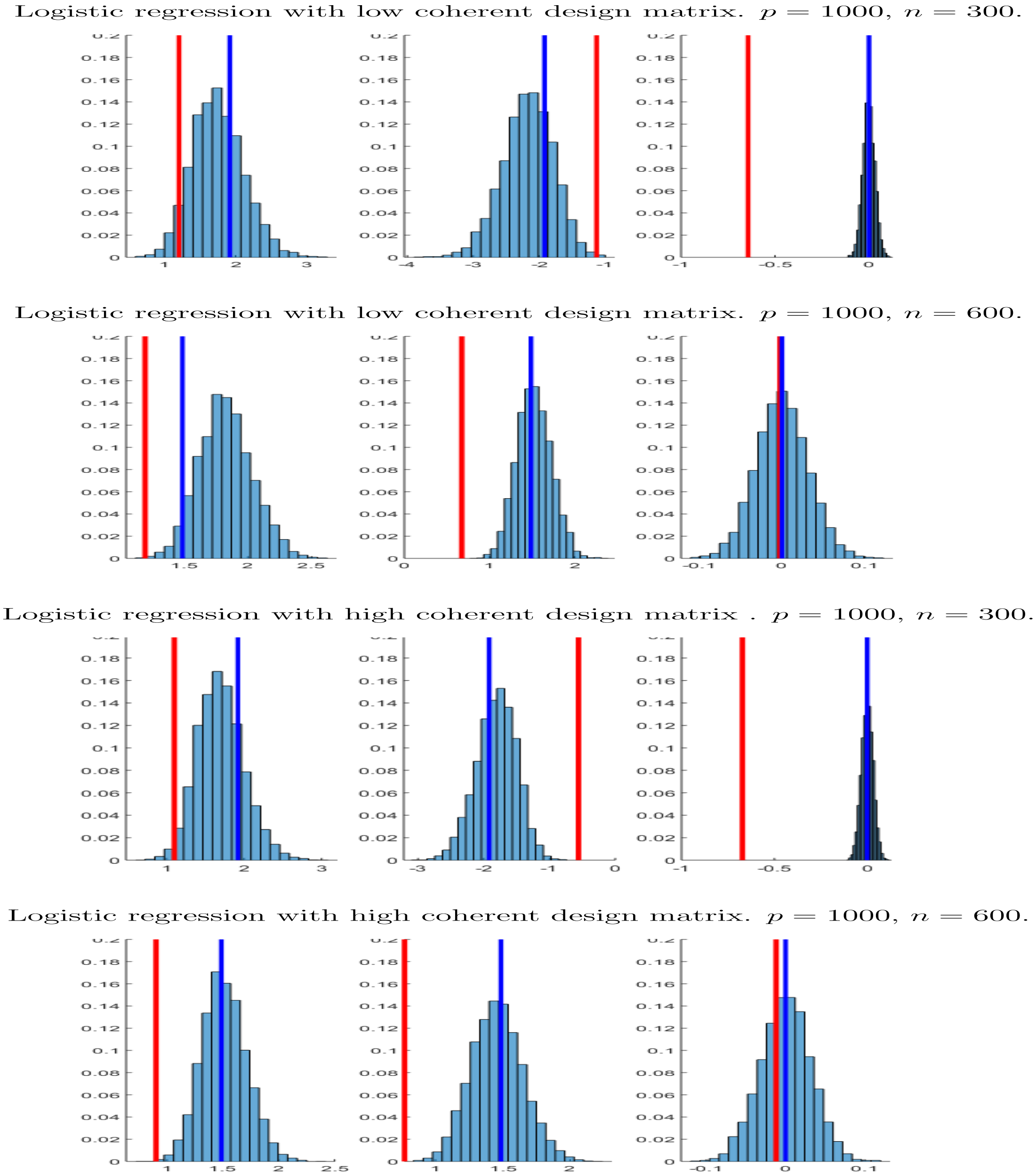
Proof. See Appendix A.9 □

The corollary covers contraction, model selection consistency and Bernstein phenomenon respectively. The restricted eigenvalue conditions (specifically conditions on $\underline{v}_2(\bar{s})$) are hard to verify in practice. The conditions have been imposed to simplify the bounds obtained and further research is required to show whether they can be relaxed.

2.5.3.1 Numerical illustration

We perform a simulation study to assess the behavior of the posterior distribution as described in Corollary 8. We set $p = 1000$, $n \in \{300, 600\}$, and we generate $Z = [Y, X] \in \{0,1\}^n \times \mathbb{R}^{n\times(p)}$ as follows. We first generate the matrix X by simulating the rows of X independently from a Gaussian distribution with correlation $\psi^{|j-i|}$ between components i and j, where $\psi \in \{0, 0.5\}$. When $\psi = 0$, the resulting matrix X has a low coherence, but the coherence increases when $\psi = 0.5$. Using X, we generate $Y_i = Ber\left(\frac{\exp(\langle\theta_\star,x_i\rangle)}{1+\exp(\langle\theta_\star,x_i\rangle)}\right)$. We build $\theta_\star$ with $s_\star = 10$ non-zeros components that we fill with draws from the uniform distribution $\pm\mathbf{U}(a, a+1)$, where $a = 4\sqrt{s_\star\log(p)/n}$.

We build Π with $u = 2$, $\rho_1 = \sqrt{\log(p)/n}$, and $\rho_0^{-1} = 1/p$. We sample from Π using Algorithm 5. We initialize the algorithm from the lasso solution. Figure 2.3 shows the (estimated) posterior distributions for the parameters θ_1, θ_2 and θ_3 from one MCMC run of $5,000$ iterations. Since we are working in a high signal-to-noise ratio setting the results are fairly consistent across replications. The true signal $\theta_\star$ is such that $\theta_{\star,1} \neq 0$ and $\theta_{\star,2} \neq 0$ while $\theta_{\star,3} = 0$. Figure 2.3 shows that the distribution covers the true value. The MCMC samples show lower variability when the design matrix has a high coherence.


Logistic regression with low coherent design matrix. $p = 1000$, $n = 300$.

Logistic regression with low coherent design matrix. $p = 1000$, $n = 600$.

Logistic regression with high coherent design matrix . $p = 1000$, $n = 300$.

Logistic regression with high coherent design matrix. $p = 1000$, $n = 600$.

Figure 2.3: Posterior inference: Logistic Regression

Posterior inference for θ_1 (first column), θ_2 (second column) and θ_3 in the logistic regression example based on one MCMC run (histogram). Vertical blue lines locate the true values of the parameters. Vertical red lines locate the lasso solution.

A further comparison is performed to study the convergence rate (using the relative error and sparsified relative error) and the recovery measured in terms of F1 score for increasing p in figure 2.4. For this particular comparison we have a low coherence design matrix and set $n \sim O(\log(p)^3)$ and number of non-zero components to the order of $\sqrt{\log(p)}$. We fix $\rho_0^{-1} \sim \frac{1}{p}$ and the results verify the fact that the mixing is slower for lower value of ρ_0^{-1}. It also illustrates the fact that the relative error rates are comparable after MCMC has converged and for strong signal settings we can achieve perfect recovery starting from the lasso solution.

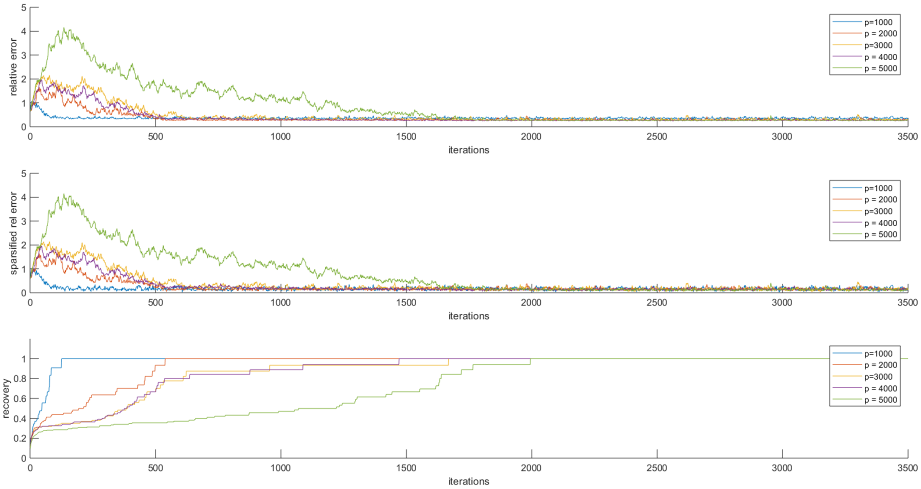


Figure 2.4: Error rates and recovery: Logistic Regression

The top panel shows the relative error defined as $\frac{\|\theta-\theta_\star\|_2}{\|\theta_\star\|_2}$. The second panel shows sparsified relative error defined as $\frac{\|\theta_\delta-\theta_\star\|_2}{\|\theta_\star\|_2}$. The bottom panel shows the F1 score defined as harmonic mean of precision and recall

2.5.4 Sparse principal component estimation

We give another illustration of the quasi-Bayesian framework with a non-standard example from sparse PCA. Principal component analysis is a widely used technique for data exploration and data reduction (Jolliffe [1986]). In order to deal with high-dimensional datasets, several works have introduced recently various versions of PCA that estimate sparse principal components (Jolliffe et al. [2003], Zou et al. [2006], Shen and Huang [2008], Lei and Vu [2015]). Extension of these ideas to a full Bayesian setting has been considered in the literature but is computationally challenging (Pati et al. [2014], Gao and Zhou [2015], Xie et al. [2018]). Using the quasi-Bayesian framework we explore here a fast regression-based approach to sparse PCA that we show works well when the sample size n is close to p and/or the spectral gap is sufficiently large. We consider the following data generating process.

D1. *The matrix $X \in \mathbb{R}^{n\times p}$ is such that the rows of X are i.i.d. from the Gaussian distribution $\boldsymbol{N}_p(0,\Sigma)$ on $\mathbb{R}^p$, with a covariance matrix Σ of the form*

$$\Sigma = \vartheta\theta_\star\theta_\star' + I_p,$$

for some sparse unit-vector $\theta_\star \in \mathbb{R}^p$, and some absolute constant $\vartheta > 0$. We set $s_\star \stackrel{\text{def}}{=} \|\theta_\star\|_0$.

Let $X = U\Lambda V'$ be the singular value decomposition (SVD) of X. Let V_1 be the first column of V. It was noted by Zou et al. [2006] that setting $y = \Lambda_{11}U_1$, it holds for all $\lambda > 0$ that

$$V_1 = \frac{\hat{b}}{\|\hat{b}\|_2}, \quad \text{where} \quad \hat{b} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

This result suggests that one can recover the first principal component V_1 by sparse regression of $y = \Lambda_{11}U_1$ on X. To implement this idea in a Bayesian framework we are naturally led to the quasi-likelihood function

$$\ell(\theta; X) = -\frac{1}{2\sigma^2}\|y - X\theta\|_2^2, \quad \theta \in \mathbb{R}^p,$$

for some constant $\sigma^2 > 0$. The resulting quasi-posterior distribution on $\Delta \times \mathbb{R}^p$ is the same as in (2.5.1):

$$\Pi(\delta, \mathrm{d}\theta | Z) \propto e^{-\frac{1}{2\sigma^2}\|y - X\theta_\delta\|_2^2}\omega(\delta)\left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}}\left(\frac{\rho_0}{2\pi}\right)^{\frac{p-\|\delta\|_0}{2}} e^{-\frac{\rho_1}{2}\|\theta_\delta\|_2^2} e^{-\frac{\rho_0}{2}\|\theta - \theta_\delta\|_2^2}\mathrm{d}\theta.$$

We analyze this quasi-posterior distribution. One challenge here is that we do not possess a good understanding of the distribution of the quasi-score function $X'(\Lambda_{11}U_1 - X\theta_\star)/\sigma^2$ due to the intricate nature of the SVD decomposition. Hence Theorem 1 cannot be applied, and thus we do not know whether the quasi-posterior distribution is automatically sparse under the prior H2. We work around this issue by hard-coding sparsity directly in the prior as follows.

C4. *We assume that*

$$\omega(\delta) \propto \boldsymbol{q}^{\|\delta\|_0}(1-\boldsymbol{q})^{p-\|\delta\|_0}\boldsymbol{1}_{\Delta_{\bar{s}}}(\delta), \quad \delta \in \Delta,$$

for some integer $\bar{s} \geq s_\star$, *where* $\mathsf{q} \in (0, 1)$ *is such that* $\frac{\mathsf{q}}{1-\mathsf{q}} = \frac{1}{p^{u+1}}$, *for some absolute constant* $u > 0$. *Furthermore we will assume that* $p \geq 9$, $p^{u/2} \geq 2e^{2\rho}$.

Since $s_\star$ is not known, how to find $\bar{s}$ in practice that satisfies $\bar{s} \geq s_\star$ is not obvious, and would require some judgment from the researcher. However in terms of computations, using D4 instead of H2 implies only a minor change to the MCMC sampler in Algorithm 2[2]. For $a \in \mathbb{R}$, $\mathsf{sign}(a) = 1$ if $a \geq 0$, and -1 otherwise.

[2] in STEP 2, if $\delta_j^{(k)} = 0$ and $\iota = 1$, we propose to do the change only if $\|\delta^{(k)}\|_0 \leq \bar{s}$.

Corollary 9. *Assume D1, D4, and choose $\sigma^2 = \vartheta$, $\rho = \sqrt{\log(p)/n}$. Suppose that $\|\theta_\star\|_\infty = O(1)$, as $p \to \infty$. There exist absolute constants C_0, C such that for $n \geq C_0(\frac{p}{\vartheta} + \bar{s}\log(p))$, we have*

$$\lim_{p\to\infty} \mathbb{E}_\star \left[\boldsymbol{1}_{\{\textsf{\textit{sign}}(\langle V_1,\theta_\star\rangle)=1\}} \Pi\left(\mathsf{B}_{\theta_\star}|X\right) + \boldsymbol{1}_{\{\textsf{\textit{sign}}(\langle V_1,\theta_\star\rangle)=-1\}} \Pi\left(\mathsf{B}_{-\theta_\star}|X\right)\right] = 1,$$

where for $\theta_0 \in \{\theta_\star, -\theta_\star\}$,

$$\mathsf{B}_{\theta_0} \overset{\text{def}}{=} \bigcup_{\delta\in\Delta_{\bar{s}}} \{\delta\} \times \left\{ \theta \in \mathbb{R}^p : \ \|\theta_\delta - \theta_0\|_2 \leq C\vartheta \sqrt{\frac{\left(\frac{p}{\vartheta} + \log(p)\right)(\bar{s} + s_\star)}{n}}, \right.$$
$$\left. \|\theta - \theta_\delta\|_2 \leq 3\sqrt{\rho_0^{-1} p} \right\}.$$

Proof. See Appendix A.10. □

It is well-known that the principal component is identified only up to a sign, which is reflected in Corollary 9. The assumption $\sigma^2 = \vartheta$ is made for simplicity, since ϑ is typically unknown. To a certain extent the procedure is robust to a misspecification of σ^2.

The contraction rate suggests that the method would perform poorly if the sample size and the spectral gap are both small, which is confirmed in the simulations. One important limitation of Corollary 9 is that the convergence rate does not have the correct dependence on the spectral gap. This is most certainly an artifact of our method of proof.

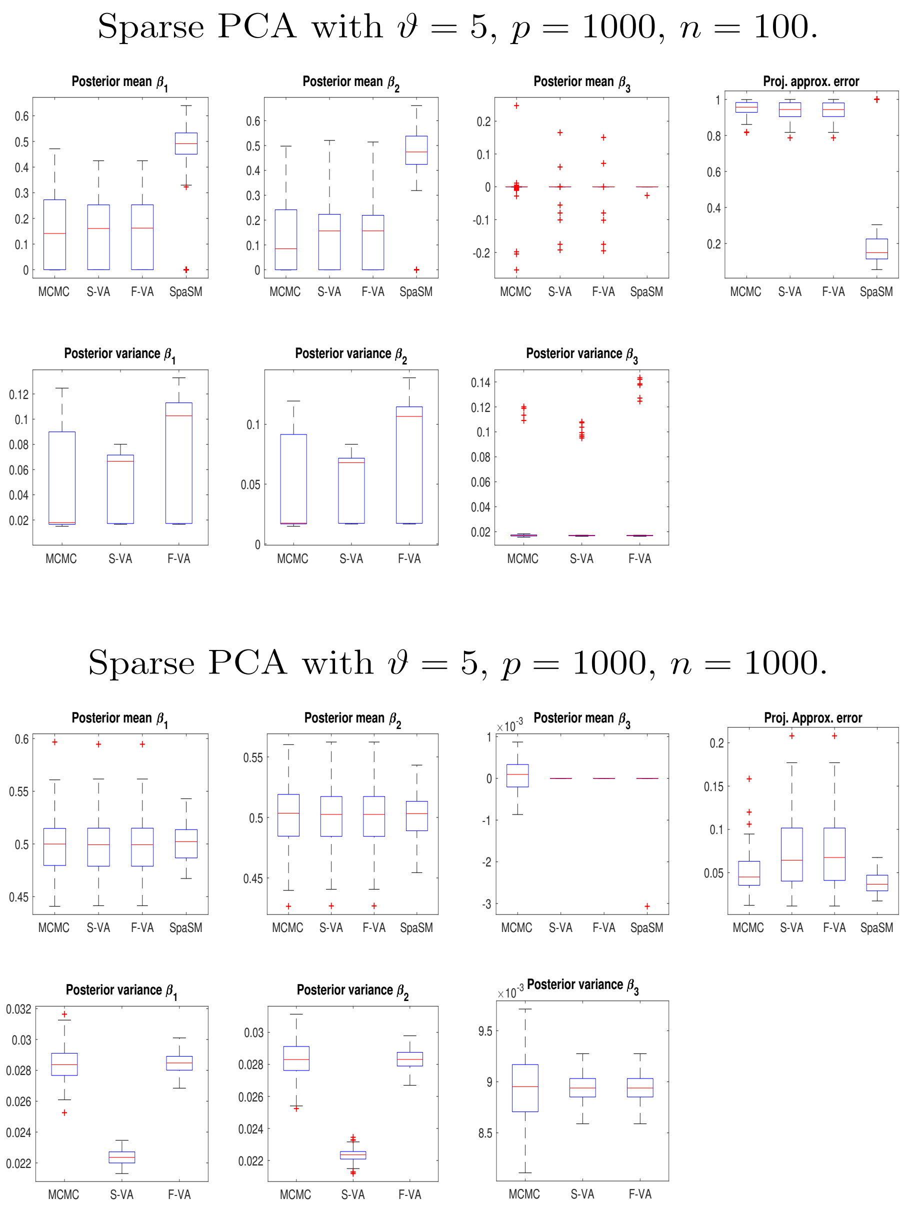
Corollary 9 does not cover model selection nor the approximation results. These results require a good control of the probability of the event $\mathcal{E}_2(\bar{s})$, which itself requires a better understanding of the distribution of singular vectors than we currently possess. We leave these issues for possible future research.

2.5.4.1 Numerical illustration

We generate a random matrix $X \in \mathbb{R}^{n\times p}$ according D1 with $p = 1000$, and $n \in \{100, 1000\}$, where $\beta_\star = (0.5, 0.5, 0, 0.5, 0.5, 0, \dots, 0)'$. We consider two levels of the spectral gap $\vartheta \in \{5, 20\}$. As above we set up the prior distribution with $u = 2$, $\rho_1 = \sqrt{log(p)/n}$, and $\rho_0^{-1} = 1/(4n)$. We use the same MCMC sampler as in the Gaussian graphical model of Section 2.5.1, that we initialize from the lasso solution, and run the 2000 iterations. We normalize the MCMC output to have unit-norm (at each iteration). We repeat all computations 100 times and use the replications to approximate the distribution of the posterior means and posterior variances of the first three components of θ (θ_1, θ_2 and θ_3). Using the 100 replications we also approximate the distribution of the error

$$\int \left\| \frac{\theta\theta'}{\|\theta\|_2^2} - \theta_\star\theta_\star' \right\|_2 \Pi(\mathrm{d}\theta|X),$$

that we call projection approximation error. To assess the quasi-likelihood method advocated here we compare its performance to that of the frequentist estimator of (Zou et al. [2006]) as implemented in the Matlab package SpaSM (Sjöstrand et al. [2018]). We present the results on Figure 2.5 and 2.6. The results supports very well the conclusions of Corollary 9.

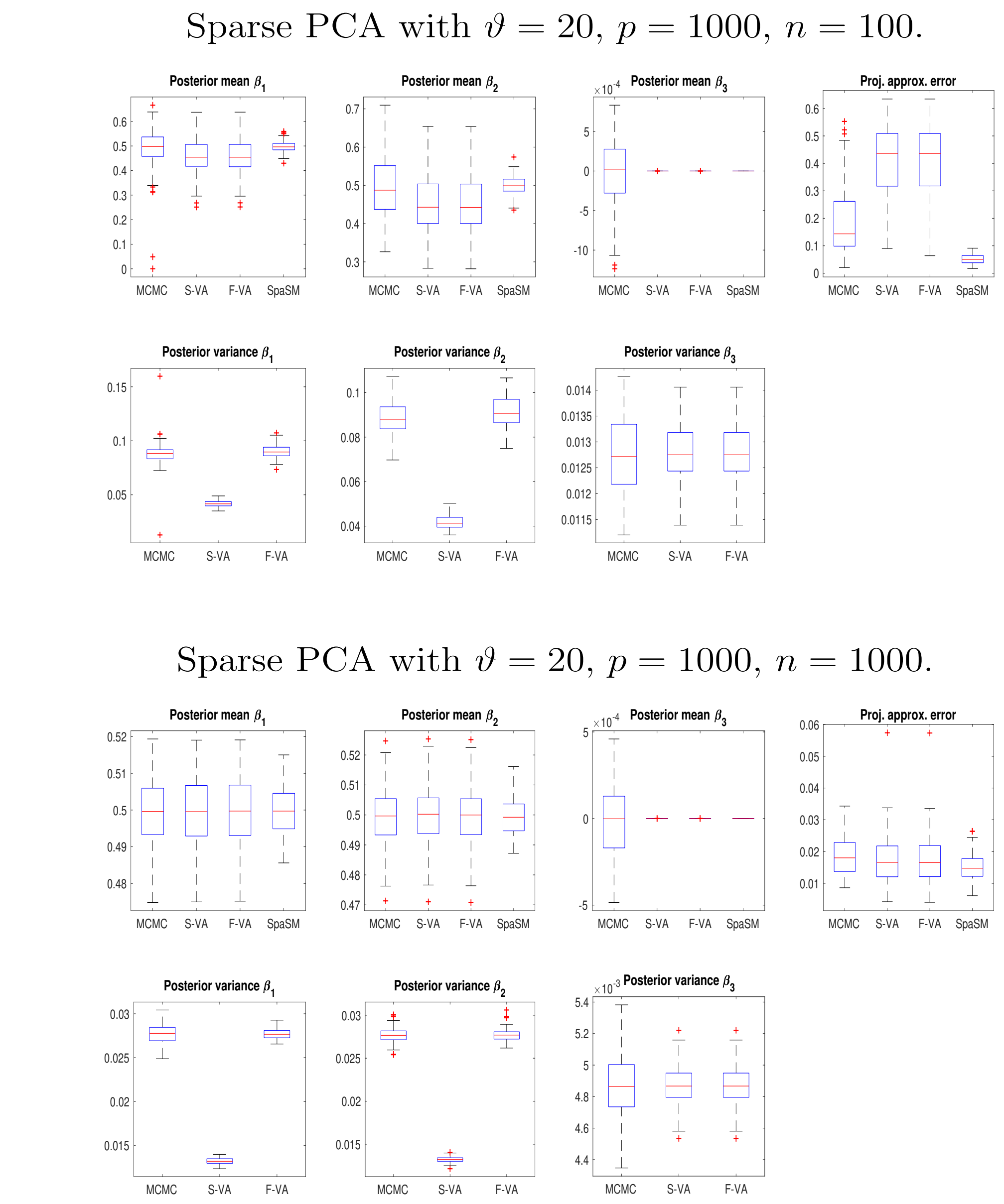


Sparse PCA with $\vartheta = 5$, $p = 1000$, $n = 100$.

Sparse PCA with $\vartheta = 5$, $p = 1000$, $n = 1000$.

Figure 2.5: Posterior inference: SParse PCA ($\vartheta = 5$)

Distributions of posterior means and variances for $\beta_1, \beta_2, \beta_3$, and distribution of the projection approx. error. Estimated from 100 replications. S-VA is skinny-VA, F-VA is full-VA. We also report similar distributions for the frequentist estimator computed by SpaSM.


Sparse PCA with $\vartheta = 20$, $p = 1000$, $n = 100$.

Sparse PCA with $\vartheta = 20$, $p = 1000$, $n = 1000$.

Figure 2.6: Posterior inference: SParse PCA ($\vartheta = 20$)

Distributions of posterior means and variances for $\beta_1, \beta_2, \beta_3$, and distribution of the projection approx. error. Estimated from 100 replications. S-VA is skinny-VA, F-VA is full-VA. We also report similar distributions for the frequentist estimator computed by SpaSM.

2.6 Algorithms for linear regression models

Both algorithms are initialized from the lasso solution and its support. The VA also needs an initial value of the matrix C which we take as $(c/n)I_p$, with $c = 0.001$.

Algorithm 2 (Gibbs sampler for (2.5.1))**.** At the k-th iteration, given $(\delta^{(k)}, \theta^{(k)})$:

1. For all j such that $\delta_j^{(k)} = 0$, draw $\theta_j^{(k+1)} \sim \mathbf{N}(0, \rho_0^{-1})$. Then draw jointly $[\theta^{(k+1)}]_\delta \sim \mathbf{N}(m^{(k)}, \Sigma^{(k)})$, where

$$m^{(k)} = \left(X'_{\delta^{(k)}} X_{\delta^{(k)}} + \sigma^2 \rho_1 I_{\|\delta^{(k)}\|_0}\right)^{-1} X'_{\delta^{(k)}} z, \quad \Sigma^{(k)} = \sigma^2 \left(X'_{\delta^{(k)}} X_{\delta^{(k)}} + \sigma^2 \rho_1 I_{\|\delta^{(k)}\|_0}\right)^{-1}.$$

2. (a) Given $\theta^{(k+1)} = \theta$, set $\delta^{(k+1)} = \delta^{(k)}$, and repeat for $j = 1, \ldots, p$. Draw $\iota \sim \mathbf{Ber}(0.5)$. If $\delta_j^{(k)} = 0$, and $\iota = 1$, with probability $\min(1, A_j)/2$ change $\delta_j^{(k+1)}$ to ι. If $\delta_j^{(k)} = 1$, and $\iota = 0$, with probability $\min(1, A_j^{-1})/2$, change $\delta_j^{(k+1)}$ to ι; where

$$A_j = \frac{\mathsf{q}}{1-\mathsf{q}} \sqrt{\frac{\rho_1}{\rho_0}} e^{-(\rho_1-\rho_0)\frac{\theta_j^2}{2}} e^{-\frac{\theta_j^2}{2\sigma^2}\|X_j\|_2^2 + \frac{\theta_j}{\sigma^2}\left(\langle X_j, Y\rangle - \sum_{i:\ \delta_i^{(k+1)}=1,\ i\neq j} \theta_i \langle X_j, X_i\rangle\right)}.$$

Algorithm 3 (Midsize VA approximation for (2.5.1) using template $\delta^{(\mathsf{i})}$)**.** Given $\alpha^{(k)}, \mu^{(k)}$, and $C^{(k)}$

1. (a) Set $\bar{\alpha} = \alpha^{(k)}$. For $j = 1, \ldots, p$ update $\bar{\alpha}_j$ as $\bar{\alpha}_j = \frac{1}{1+R_j}$, where

$$R_j = \frac{1-\mathsf{q}}{\mathsf{q}} \sqrt{\frac{\rho_0}{\rho_1}} e^{(\rho_1-\rho_0)\frac{\widehat{\theta_j^2}}{2}} e^{\frac{1}{2\sigma^2}\left[\widehat{\theta_j^2}\|X_j\|_2^2 - 2\mu_j^{(k)}\left\langle X_j, y - \sum_{i\neq j} \mu_i^{(k)} \bar{\alpha}_i X_i\right\rangle + S_j\right]},$$

where $\widehat{\theta_j^2} = (\mu_j^{(k)})^2 + C_{jj}^{(k)}$, and $S_j = 2\sum_{i\neq j} \bar{\alpha}_i C_{ij} \langle X_j, X_i\rangle$.

   (b) Set $\alpha^{(k+1)} = \bar{\alpha}$.

2. (a) For each j such that $\delta_j^{(\mathsf{i})} = 0$, set

$$C_{jj}^{(k+1)} = \frac{1}{\left(\rho_1 + \frac{\|X_j\|_2^2}{\sigma^2}\right)\alpha_j^{(k+1)} + \rho_0(1-\alpha_j^{(k+1)})},$$

and

$$\mu_j = \frac{C_{jj}^{(k+1)}}{\sigma^2}\alpha_j^{(k+1)}\left\langle X_j, y - \sum_{i\neq j}\alpha_i^{(k+1)}\bar{\mu}_i X_i\right\rangle.$$

(b) If $\|\delta^{(\mathsf{i})}\|_0 > 0$ do the following. Set $\tilde{y} = y - \sum_{j:\delta_j^{(\mathsf{i})}=0}\alpha_j^{(k+1)}\mu_j^{(k+1)}X_j$. Form the matrix $M \in \mathbb{R}^{p\times p}$ such that $M_{ij} = \alpha_i^{(k+1)}\|X_i\|_2^2$, if $i = j$, and $M_{ij} = \alpha_i^{(k+1)}\alpha_j^{(k+1)}\langle X_i, X_j\rangle$ if $i \neq j$. Let $\Lambda \in \mathbb{R}^{p\times p}$ be the diagonal matrix such that $\Lambda_{jj} = \alpha_j^{(k+1)}\rho_1 + \rho_0(1-\alpha_j^{(k+1)})$. Then we update $C^{(k)}$ to

$$[C^{(k+1)}]_{\delta^{(\mathsf{i})},\delta^{(\mathsf{i})}} = \left(\left[\Lambda + \frac{1}{\sigma^2}M\right]_{\delta^{(\mathsf{i})},\delta^{(\mathsf{i})}}\right)^{-1},$$

and we update $\mu^{(k)}$ to

$$[\mu^{(k+1)}]_{\delta^{(\mathsf{i})}} = \left([C^{(k+1)}]_{\delta^{(\mathsf{i})},\delta^{(\mathsf{i})}}\right)\left[\mathsf{diag}(\alpha^{(k+1)})\right]_{\delta^{(\mathsf{i})},\delta^{(\mathsf{i})}} X'_{\delta^{(\mathsf{i})}}\tilde{y},$$

where $\mathsf{diag}(\alpha^{(k+1)})$ is the diagonal matrix with diagonal given by $\alpha^{(k+1)}$.

Remark II.12. Setting $\delta^{(\mathsf{i})} = \mathbf{0}_p$ in the algorithm above yields the mean field variational approximation (skinny-VA). And taking $\delta^{(\mathsf{i})}$ as the vector will all components equal to 1 yields the full variational approximation (full-VA).

CHAPTER III

A quasi-bayesian method for fitting Potts Model

3.1 Introduction

The second part of this dissertation is focused on implementing a tractable Bayesian quasi-likelihood based approach for fitting high-dimensional Potts or Ising models. We consider particularly the setting where the data can take only finitely many values. This is motivated by the widespread availability of this type of data in areas of psychology, image processing, computer science, social sciences, bioinformatics, to name a few. For instance, Banerjee et al. [2008] used an Ising model to find association between US senators from their binary voting records. Ekeberg et al. [2013] used a Potts model to predict contact between amino acids in protein chains. In Epskamp et al. [2017, 2018], the authors worked the reader through statistical procedures for estimating psychological networks in personality research. The Ising model Ising [1925] was originally formulated in the Physics literature as a simple model for interacting magnetic spins on a lattice. The Potts model is a generalization of the Ising model in which spins can take more than two values with more complex dependencies. These models are widely used in the applications for teasing out direct and undirected dependencies between large collections of nodes in a graph. The purpose of this work is to construct robust and scalable bayesian procedures for fitting these models. We focus on settings where the underlying network is sparse and we address the problem by introducing an auxiliary selection variable that represents the structure of the network. We use a form of quasi-likelihood that takes product of the conditional distribution of each node of the graph, conditioning on the other nodes and a spike and slab prior distribution that is separable across the nodes. In case of Ising models specifically, the resulting quasi-posterior distribution can then be written as a product of logistic regression posterior distributions that we sample independently. Using this approach on a multi-core system yields a significant reducing in computing time. Our method is roughly based on the theoretical findings in Chapter II. It simultaneously estimates the model parameters and the underlying structure of the graph. We develop scalable Markov Chain Monte Carlo (MCMC) algorithms that can be implemented in parallel thus significantly reducing the computational cost of the method. At the end our methodology provides the user with both point estimates and quasi-credible intervals for the model parameters. We run extensive simulations to check the accuracy of the method, and we illustrate its practical applicability using an example from personality research.

The rest of the chapter is organized as follows: Section 3.2 introduces the methodology. In Section 3.3, we propose two scalable MCMC algorithms to deal with the resulting quasi-posterior distribution. Section 3.4 illustrates the performance of our method through simulation results. Finally, in Section 3.5, we present an application of our method in the context of psychological data through the analysis of the 16 Personality Factors (16PF) dataset.

3.2 Quasi-posterior distribution of the Potts model under spike and slab prior

An m-colored Potts model parametrized by a sparse symmetric matrix θ is a probability mass function on $\mathcal{Z} = \{0, 1, \cdots, m-1\}^p$ given by

$$f(z_1, \cdots z_p|\theta) = \frac{1}{\Psi(\theta)} \exp\Big\{ \sum_{r=1}^{p} \theta_{rr} C(z_r) + \sum_{r=1}^{p} \sum_{j<r}^{p} \theta_{rj} C(z_r, z_j) \Big\}. \tag{3.2.1}$$

Here $\Psi(\theta) = \sum_{z \in \mathcal{Z}} \exp\Big\{ \sum_{r=1}^{p} \theta_{rr} C(z_r) + \sum_{r=1}^{p} \sum_{j<r}^{p} \theta_{rj} C(z_r, z_j) \Big\}$ is the normalizing constant. The mean field function $C(.)$ describes the marginal information on z_r while the coupling function $C(.,.)$ as suggested by the name describes the interaction between z_r and z_j. A special case of 3.2.1 is the Ising Model where m is 2, and hence $\mathcal{Z} = \{0, 1\}^p$. In case of the Ising model the mean field and the coupling functions are typically taken as identity ($C(z_r) = z_r$) and multiplicative ($C(z_r, z_j) = z_r z_j$) respectively.

The problem of interest in this work is the estimation and recovery of the sparse matrix θ based on n sample observations $\{z^i\}_{i=1}^n$, where $z^i = (z_1^i, \cdots, z_p^i) \in \mathcal{Z}$ is the i_{th} observation. We use $Z \in \{0, \dots, m-1\}^{n \times p}$ to denote the matrix of observations, where the i-th row of Z is z^i. The likelihood of θ can then be expressed as

$$\mathcal{L}^n(\theta|Z) = \prod_{i=1}^{n} f(z^i|\theta) = \prod_{i=1}^{n} \frac{1}{\Psi(\theta)} \exp\Big\{ \sum_{r=1}^{p} \theta_{rr} C(z_r^i) + \sum_{r=1}^{p} \sum_{j<r}^{p} \theta_{rj} C(z_r^i, z_j^i) \Big\}.$$

In a high-dimensional setting (typically $p > n$, $\frac{\log(p)}{n} \to 0$), likelihood based inference on θ is computationally intractable because of the normalization constant $\Psi(\theta)$. Note that, the number of summands in the normalizing constant $\Psi(\theta)$ is exponential in p, and quickly blows up for even moderate values of p.

3.2.1 Quasi(Pseudo)-likelihood

Following an approach widely adopted in the high-dimensional frequentist literature, we explore the use of quasi(pseudo)-likelihoods in the Bayesian treatment of discrete graphical models. The conditional distribution for the r_{th} node (given all other nodes) in a Potts model for the i_{th} observation z^i can be written as

$$f(z_r^i|z_{\backslash r}^i, \theta_r) = \frac{1}{\Psi_r^i(\theta_r)} \exp\left\{\theta_{rr}C(z_r^i) + \sum_{j\neq r}\theta_{rj}C(z_r^i, z_j^i)\right\}, \tag{3.2.2}$$

where $z_{\backslash r}^i = (z_1^i, \cdots, z_{r-1}^i, z_{r+1}^i, \cdots z_p^i)'$ and $\theta_r = (\theta_{r1}, \cdots, \theta_{rp})'$ is the r_{th} column of θ. The normalizing constant of this conditional distribution is given by

$$\Psi_r^i(\theta_r) = \sum_{s=0}^{m-1} \exp\left(\theta_{rr}C(s) + \sum_{j\neq r}\theta_{rj}C(s, z_j^i)\right).$$

Computing $\Psi_r^i(\theta_r)$ requires $O(p \times m)$ units of operations and hence is scalable when m is small. We denote the r_{th} conditional log-likelihood as

$$\ell_r^n(\theta_r|Z) = \sum_{i=1}^{n}\left[\theta_{rr}C(z_r^i) + \sum_{j\neq r}\theta_{rj}C(z_r^i, z_j^i) - \log\left(\Psi_r^i(\theta_r)\right)\right].$$

Following Meinshausen and Buhlmann [2006], Ravikumar et al. [2010], we consider the log pseudo-likelihood of θ given by

$$\ell^n(\theta|Z) = \sum_{r=1}^{p}\ell_r^n(\theta_r|Z). \tag{3.2.3}$$

Note that the ability to write the log pseudo-likelihood $\ell^n(\theta|Z)$ as a sum of log conditional likelihoods $\ell_r^n(\theta_r|Z)$ allows us to transform the inference on $\theta \in \mathbb{R}^{p\times p}$ into p separable sub-problems on $\mathbb{R}^p$. Parallel treatment of each of these regression problems when deploying a multi-core computer increases computational efficiency but comes at a cost of loss in symmetry in the estimated matrix θ. We get two estimates for each component θ_{ij} from the computations involving nodes i and j respectively. Following Meinshausen and Buhlmann [2006] we resolve this issue at the post-inference stage by taking an aggregate of the two estimates which shall be discussed in details in the later sections.

3.2.2 Spike and slab prior

To take advantage of the factorized form of the pseudo-likelihood function from (3.2.3) we will assume in our prior distribution that the columns of θ are independent. We note that it is a common practice in Bayesian data analysis to ignore unknown dependence structure among parameters in the prior distribution when dealing with multivariate parameters. These dependences are then learned from the data in the posterior distribution. As mentioned before, the lack of symmetry is dealt with at the post-inference stage.

As a prior distribution for θ_r we propose to use a relaxed form of the spike and slab prior (Mitchell and Beauchamp [1988],George and McCulloch [1997]). More specifically, for each parameter $\theta_r \in \mathbb{R}^p$, $r = 1, \cdots, p$, we introduce a selection parameter $\delta_r = (\delta_{r1}, \cdots, \delta_{rp}) \in \Delta$, where $\Delta = \{0,1\}^p$. We assume that the component of δ_r have independent Bernoulli prior distributions, so that the joint distribution of δ_r writes

$$\omega_{\delta_r} = \prod_{j=1}^{p} q^{\delta_{rj}}(1-q)^{\delta_{rj}} \ ; \ q = p^{-(u+1)} \ ; \ u > 0 \tag{3.2.4}$$

where u is a hyper-parameter. The conditional distribution of θ_r given δ_r is given by

$$\begin{aligned}\theta_{rj}|\{\delta_{rj}=1\} &\sim \mathbf{N}(0,\rho);\ \rho>0\\ \theta_{rj}|\{\delta_{rj}=0\} &\sim \mathbf{N}(0,\gamma);\ \gamma>0,\end{aligned}\tag{3.2.5}$$

We introduce the notations $\theta_{r\delta_r}=(\theta_{rj}\ s.t.\ \delta_{rj}=1)\in\mathbb{R}^{\|\delta_r\|_1}$, $\delta_r^c=1-\delta_r$, $\|z\|_1=\sum_{j=1}^p|z_j|$ and $\|z\|_2=\sqrt{\sum_{j=1}^p z_j^2}$. Using this notation, and writing $\delta=(\delta_1,\ldots,\delta_p)$, $\theta=(\theta_1,\ldots,\theta_p)$, the joint prior distribution of $(\delta,\theta)\in\Delta^p\times\mathbb{R}^{p\times p}$ is given by

$$\pi(\delta,d\theta)=\prod_{r=1}^p\pi(\delta_r,d\theta_r).$$

The prior distribution $\pi(\delta_r,d\theta_r)$ on $\Delta\times\mathbb{R}^p$ can be written as

$$\pi(\delta_r,d\theta_r)\propto\omega_{\delta_r}\,(2\pi\rho)^{-\frac{||\delta_r||_1}{2}}\,(2\pi\gamma)^{\frac{||\delta_r||_1}{2}}\exp\left(-\frac{1}{2\rho}\sum_{j:\,\delta_{rj}=1}\theta_{rj}^2-\frac{1}{2\gamma}\sum_{j:\,\delta_{rj}=0}\theta_{rj}^2\right)d\theta_r.\tag{3.2.6}$$

3.2.3 Quasi-posterior distribution

Following Chapter II, we combine the prior distribution in (3.2.6) together with the pseudo-likelihood $\ell_r(\cdot|Z)$ and consider the quasi-posterior distribution for the r-th column of θ on $\Delta\times\mathbb{R}^p$ given by

$$\Pi_n(\delta_r,d\theta_r|Z)\propto\omega_{\delta_r}\left(\frac{\sqrt{\gamma}}{\sqrt{\rho}}\right)^{||\delta_r||_1}\exp\left(\ell_r^n(\theta_{r\delta_r}|Z)-\frac{1}{2\rho}\sum_{j:\,\delta_{rj}=1}\theta_{rj}^2-\frac{1}{2\gamma}\sum_{j:\,\delta_{rj}=0}\theta_{rj}^2\right)d\theta_r.\tag{3.2.7}$$

Note the use of $\theta_{r\delta_r}$ (the sparsified version of θ_r) in the quasi-likelihood. Although we use the same standard Gaussian-Gaussian spike-and-slab prior (as in for instance George and McCulloch [1997], Narisetty and He [2014]), the quasi-posterior in (3.2.7) differs from those considered in the aforementioned paper due to the sparsification of $\theta_{r\delta_r}$ in the quasi-likelihood. The idea is borrowed from Chapter 2 to facilitate computation and more closely approximate the quasi-posterior distribution obtained from spike-and-slab with point-mass at the origin. We multiplicatively combine these p quasi-posterior distributions to obtain the full quasi-posterior distribution on (δ, θ) given by

$$\Pi_n(\delta, d\theta|Z) = \prod_{r=1}^{p} \Pi_n(\delta_r, d\theta_r|Z). \tag{3.2.8}$$

3.2.4 Choice of hyper-parameters

The behavior of (3.2.7) depends by and large on the choice of the hyper-parameter γ, ρ and u. We refer the readers to Chapter II for a detailed discussion. In our algorithms we set q in (3.2.4) at $q = \frac{1}{p^{1+u}}$, for some constant $u > 0$. We have found that the inference is typically very robust to any choice of u between 1 and 2.

The hyper-parameter γ is the prior variance of the inactive component, whereas ρ is the prior variance of the active components. For positive constants c_0, c_1, choose

$$\gamma = \frac{c_0}{\max(n,p)}, \quad \text{and} \quad \rho = c_1\sqrt{\frac{n}{\log(p)}}.$$

3.2.5 Post estimation symmetrization

As mentioned above our procedure can lead to two different set of estimates $\hat{\theta}_{ij}$ and $\hat{\theta}_{ji}$ for the same parameter θ_{ij}. For the sake of interpretation it is useful to provide a single estimate and credible interval. We propose a post-estimation symmetrization resulting in a singular estimate

$$\tilde{\theta}_{ij} = \frac{\hat{\theta}_{ij} + \hat{\theta}_{ji}}{2}. \tag{3.2.9}$$

Similarly, the credible region corresponding to the parameter θ_{ij} is constructed as union of the 95% credible intervals θ_{ij} and θ_{ji}. Taking the union is a conservative approach as opposed to taking the intersection. However it always provides a concrete interval or set unlike the intersection in which case the credible intervals may be too short or in some cases even result in null set. A more direct inference on the presence of edge between nodes i and j can be made from the indicator variable δ_{ij}. In the same spirit as above we estimate δ_{ij} using

$$\tilde{p}_{ij} = P(edge\ between\ node\ i\ and\ j|Z) = \frac{1}{2}\left(\hat{P}(\delta_{ij} = 1|Z) + \hat{P}(\delta_{ji} = 1|Z)\right). \tag{3.2.10}$$

3.3 MCMC Sampling Algorithms

In this section we shall discuss in details the construction of Markov Chain Monte Carlo (MCMC) algorithms to draw Monte Carlo samples from the posterior distribution (3.2.8). By virtue of independence, it is enough to draw sample for each of the joint variable (θ_r, δ_r). Large efficiency gain is possible by performing these simulations in parallel. In general we adopt a Metropolis-Hastings within Gibbs approach to create our samplers.

We describe in Section 3.3.1 a general Metropolis Adjusted Langevin Algorithm (MALA) to sample from (3.2.8). In case of Ising model, one can also take advantage of the fact that the conditional distributions are logistic regression models and employ the Polya-Gamma(PG) sampler of Polson et al. [2012] for sampling (Section 3.3.2). We compare the two schemes in Section 3.4.1.

3.3.1 A Metropolis Adjusted Langevin sampler

The algorithm updates the active components $\theta_{r\delta_r}$ given $(\delta_r, \theta_{r\delta_r^c})$, then updates the inactive components $\theta_{r\delta_r^c}$ given $(\delta_r, \theta_{r\delta_r})$, and finally updates δ_r given (θ_r). Here we have used the notations $\theta_r = [\theta_{r\delta_r}, \theta_{r\delta_r^c}]$, where $\theta_{r\delta_r}$ regroups the components of θ_r for which $\delta_{rj} = 1$, and $\theta_{r\delta_r^c}$ regroups the remaining components. We refer the reader to Robert and Casella [2004a], Liu [2001] for an introduction to basic MCMC algorithms.

Update of active parameters

Suppose that δ_r is such that $0 < ||\delta_r||_1 < p$. We update $\theta_{r\delta_r}$ by a Metropolis Adjusted Langevin Algorithm (Atchadé [2006]). Other algorithms including Hamiltonian Monte Carlo could be used as well. We define

$$h(\delta_r, \theta_r; z) = \left[\ell_r^n(\theta_{r\delta_r}|z) - \frac{1}{2\rho}||\theta_{r\delta_r}||_2^2 - \frac{1}{2\gamma}||\theta_{r\delta_r^c}||_2^2\right]. \tag{3.3.1}$$

The function $\theta_r \to h(\delta_r, \theta_r; z)$ has a gradient given by

$$\nabla_{\theta_r} h_\gamma(\delta_r, \theta_r; z) = \nabla_{\theta_{r\delta_r}} \ell_r^n(\theta_{r\delta_r}|z) - \frac{1}{\rho}\theta_{r\delta_r} - \frac{1}{\gamma}\theta_{r\delta_r^c}.$$

Following (Atchadé [2006]), we further truncate the gradient by introducing

$$G(\delta_r, \theta_r; z) \overset{\text{def}}{=} \frac{c}{c \vee \|\nabla_{\theta_r} h(\delta_r, \theta_r; z)\|_2} \nabla_{\theta_r} h(\delta_r, \theta_r; z), \tag{3.3.2}$$

for some positive constant c, where $a \vee b = \max(a, b)$. We update (one at the time) the selected components of θ_r as follows. Given j such that $\delta_{rj} = 1$, we propose

$$\theta_{rj}^{prop}|\theta_r \sim \mathbf{N}\Big(\theta_{rj} + \frac{\sigma}{2}[G(\delta_r, \theta_r; z)]_j, \sigma^2\Big), \tag{3.3.3}$$

where σ is some constant step size and $[G(\delta_r, \theta_r, \rho_r; z)]_j$ represents the j_{th} component of $G(\delta_r, \theta_r; z)$. Let $g(\theta_{rj}^{prop}|\theta_r)$ denote the density of the proposal distribution in (3.3.3). We also define $\theta_r^{prop} = [\theta_{r1}, \cdots \theta_{r(j-1)}, \theta_{rj}^{prop}, \theta_{r(j+1)}, \cdots \theta_{rp}]$ and the acceptance probability as

$$\mathsf{Acc}_{rj} = \min\left(1, \frac{g(\theta_{rj}|\theta_r^{prop})}{g(\theta_{rj}^{prop}|\theta_r)} \times \frac{\Pi_n(\delta_r, \theta_r^{prop}|Z)}{\Pi_n(\delta_r, \theta_r|Z)}\right). \tag{3.3.4}$$

With probability Acc_{rj} we set $\theta_{rj} = \theta_{rj}^{prop}$, and with probability $1 - \mathsf{Acc}_{rj}$, we do nothing. In our simulations the step size σ is kept constant. Alternatively, it can also be updated for each θ_{rj} in the spirit of an adaptive MCMC scheme if so desired.

Finally we note that, under sparse prior the number of active parameters in each node is small. Hence the active parameters at a node can be updated one by one without loss in computational efficiency.

Independent update for inactive parameters

Note that for the stated posterior distribution 3.2.7, given δ_r, the inactive components $\theta_{r\delta_r^c}$ can be updated from their full conditional distribution given by

$$\theta_{r\delta_r^c} \sim \mathbf{N}(\mathbf{0}, \gamma I_{p-\|\delta_r\|_1}). \tag{3.3.5}$$

Bernoulli sampler for selection parameters

Equation (3.2.7) is used to derive the one by one Gibbs update of the δ_{rj}'s. For each

$j = 1, \cdots p$, we define $\check{\delta}_r = (\delta_{r1}, \cdots, \delta_{r(j-1)}, \delta_{rj}^c, \delta_{r(j+1)}, \cdots, \delta_{rp})$ and set

$$\tau_{rj} = \min\left(1, \frac{\left(\frac{q}{1-q}\right)^{\|\check{\delta}_r\|_0} \left(\frac{\gamma}{\rho}\right)^{\frac{\|\check{\delta}_r\|_0}{2}} e^{h(\check{\delta}_r, \theta_r; z)}}{\left(\frac{q}{1-q}\right)^{\|\delta_r\|_0} \left(\frac{\gamma}{\rho}\right)^{\frac{\|\delta_r\|_0}{2}} e^{h(\delta_r, \theta_r; z)}}\right) \tag{3.3.6}$$

We change δ_{rj} to δ_{rj}^c based on a flip of probability τ_{rj}

The overall MCMC algorithm, hereafter referred to as MALA can be summarized as follows.

Algorithm 4. MALA sampler

For each node $r \in \{1, \cdots, p\}$ do the following.

1. Initialize with $(\theta_r^{(0)}, \delta_r^{(0)})$

2. At the t-th iteration, given $\delta_r^{(t-1)} = \check{\delta}$ and $\theta_r^{(t-1)} = \check{\theta}$, do

   (a) For each j such that $\check{\delta}_j = 1$, we update $\check{\theta}_j$ using the MALA algorithm described in (3.3.3) and (3.3.4).

   (b) Update $\check{\theta}_{\check{\delta}^c} \sim \mathbf{N}(0, \gamma I_{p-\|\check{\delta}\|_0})$

   (c) Set $\theta_r^{(t)} = \check{\theta}$. For each j in $\{1, \ldots, p\}$, we update $\check{\delta}_j$ based on a binary flip of probability τ_{rj} as defined in (3.3.6). Set $\delta_r^{(t)} = \check{\delta}$.

3.3.2 A Polya-Gamma sampler for Ising models

The Polya-Gamma sampler is a data-augmentation technique which introduces latent Polya-gamma variables to obtain an efficient Gibbs sampler for Bayesian logistic regression (Polson et al. [2012]). To see how this is used here, note that the conditional posterior of the active parameters for the r_{th} node is

$$\Pi_n(\theta_{r\delta_r}|\delta_r, \theta_{r\delta_r^c}; Z) \propto \exp\left(\ell_r^n(\theta_{r\delta_r}|Z) - \frac{1}{2\rho}\sum_{j:\delta_{rj=1}} \theta_{rj}^2\right), \tag{3.3.7}$$

which is the same as the posterior distribution in a logistic regression of variable z_r over the variables z_j for which $\delta_{rj} = 1,\ j \neq r$, using all available data samples. Given r, δ_r, we write $x(r)_{\delta_r}^{(i)} = (z_1^{(i)}, \cdots, z_{r-1}^{(i)}, 1, z_{r+1}^{(i)}, \cdots, z_p^{(i)})_{\delta_r} \in \{0,1\}^{\|\delta_r\|_1}$, $Z_r = (z_r^{(1)}, \cdots, z_r^{(n)})' \in \{0,1\}^n$ and use $X(r)_{\delta_r} \in \{0,1\}^{n\times\|\delta_r\|_1}$ to denote the matrix of n observations $\{x(r)_{\delta_r}^{(i)}\}_{i=1}^n$.

Hence to sample from (3.3.7) we follow a Gibbs update of first drawing independently Polya-Gamma random variables using

$$W_i|\theta_{r\delta_r} \sim \mathbf{PG}(1, |\langle x(r)_{\delta_r}^{(i)}, \theta_{r\delta_r}\rangle|);\ i = 1, \cdots, n \tag{3.3.8}$$

Note that $\langle a, b\rangle$ denotes the inner product between two vectors a, b. The second step is to update $\theta_{r\delta_r}$ given these Polya-Gamma variables using

$$\begin{aligned}
\theta_{r\delta_r} &\sim \mathbf{N}(\mu, \Sigma) && (3.3.9)\\
\mu &= \Sigma\left(X(r)_{\delta_r}^T(Z_r - \frac{1}{2}1_n)\right) && (3.3.10)\\
\Sigma &= \left(X(r)_{\delta_r}^T\Omega X(r)_{\delta_r} + \frac{1}{\rho}I_{\|\delta_r\|_0}\right)^{-1} && (3.3.11)\\
\Omega &= \mathsf{diag}(W_1, \cdots, W_n) && (3.3.12)
\end{aligned}$$

Independent update for inactive parameters

As in (3.3.5) given δ_r, the inactive components $\theta_{r\delta_r^c}$ can be updated independently and simultaneously from $\mathbf{N}(\mathbf{0}, \gamma I_{p-\|\delta_r\|_1})$

Bernoulli sampler for selection parameters

As before, (3.2.7) is used to derive the one by one Gibbs update of the δ_{rj}'s. For the Polya-Gamma (PG) sampler, the calculations of the Bernoulli probability of the update can be simplified. For each $j = 1, \cdots p$, we define

$$\tau_{rj} = \log\left(\frac{1-q}{q}\right) - \frac{1}{2}\log\left(\frac{\gamma}{\rho}\right) + \frac{1}{2}\left(\frac{1}{\rho} - \frac{1}{\gamma}\right)\theta_{rj}^2 - \frac{1}{2}\Big([X(r)]'_{\cdot j}\Omega[X(r)]_{\cdot j}\Big)\theta_{rj}^2 \\ - \theta_{rj}\left\langle [X(r)]_{\cdot j}, \left(Z_r - \frac{1}{2}1_n\right)\right\rangle - \Big\langle \theta_{r\delta_r}, [X(r)]'_{\cdot j}\Omega X(r)_{\delta_r}\Big\rangle \quad (3.3.13)$$

where $X(r)$ denotes the full matrix $X(r)_{\mathbf{1}_p}$ and $[X(r)]_{\cdot j}$ denotes the j_{th} column of $X(r)$.

When $\delta_{rj} = 1$ we flip it to 0 with probability $\min(1, e^{\tau_{rj}+\theta_{rj}[X(r)]'_{\cdot j}\Omega[X(r)]_{\cdot j}})$. On the other hand if $\delta_{rj} = 0$ we flip it to 1 with probability $\min(1, e^{-\tau_{rj}})$. The Polya-Gamma MCMC algorithm (hereafter PG sampler) can be summarized as follows.

Algorithm 5. PG sampler

For each node $r \in \{1, \cdots, p\}$ do the following.

1. Initialize with $(\theta_r^{(0)}, \delta_r^{(0)})$

2. At the t-th iteration, given $\delta_r^{(t-1)} = \check{\delta}$ and $\theta_r^{(t-1)} = \check{\theta}$, do

   (a) we update $\check{\theta}_{\check{\delta}}$ using the Polya-Gamma algorithm described in (3.3.8 - 3.3.12).

   (b) Update $\check{\theta}_{\check{\delta}^c} \sim N(0, \gamma I_{p-\|\check{\delta}\|_0})$

   (c) Set $\theta_r^{(t)} = \check{\theta}$. For each j in $\{1, \ldots, p\}$

   IF $\check{\delta}_j = 1$

   we flip it to 0 with probability $\min(1, e^{\tau_{rj}+\theta_{rj}[X(r)]^T_{\cdot j}\Omega[X(r)]_{\cdot j}})$

```
        ELSE
        flip it to 1 with probability $\min(1, e^{-\tau_{rj}})$.
        Here $\tau_{rj}$ is as defined in (3.3.13). Set$\delta_r^{(t)} = \check{\delta}$.
```

Before moving on to simulation studies it is worth mentioning a few recent works on estimation of Potts Model like (Moores et al. [2020], Rosu et al. [2015]). Moores et al. [2020] uses specific form of the Potts Model to develop sufficient statistics that can be used to construct surrogate likelihoods. While a direct comparison is not included in this dissertation, the idea is worth further investigation in terms of computational speed and applicability to a more general Potts Model. Rosu et al. [2015] on the other hand pre-computes the partition function on a fine grid. The computations in this case though utilizing the true likelihood, relies on the the granularity of the mesh and prior knowledge on the support of the parameter.

3.4 Simulation studies

We first present a comparison of the performance of Algorithms 4 and 5 in terms of relative error and time complexity using a logistic regression with different sample sizes (n) and dimension (p) of the parameter of interest. Secondly we generate data from Ising model with two different structures of θ and compare the error rates and recovery of the quasi-posterior samples for different data size (n). Lastly to show scalability of the algorithm, we construct credible intervals based on the posterior samples for a network parametrized by a large 300×300 matrix θ based on 2000 observations and check the percentage of active parameters that are covered by the credible intervals.

3.4.1 Comparison of PG and MALA for logistic regression

We first present results comparing the two algorithms based on logistic regression in Figure 3.1. The data was generated based on a parameter $\theta_\star \in \mathbb{R}^p$ which had 10 active signals of absolute strength approximately $4\sqrt{10\frac{\log(p)}{n}}$ with a positive or negative sign randomly assigned to them. The regressors were drawn from independent Gaussian distribution and adjusted to have $\|X_j\|_2^2 = n,\ j = 1, \cdots, p$. We used $\rho = \sqrt{\frac{n}{\log(p)}}$, $\gamma = \frac{1}{n \vee p}$ and $u = 2$. We define the relative error and recovery as follows

$$\text{relative error at iteration } t: \ e^{(t)} \overset{\text{def}}{=} \frac{||\theta^{(t)} - \theta_\star||_2}{\|\theta_\star\|_2} \tag{3.4.1}$$

$$\text{F1 score at iteration } t: \ \text{F1}^{(t)} \overset{\text{def}}{=} \frac{2 * TA^{(t)} * PA^{(t)}}{TA^{(t)} + PA^{(t)}} \tag{3.4.2}$$

Here,

$TA^{(t)}$ = proportion of true active out of predicted active elements of δ at iteration t and

$PA^{(t)}$ = proportion of predicted active out of truly active elements of δ at iteration t.

We run both algorithm for 5000 iterations. Figure 3.1 shows the relative error (averaged over the number of iterations), as well as the total computation time. The comparison of the computational complexity is valid since both the samplers started from the same initializations and ran for the same number of iterations. Moreover, the mixing of the chains are similar and this is further substantiated by the fact that the average relative errors from the two samplers remain close in 3.1. Since (δ, θ) arise from the same quasi-posterior distribution in both algorithms, the closeness in relative errors is indicative of comparable mixing of the two samplers. Hence the only point of comparison between the samplers is their performance in terms of computational complexity. The notable conclusion is that the time complexity for the Polya-Gamma sampler degrades compared to the MaLa sampler when the sample size n is much larger than the dimension. This is due to the fact that sampling n Polya-Gamma variables at each iteration increases the computation cost of the algorithm significantly.

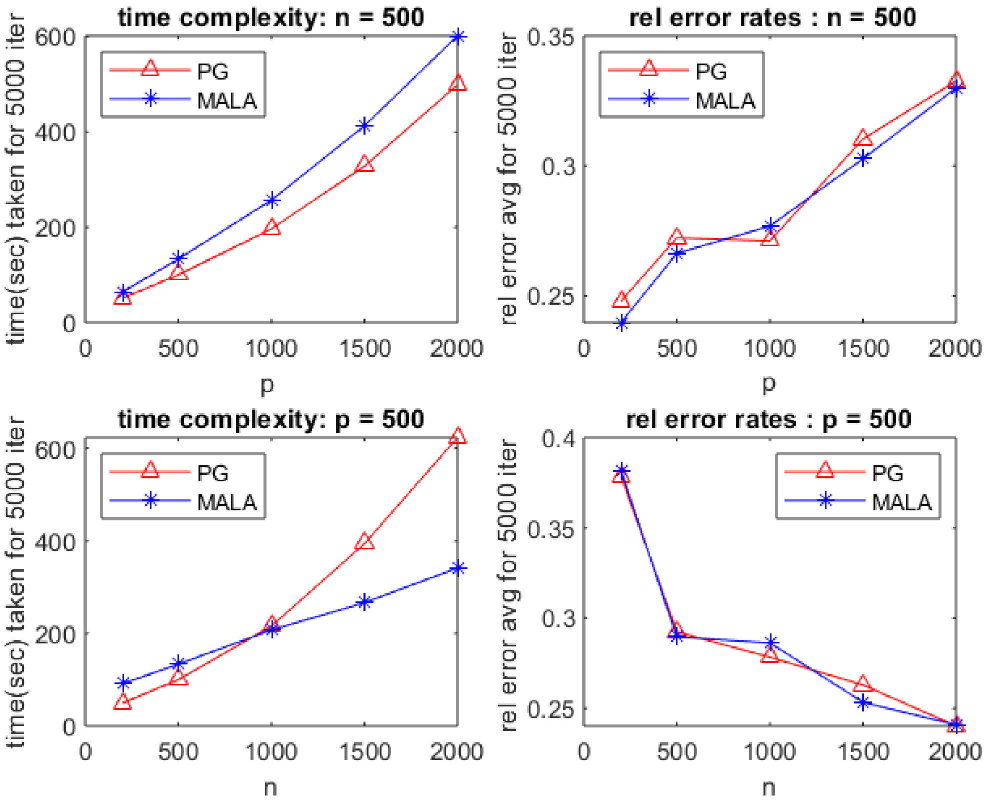


Figure 3.1: Comparison of MALA 4 and PG 5 for Logistic Regression based on 5000 iterations

3.4.2 Numerical experiments using the Ising model

The next set of results are based on the whole Ising Model. Here we present results based on two networks, one where the structure is completely random (Figure

3.2) and the other where it consists of clusters along the diagonals (Figure 3.3).

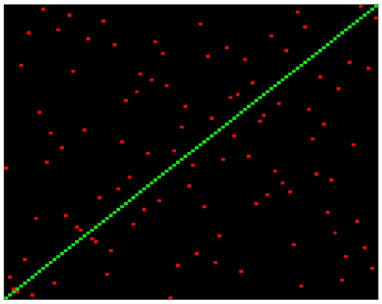


Figure 3.2: Heatmap of $\theta_\star$ [network 1]

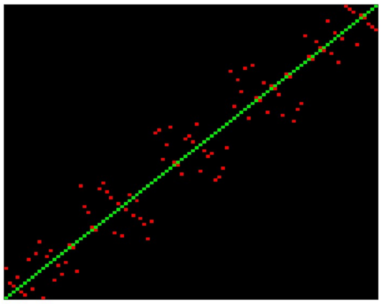


Figure 3.3: Heatmap of $\theta_\star$ [network 2]

Figure 3.3: The red and green dots indicate positive and negative values of θ_{ij} respectively

We introduce the norm $\|\theta\|_0$ as a measure of sparsity where

$$\|\theta\|_0 = \sum_{r=1}^{p} \sum_{j=1}^{p} \mathbb{1}[\theta_{rj} \neq 0].$$

For each of the two networks, the generating matrix $\theta_\star$ is symmetric in $\mathbb{R}^{100\times 100}$. Both the networks have 100 non-zero values along the diagonal of $\theta_\star$ and 50 active edges out of 4950 edges, resulting in $\|\theta_\star\|_0 = 200$.

The Ising model is well known to exhibit a phase transition phenomenon Georgii [1988]. The phase transition properties of the Ising model may lead to nodes on graph with low or no variability for certain choices of parameter $\theta_\star$ Li and Zhang [2010]. We carefully chose $\theta_{\star ij}$ to avoid these scenarios. The diagonal elements of $\theta_\star$ were chosen to be -2 and the non-zero off-diagonal $\theta_{\star ij}$'s to be 4. We generate the data from the Ising model using a Gibbs sampler.

The initialization of the MCMC values can be done randomly but the mixing will be much slower in this case. We choose to use the frequentist estimate as initial value at each node r obtained through a proximal gradient descent on the corresponding conditional likelihood Parikh and Boyd [2013]. This ensures that the MCMC sampler converges almost immediately. As we noted in Figure 3.1 the PG and MALA sampler produce similar error rates for logistic regression. Hence we present the results of the PG sampler only in case of the Ising Model for the sake of brevity.

To measure convergence of the MCMC we use the relative error (3.4.1) for each node r referring to then as $e_r^{(t)}$ for the t_{th} iteration and define

$$\text{relative error at iteration } t \text{ averaged across nodes}: \ e^{(t)} \overset{\text{def}}{=} \frac{\sum_{r=1}^{p} e_r^{(t)}}{p},$$

Similarly, using (3.4.2)

$$\text{F1 score at iteration } t \text{ averaged across nodes}: \ \text{F1}^{(t)} \overset{\text{def}}{=} \frac{\sum_{r=1}^{p} F1_r^{(t)}}{p}.$$

F1 score is the combined measure of the power of a method and it's control over false discoveries. A high F1 score indicates low type 1 error and high power.

3.4.3 Behavior of the quasi-posterior distribution with increasing sample size

We study here the behavior of the quasi-posterior distribution as the sample size increases. We generate n independent samples from the Ising model with parameter $\theta_\star \in \mathbb{R}^{100\times 100}$, for $n \in \{200, 500, 1000\}$, where $\theta_\star$ is as described above. Using the simulated data, we ran the PG sampler for 5,000 iterations with $\gamma = \frac{0.1}{p}$, $\rho = \sqrt{\frac{n}{log(p)}}$

and $u = 2$. We initialize the PG sampler using the frequentist lasso estimate. The relative errors and F1 scores averaged both over the nodes and the last 1,000 iterations are presented in Table 3.1. We can see a substantial increase in performance when the sample size grows from 200 to 500 and there is not much gain in terms of precision of estimate as sample size is increased further to 1,000. The quasi-Bayesian approach appears to perform equally well for the two types of network.

		Average Relative Error	Average F1 score
Network 1 $p = 100$	$n = 200$	0.2187	0.9336
	$n = 500$	0.0992	0.9960
	$n = 1,000$	0.0704	0.9955
Network 2 $p = 100$	$n = 200$	0.1698	0.9689
	$n = 500$	0.0846	1.0000
	$n = 1,000$	0.0690	0.9960

Table 3.1: Table showing average relative errors and average F1 scores (recovery) for the two networks and different sample sizes.

3.4.4 Behavior of credible intervals for a network with 300 nodes

We generate a larger network with 300 nodes and 2,000 observations. The network structure is similar to network 2 with block structure along the diagonals but also some sparse active edges along the anti-diagonal resulting in $\max_{r=1,\cdots,p} \|\delta_{\star r}\|_0 = 3$. Here $\theta_\star$ is symmetric in $\mathbb{R}^{300\times 300}$ with $||\theta_\star||_0 = 660$. The non-zero off-diagonal values of $\theta_\star$ are set at 4 and the diagonals of $\theta_\star$ are either -2 or -4. The settings were changed slightly again keeping in mind the phase transition properties of the Ising Model. In this setup, we specifically look at the credible intervals estimated through the MCMC samples using the PG sampler with $\gamma = \frac{0.1}{p}$, $\rho = \sqrt{\frac{n}{\log(p)}}$ and

$u = 2$. We run the PG sampler for 30,000 iterations and take the initial 10,000 iterations as burn-in. After the burn-in, the estimates of each θ_{ij} are obtained by taking the mean of 500 samples, keeping the sample from every 40_{th} iteration. The relative error for these 500 samples averaged across the 300 nodes is **0.0078** while the recovery(F1 score) is calculated to be **1.0000**. We obtain the final estimate of $\tilde{\theta}$ after symmetrization of the estimates as mentioned in (3.2.9). For the credible interval of $\theta_{\star ij}$ we use the union of the 95% credible intervals of θ_{ij} and that of θ_{ji}. Figures 3.4 and 3.5 show the credible intervals of the active and inactive θ_{ij} separately. We also include the estimates and the true value of the parameter to show the accuracy of the estimates. In 97% cases the active parameters are covered by the union credible intervals while in 3% cases they fall just outside. The inactive parameters have credible intervals symmetric around 0. The average credible intervals for each of the 4 distinct true parameter values are given in table 3.2 .

True parameter value	Average Credible Interval
0	(-0.037,0.037)
-4	(-4.43,-3.62)
-2	(-2.21,-1.82)
4	(3.66,4.40)

Table 3.2: Table showing Credible Intervals average for each of the four unique parameter values in the matrix $\theta_\star$

The total computing time of our method for this network with 300 nodes and 2000 observations was approximately 600 CPU-hours where each node ran for 30000 iterations. We parallelized the MCMC into 80 parallel processes and the simulation was completed in approximately 8 hours. Given this, we can say that our method is computationally scalable in these data dimensions.

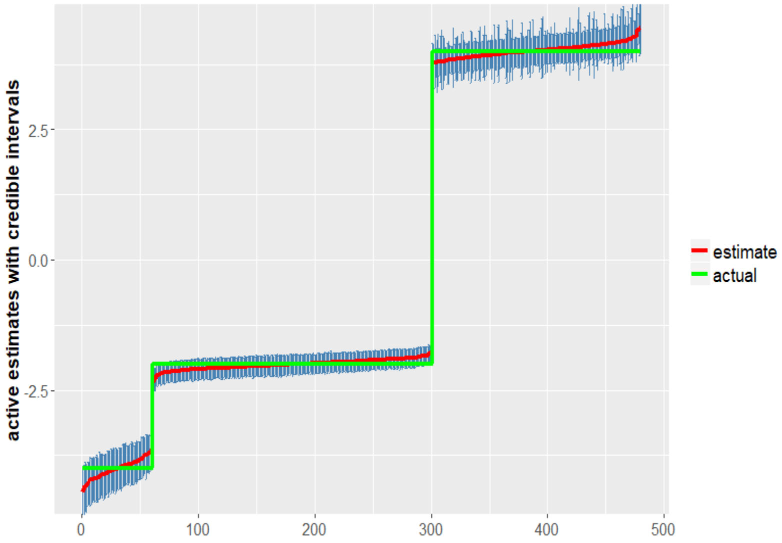


Figure 3.4: Credible intervals of active θ_{ij} in order of strength of estimates

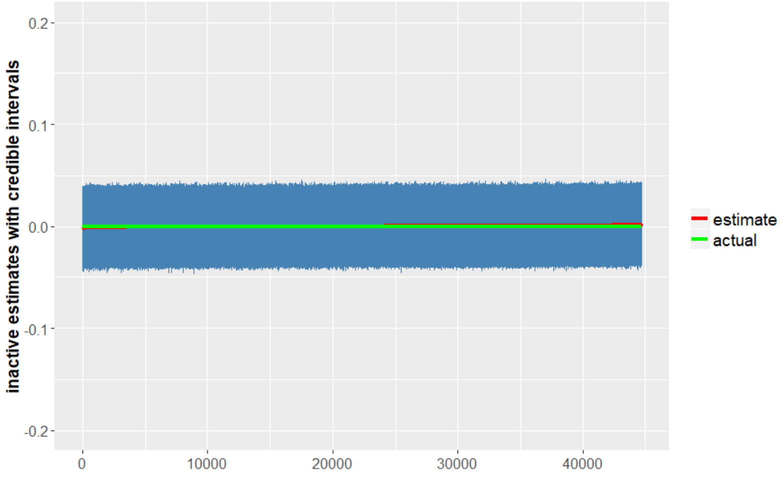


Figure 3.5: Credible intervals of inactive θ_{ij} in order of strength of estimates

3.5 Real data analysis

According to British psychologist Raymond Cattell, variations in human personality is best explained by a model containing sixteen variables (personality factors/traits) Cattell and Mead [2008]. The data that we have analyzed (source: https://openpsychometrics.org/_rawdata/), comes from an interactive questionnaire of 163 questions designed to measure Cattell's 16 Personality Factors (16PF). For each question, a self-assigned score indicates how accurate it is on a scale of (1) disagree (2) slightly disagree (3) neither agree nor disagree (4) slightly agree (5) agree. Additionally, some other information is collected which includes the test taker's home country, the source from which (s)he got information about the test, her/his perceived accuracy about the answers (s)he provided, age, gender and time elapsed to complete the test. In our analysis, we focused on women in the age group of 30 to 50, who had a self-reported accuracy $\geq 75\%$ and finished the test within half an hour.

The selected data had 4,162 individuals answering 163 questions. Some of the observations had missing values which are represented as 0. The proportion of missing values varied from 0.4% to 1% across different questions. The missing values were treated as missing at random and each of them were substituted by a value between 1 to 5. This value was sampled from the marginal distribution of scores for that particular question (covariate).

Table 3.3 describes the 16 primary factors. Each factor has 10 questions associated with it except trait B (Reasoning) which has 13 questions leading to a total of 163 questions.

Trait Name	Trait Code
Warmth	A
Reasoning	B
Emotional Stability	C
Dominance	E
Liveliness	F
Rule-Consciousness	G
Social Boldness	H
Sensitivity	I
Vigilance	L
Abstractedness	M
Privateness	N
Apprehension	O
Openness to change	Q1
Self-reliance	Q2
Perfectionism	Q3
Tension	Q4

Table 3.3: 16 PF Primary Factors

We aim to model the network of 163 questions through a Potts model with 163 nodes. Each of the questions are evaluated on a scale of 1 to 5, resulting in a 5-colored Potts model. Our objective is to understand the associations between the questions by estimating the parameter matrix θ in the Potts model (3.2.1).We set the coupling function $C(z_r, z_j) = \frac{z_r z_j}{(4)^2}$ and marginal term $C(z_r) = (\frac{z_r}{4})^2$, where $z_r \in (0, 1, \cdots, 4)$ after shifting the origin to 0. The denominators in these terms help stabilize the computation of the log-likelihoods and the derivatives required in our MCMC computations. We run the MALA sampler (Algorithm 4) using $\rho = \sqrt{n/\log(p)}$, $\gamma = \frac{1}{n}$ and $u = 2$, with a burn-in of $10,000$ iterations. The MCMC runs for $50,000$ more iterations and we keep every 50_{th} iteration to obtain a $1,000$ MCMC samples.

We define

$$\hat{\theta}_{ij} = \frac{1}{1000}\sum_{t=1}^{1000}\theta_{ij}^{(t)} \tag{3.5.1}$$

$$\hat{P}(\delta_{ij}) = \frac{1}{1000}\sum_{t=1}^{1000}\mathbb{I}(\delta_{ij}^{(t)} = 0) \tag{3.5.2}$$

The final strength of association between node (i, j) based on $1,000$ samples is then measured through a single value $\tilde{\theta}_{ij}$ evaluated as in (3.2.9) which has values in the range of $(-21, 21)$. The heatmap of the strength of association $(\tilde{\theta})$ is given in Figure 3.6. The cluster of strong signals around the diagonal represents association between questions relating to the same personality trait while the sparse off-diagonal strong signals represent association between question that are related to two different personality traits. The percentage of estimates with $\hat{P}(\delta_{ij}) = 0$ is around (94%).

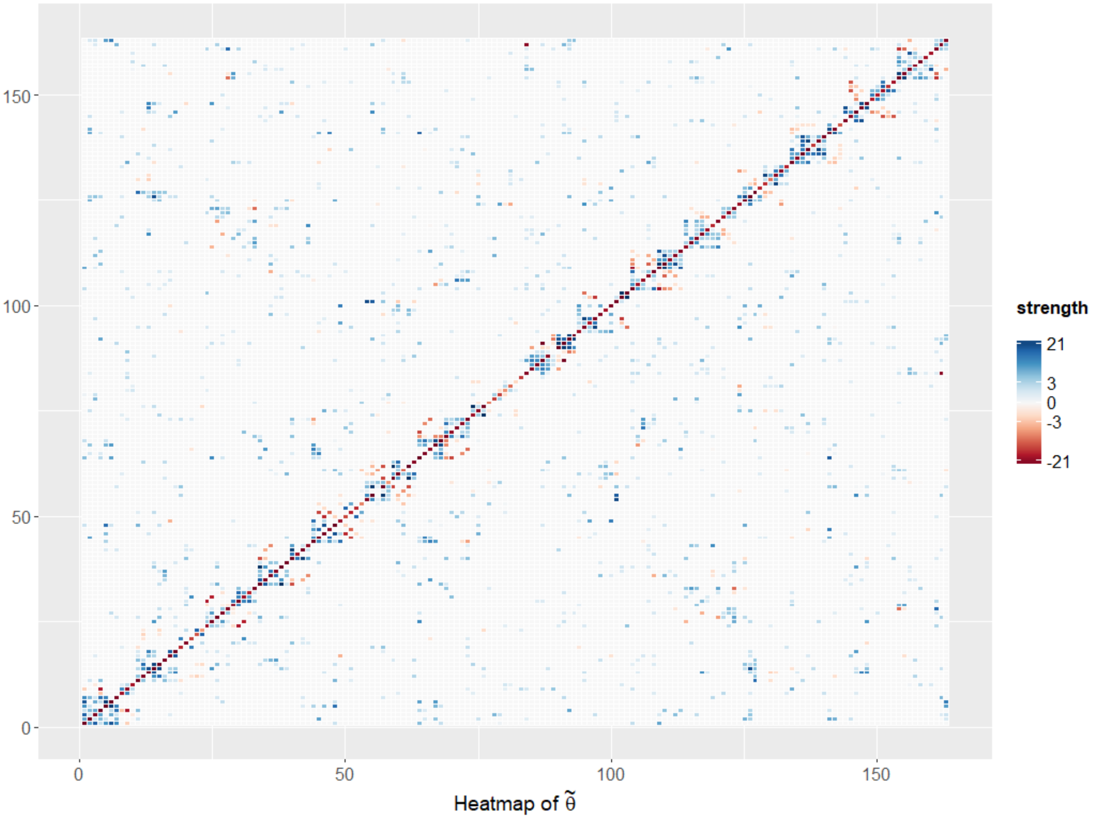


Figure 3.6: Heatmap of symmetrized $\hat{\theta}$ ((3.2.9)).

The credible region for the estimate of θ_{ij} are evaluated as union of the 95% credible intervals of θ_{ij} and θ_{ji}, obtained from the respective set of MCMC samples. Figure 3.7 shows the estimated credible intervals for all the parameters (θ_{ij}). It demonstrates the fact that for most inactive parameters the credible set is a very small interval around 0 which given the scale of the image appears as a straight line. Figure 3.8 is a zoomed in version of Figure 3.7 corresponding to parameters whose credible intervals do not contain 0.

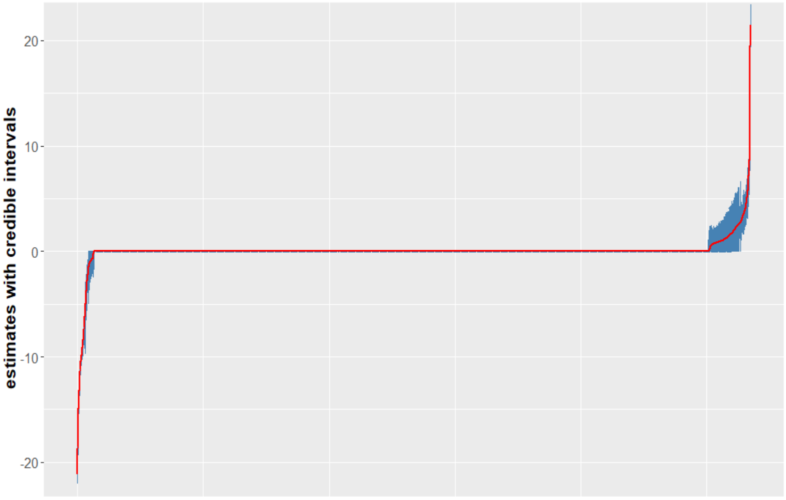


Figure 3.7: Credible intervals in order of strength for 16PF data

$\tilde{\theta}_{ij}$ (red) with credible intervals (blue) in order of strength of estimates

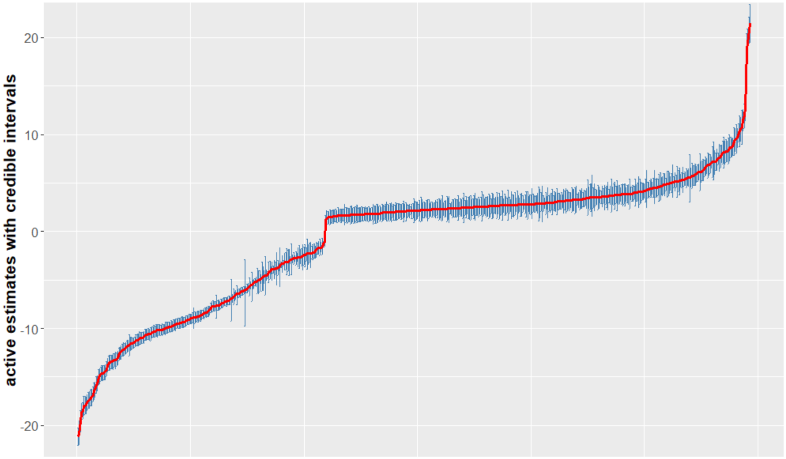


Figure 3.8: Fig 3.7 zoomed in for credible intervals not containing 0

We introduce Figure 3.9 to show the concordance between the estimates $\hat{\theta}_{ij}$ and $\hat{\theta}_{ji}$ for those estimates whose union credible intervals do not contain 0. The figure shows a high level of concordance.

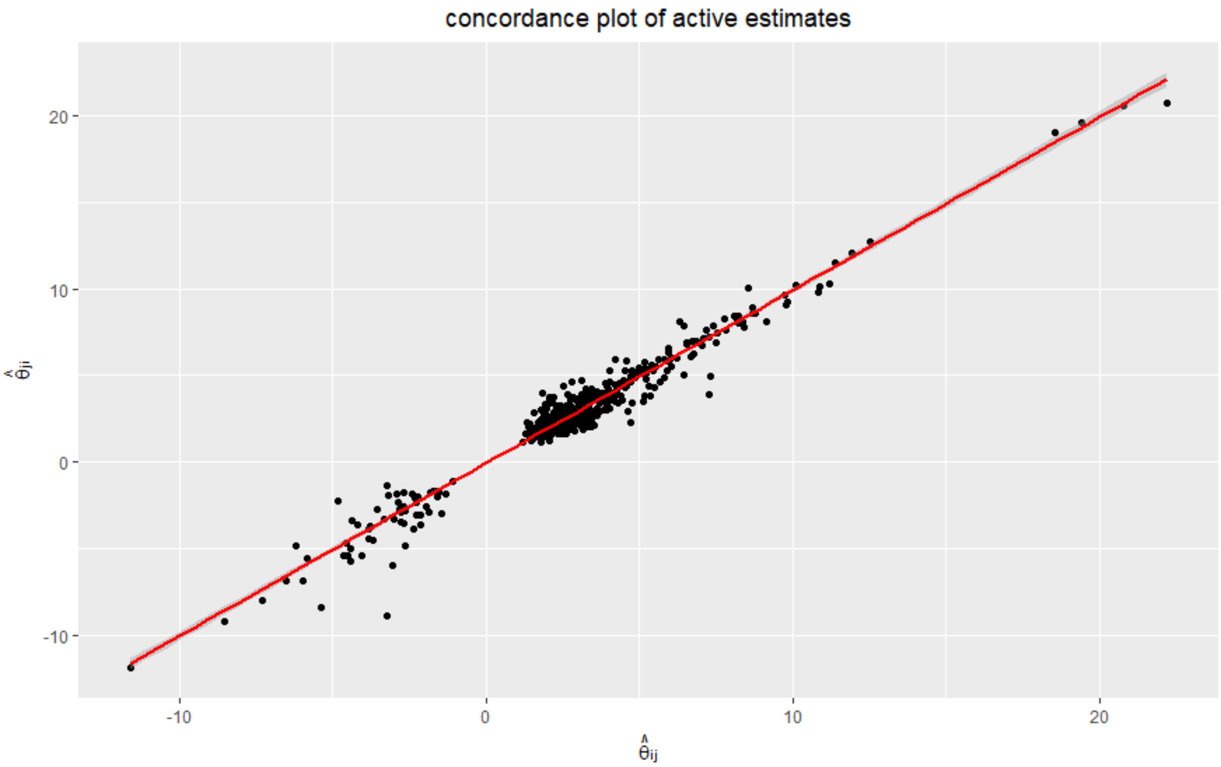


Figure 3.9: Concordance plot for estimates (16PF data) with credible intervals not containing 0. Fitted line in red has intercept: -0.08 and slope: 0.9995

Cattell and Mead [2008] used several techniques including factor analysis to establish that personality structure is hierarchical, with primary and secondary level traits. The primary level consists of the 16 personality traits (used in our analysis). The secondary level consists of a version of the Big Five Traits corresponding to broader human qualities. They are obtained by factor-analyzing the correlation matrix of the 16 primary-level personality traits.

The grouping of the 16 primary factors into the Big Five Traits are shown in Table 3.4. Reasoning (trait B) stands alone without any association to the Big Five Traits.

Big Five Traits	Associated 16PF Traits
Introversion/Extroversion	A, F, H, N, Q2
Low anxiety/High Anxiety	C, L, O, Q4
Receptivity/Tough-Mindedness	A, I, M, Q1
Accommodation/Independence	E, H, L , Q1
Lack of Restraint/Self Control	F, G, M, Q3
–	B

Table 3.4: Grouping of the 16 primary factors into the Big Five Traits

With the results of the analysis we now wish to see if the 16 primary factors show similar associations as the ones established in Table 3.4, thus providing a validation to the inference. In order to do so, we start the probability of edge between the questions (i, j) given by $\tilde{p}_{ij}$ (3.2.10) and (3.5.2). We summarize the estimates of probability of edge between 163 questions into a smaller 16×16 matrix ϕ corresponding to the 16 traits. We define the set $S_i = \{\text{questions under trait } i\}$ and n_{ij} to be the total number of possible edges between trait i and trait j. We define the matrix ϕ as

$$\phi_{ij} = \frac{1}{n_{ij}} \sum_{k \in S_i, l \in S_j} \tilde{p}_{kl} \quad .$$

The off-diagonal elements of the matrix ϕ measure the average probability of association between each pair of traits. The element-wise reciprocal of this matrix gives us a pseudo-distance measure between the 16 traits which is used to form a hierarchical clustering using Ward's method (ward.D2 in *stats:hclust* in R). Since we did not use model based clustering, it is not possible to present probabilities of the the traits belonging to a cluster. However the dendogram in 3.10 showing the structure of the hierarchical clustering offers some insight on how the traits are connected. As for example, the self control cluster and receptivity cluster share the trait M, and are closely connected through that trait. Similarly the closeness of the introversion cluster to the receptivity and self control cluster is due to the shared traits A and F. However a hard clustering on our traits results in a loss of this trait overlapping information, and we are left with separate non-overlapping clusters marked by the red blocks in the dendogram 3.10.

Figure 3.10 shows the results of the clustering. We see that our method perfectly recovers the low-anxiety/high-anxiety (C,L,O,Q4) cluster [Table 3.4]. It also nearly recovers Introversion/Extroversion (A, H, N, Q2)[Table 3.4]. The trait F(liveliness) [Table 3.3] which is common to both Introversion/Extroversion and Lack of Restraint/Self-Control in Table 3.4 is shown to be clustered more strongly with the later group and we also recover most of the Lack of Restraint/Self Control Cluster (F,G,M). In our clustering (I,Q1) are also placed together which is substantiated by the fact that they are common to the Receptivity/Tough-Mindedness cluster [Table3.4]. Additionally we find that given the data and the demographics with which we chose to work our method identifies a new cluster (E,Q3,B) which may lead to possible novel insights for this particular demographic warranting further investigations. Thus we see that several groupings in Table 3.4 corresponding to the Big Five Traits are reflected in our method.

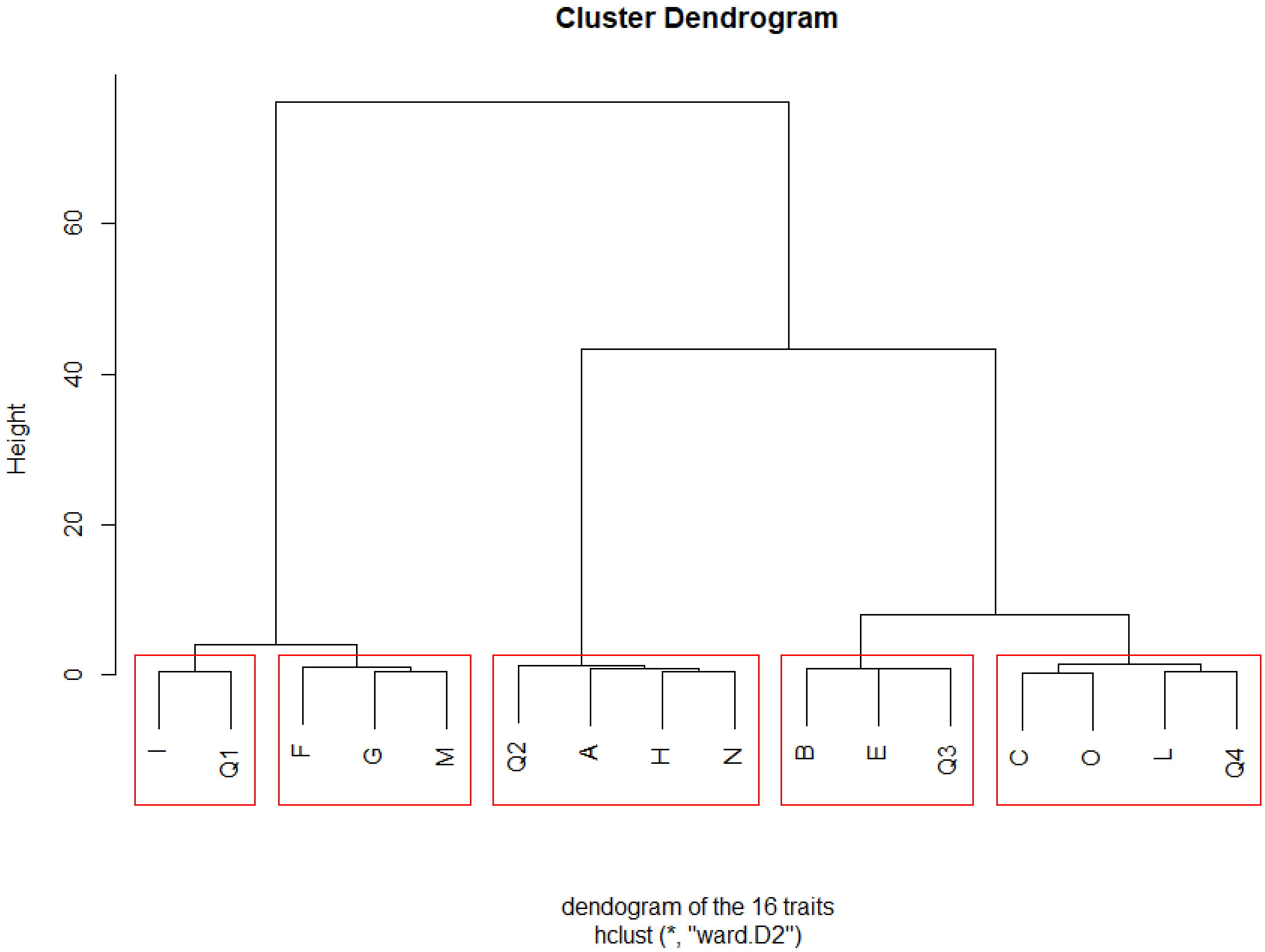


Figure 3.10: Dendogram identifying clusters of the 16 traits

CHAPTER IV

Conclusion

This dissertation explores the scope of using quasi-likelihood based methods in high dimensional Bayesian inference and opens up further avenues of research in this context. We have provided various results on the general quasi-posterior distribution. We illustrated the applicability of these results using specific examples including linear regression, logistic regression, Gaussian graphical models and sparse PCA models. These results cover several properties of the quasi-posterior distribution, including posterior sparsity, contraction rates, selection consistency, Bernstein Von Mises phenomenon. We also provide a quantification of variational approximation accuracy, since variational approximations are relevant in terms of the improvement in computational speed.

The dissertation also shows that the use of a pseudo (quasi)-likelihood and a prior distribution that factorizes across the columns of the parameter matrix can enable us to side-step the intractable normalization constant of the Potts model and perform computations in parallel for each node of the graph. We have shown in our simulations that for appropriate choices of the hyper-parameters, the method recovers the true data-generating parameters and achieves high recovery even for moderate sample size. The proposed MCMC algorithms can easily handle graphs with thousand nodes, and possibly more if access to a computer with a large number of cores is available.

However, there is still scope of further research to understand these results and improve them. As mentioned before, several assumptions are difficult to verify for non-Gaussian models. Though we have presented several results on logistic regression, they can be further improved by relaxing the assumptions required. One of the immediate focus in this stage might be to see how far the results are applicable to generalized linear models and models with sub-gaussian tails. Another direction will be to study the applicability of the results obtained for logistic regression in case of Ising Models. Though we have seen through simulations that a quasi-likelihood method provides valid estimates and intervals for Ising models, the exact rates and proofs would only further the validity of the method.

I appreciate your patience and interest and hope this doctoral research would be helpful to further the statistical understanding of the entire community.

APPENDICES

APPENDIX A

Proofs of the main results

A.1 Some preliminary lemmas

Let $\mu_\delta(\mathrm{d}\theta)$ denote the product measure on $\mathbb{R}^p$ given by

$$\mu_\delta(\mathrm{d}\theta) \stackrel{\mathrm{def}}{=} \prod_{j=1}^{p} \mu_{\delta_j}(\mathrm{d}\theta_j),$$

where $\mu_0(\mathrm{d}x)$ is the Dirac mass at 0, and $\mu_1(\mathrm{d}x)$ is the Lebesgue measure on $\mathbb{R}$. We start with a useful lower bound on the normalizing constant.

Lemma 10. *Assume H1-H2. For $z \in \mathcal{Z}$, let $C(z)$ denote the normalizing constant of $\Pi(\cdot|z)$. For $z \in \mathcal{E}_0$, we have*

$$C(z) \geq \omega(\delta_\star) e^{\ell(\theta_\star;z)} e^{-\frac{\rho_1}{2}\|\theta_\star\|_2^2} \left(\frac{\rho_1}{\bar{\kappa}+\rho_1}\right)^{\frac{\|\theta_\star\|_0}{2}}. \tag{A.1}$$

Proof. The proof is very similar to the proof of Lemma 11 of Atchade [2017]. We set

$$\bar{\omega}(\delta) \overset{\text{def}}{=} \omega(\delta) \left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \left(\frac{\rho_0}{2\pi}\right)^{\frac{p-\|\delta\|_0}{2}}.$$

Fix $z \in \mathcal{E}_0$. Then Π is well-defined, and we have

$$\begin{aligned} C(z) &= \sum_{\delta\in\Delta} \bar{\omega}(\delta) \int_{\mathbb{R}^p} e^{-\ell(\theta_\delta;z)-\frac{\rho_1}{2}\|\theta_\delta\|_2^2-\frac{\rho_0}{2}\|\theta-\theta_\delta\|_2^2}\mathrm{d}\theta \\ &\geq \bar{\omega}(\delta_\star) \int_{\mathbb{R}^p} e^{-\ell(\theta_{\delta_\star};z)-\frac{\rho_1}{2}\|\theta_{\delta_\star}\|_2^2-\frac{\rho_0}{2}\|\theta-\theta_{\delta_\star}\|_2^2}\mathrm{d}\theta \\ &= \bar{\omega}(\delta_\star)(2\pi\rho_0^{-1})^{\frac{p-\|\delta_\star\|_0}{2}} \int_{\mathbb{R}^p} e^{\ell(u;z)-\frac{\rho_1}{2}\|u\|_2^2}\mu_{\delta_\star}(\mathrm{d}u). \end{aligned}$$

Setting $G \overset{\text{def}}{=} \nabla\ell(\theta_\star; z)$, we have for all $u \in \mathbb{R}^p_{\delta_\star}$ and $z \in \mathcal{E}_0$,

$$\ell(u; z) - \ell(\theta_\star; z) - \langle G, u - \theta_\star\rangle \geq -\frac{\bar{\kappa}}{2}\|u - \theta_\star\|_2^2,$$

which implies that

$$C(z) \geq \omega(\delta_\star) \left(\frac{\rho_1}{2\pi}\right)^{s_\star/2} e^{\ell(\theta_\star;z)-\frac{\rho}{2}\|\theta_\star\|_2^2} \int_{\mathbb{R}^p} e^{\langle G,u-\theta_\star\rangle-\frac{\bar{\kappa}}{2}\|u-\theta_\star\|_2^2+\frac{\rho_1}{2}\|\theta_\star\|_2^2-\frac{\rho_1}{2}\|u\|_2^2}\mu_{\delta_\star}(\mathrm{d}u).$$

For all $u \in \mathbb{R}^p_{\delta_\star}$, $(1/2)(\|\theta_\star\|_2^2 - \|u\|_2^2) = -\frac{1}{2}\|u - \theta_\star\|_2^2 - \langle\theta_\star, u - \theta_\star\rangle$. Therefore,

$$\begin{aligned} &\int_{\mathbb{R}^p} e^{\langle G,u-\theta_\star\rangle-\frac{\bar{\kappa}}{2}\|u-\theta_\star\|_2^2+\frac{\rho_1}{2}\|\theta_\star\|_2^2-\frac{\rho_1}{2}\|u\|_2^2}\mu_{\delta_\star}(\mathrm{d}u) \\ &\qquad = \int_{\mathbb{R}^p} e^{\langle G-\rho_1\theta_\star,u-\theta_\star\rangle-\frac{\bar{\kappa}+\rho_1}{2}\|u-\theta_\star\|_2^2}\mu_{\delta_\star}(\mathrm{d}u) = \left(\frac{2\pi}{\bar{\kappa}+\rho_1}\right)^{\frac{s_\star}{2}} e^{\frac{\bar{\kappa}+\rho_1}{2}\|G-\rho_1\theta_\star\|_2^2}, \end{aligned}$$

and (A.1) follows easily.

□

Our proofs rely on the existence of some generalized testing procedures that we develop next, following ideas from Atchade [2017]. More specifically we will make use of the following result which follows by combining Lemma 6.1 and Equation (6.1) of Kleijn and van der Vaart [2006].

Lemma 11 (Kleijn-Van der Vaart (2006))**.** *Let $(\mathcal{X}, \mathcal{B}, \lambda)$ be a measure space with a sigma-finite measure λ. Let p be a density on $\mathcal{X}$, and $\mathcal{Q}$ a family of integrable real-valued functions on $\mathcal{X}$. There exists a measurable $\phi : \mathcal{X} \to [0,1]$ such that*

$$\sup_{q \in \mathcal{Q}} \left[\int \phi p \mathrm{d}\lambda + \int (1-\phi) q \mathrm{d}\lambda \right] \leq \sup_{q \in \mathsf{conv}(\mathcal{Q})} \mathcal{H}(p,q),$$

where $\mathsf{conv}(\mathcal{Q})$ is the convex hull of $\mathcal{Q}$, and $\mathcal{H}(q_1,q_2) \overset{\text{def}}{=} \int \sqrt{q_1 q_2} \mathrm{d}\lambda$.

We introduce the quasi-likelihood

$$f_\theta(z) \overset{\text{def}}{=} e^{\ell(\theta;z)}, \;\; \theta \in \mathbb{R}^p, \; z \in \mathcal{Z}.$$

For $\theta_1 \in \mathbb{R}^p$, we recall that

$$\mathcal{L}_{\theta_1}(\theta; z) \overset{\text{def}}{=} \ell(\theta; z) - \ell(\theta_1; z) - \langle \nabla \ell(\theta_1; z), \theta - \theta_1 \rangle, \; \theta \in \mathbb{R}^p.$$

We develop the test in a slightly more general setting. More specifically , in order to handle the PCA example we will allow the mode of $\ell(\cdot; z)$ to depend on z.

Let $\delta_\star$ be some sparse element Δ. Let $\Theta_\star$ be a finite nonempty subset of $\mathbb{R}^p_{\delta_\star}$ (the set of possible contraction points). Let $\bar{\rho} > 0$ be a constant, $\bar{s} \geq 1$ an integer, and r a rate function. For each $\theta_\star \in \Theta_\star$, we define

$$\mathcal{E}_{\mathsf{t},\theta_\star} \overset{\text{def}}{=} \Big\{ z \in \mathcal{Z} :\ \|\nabla \log f_{\theta_\star}(z)\|_\infty \leq \frac{\bar\rho}{2},\\ \text{and for all } \delta \in \Delta_{\bar s},\ \theta \in \mathbb{R}^p_\delta,\ \mathcal{L}_{\theta_\star}(\theta; z) \leq -\frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_2) \Big\},$$

which roughly represents the set of data points for which $\Pi(\cdot|z)$ could contract towards $\theta_\star$.

Lemma 12. *Set $s_\star \overset{\text{def}}{=} \|\delta_\star\|_0$, and*

$$\epsilon \overset{\text{def}}{=} \inf\left\{ z > 0 :\ \mathsf{r}(x) - 2\bar\rho(s_\star + \bar s)^{1/2} x \geq 0, \ \ \textit{for all } x \geq z \right\}.$$

Let $f_\star$ be a density on $\mathcal{Z}$, and $M > 2$ a constant. There exists a measurable function $\phi : \mathcal{Z} \to [0,1]$ such that

$$\int_{\mathcal{Z}} \phi(z) f_\star(z) \mathrm{d}z \leq \frac{2|\Theta_\star|(9p)^{\bar s} e^{-\frac{M}{8}\bar\rho(s_\star + \bar s)^{1/2}\epsilon}}{1 - e^{-\frac{M}{8}\bar\rho(s_\star+\bar s)^{1/2}\epsilon}},$$

where $|\Theta_\star|$ denotes the cardinality of $\Theta_\star$. Furthermore, for any $\delta \in \Delta_{\bar s}$, any $\theta \in \mathbb{R}^p_\delta$ such that $\|\theta - \theta_\star\|_2 > jM\epsilon$ for some $j \geq 1$, and some $\theta_\star \in \Theta_\star$, we have

$$\int_{\mathcal{E}_{t,\theta_\star}} (1 - \phi(z)) \frac{f_\theta(z)}{f_{\theta_\star}(z)} f_\star(z) \mathrm{d}z \leq e^{-\frac{1}{8} r\left(\frac{jM\epsilon}{2}\right)}.$$

Proof. Define

$$\bar q_{\theta_\star, u}(z) \overset{\text{def}}{=} \frac{f_u(z)}{f_{\theta_\star}(z)} f_\star(z) \mathbf{1}_{\mathcal{E}_{\mathsf{t},\theta_\star}}(z), \quad \theta_\star \in \Theta_\star, \ \ u \in \mathbb{R}^p, \ \ z \in \mathcal{Z}.$$

Using the properties of the event $\mathcal{E}_{\mathsf{t},\theta_\star}$, we note that for $\delta \in \Delta_{\bar{s}}$, and $u \in \mathbb{R}^p_\delta$ we have

$$\int_{\mathcal{Z}} \bar{q}_{\theta_\star,u}(z)\mathrm{d}z = \int_{\mathcal{E}_{\mathsf{t},\theta_\star}} e^{\langle \nabla \ell(\theta_\star;z), u-\theta_\star\rangle + \mathcal{L}_{\theta_\star}(u;z)} f_\star(z)\mathrm{d}z \leq e^{\frac{\bar{\rho}}{2}\|u-\theta_\star\|_1} < \infty. \tag{A.2}$$

Fix $\eta \geq 2\epsilon$ arbitrary. Fix $\theta_\star \in \Theta_\star$, $\delta \in \Delta_{\bar{s}}$, and fix $\theta \in \mathbb{R}^p_\delta$ such that $\|\theta - \theta_\star\|_2 > \eta$. Let

$$\mathcal{P} = \mathcal{P}_{\theta_\star,\delta,\theta} \overset{\text{def}}{=} \left\{ \bar{q}_{\theta_\star,u} :\ u \in \mathbb{R}^p_\delta,\ \|u - \theta\|_2 \leq \frac{\eta}{2} \right\}.$$

According to Lemma 11, applied with $p = f_\star$, and $\mathcal{Q} = \mathcal{P}$, there exists a test function $\phi_{\theta_\star,\delta,\theta}$ (that we will write simply as ϕ for convenience) such that

$$\sup_{q\in\mathcal{P}} \left[\int \phi f_\star + \int (1-\phi) q \right] \leq \sup_{q \in \mathsf{conv}(\mathcal{P})} \int_{\mathcal{Z}} \sqrt{f_\star(z) q(z)}\mathrm{d}z. \tag{A.3}$$

Any $q \in \mathsf{conv}(\mathcal{P})$ can be written as $q = \sum_j \alpha_j \bar{q}_{\theta_\star,u_j}$, where $\sum_j \alpha_j = 1$, $u_j \in \mathbb{R}^p_\delta$, $\|u_j - \theta\|_2 \leq \eta/2$. Notice that this implies that $\|u_j - \theta_\star\|_2 > \eta/2 \geq \epsilon$. Therefore, by Jensen's inequality, the first inequality of (A.2), and the properties of the set $\mathcal{E}_{\mathsf{t},\theta_\star}$, we get

$$\begin{aligned}
\int_{\mathcal{Z}} \sqrt{f_\star(z) q(z)}\mathrm{d}z &\leq \sqrt{\sum_j \alpha_j \int_{\mathcal{E}_{\mathsf{t},\theta_\star}} \frac{f_{u_j}(z)}{f_{\theta_\star}(z)} f_\star(z)\mathrm{d}z} \\
&\leq \sqrt{\sum_j \alpha_j e^{\frac{\bar{\rho}}{2}\|u_j-\theta_\star\|_1 - \frac{1}{2}\mathsf{r}(\|u_j-\theta_\star\|_2)}}, \\
&\leq \sqrt{\sum_j \alpha_j e^{-\frac{1}{4}\mathsf{r}(\|u_j-\theta_\star\|_2)}} \\
&\leq e^{-\frac{1}{8}\mathsf{r}\left(\frac{\eta}{2}\right)}.
\end{aligned}$$

Consequently, (A.3) yields

$$\sup_{q\in\mathcal{P}}\left[\int \phi f_\star + \int (1-\phi)q\right] \le e^{-\frac{1}{8}\mathsf{r}\left(\frac{\eta}{2}\right)}. \tag{A.4}$$

For $M > 2$, write $\cup_{\theta_\star}\cup_\delta \{\theta\in\mathbb{R}^p_\delta:\ \|\theta-\theta_\star\|_2 > M\epsilon\}$ as $\cup_{\theta_\star}\cup_\delta\cup_{j\ge1}\mathcal{A}_\epsilon(\theta_\star,\delta,j)$, where the unions in δ are taken over all δ such that $\|\delta\|_0\le\bar{s}$, and

$$\mathcal{A}_\epsilon(\theta_\star,\delta,j) \overset{\text{def}}{=} \left\{\theta\in\mathbb{R}^p_\delta:\ jM\epsilon < \|\theta-\theta_\star\|_2 \le (j+1)M\epsilon\right\}.$$

For $\mathcal{A}_\epsilon(\theta_\star,\delta,j)\neq\emptyset$, let $\mathcal{S}(\theta_\star,\delta,j)$ be a maximally $(jM\epsilon/2)$-separated point in $\mathcal{A}_\epsilon(\theta_\star,\delta,j)$. It is easily checked that the cardinality of $\mathcal{S}(\theta_\star,\delta,j)$ is upper bounded by $9^{\|\delta\|_0}\le 9^{\bar{s}}$ (see for instance Ghosal et al. [2000] Example 7.1 for the arguments). For $\theta\in\mathcal{S}(\theta_\star,\delta,j)$, let ϕ denote the test function obtained above with $\eta = jM\epsilon$. From (A.4), this test satisfies

$$\sup_{u\in\mathbb{R}^p_\delta,\ \|u-\theta\|_2\le\frac{jM\epsilon}{2}}\left[\int_{\mathcal{Z}}\phi(z)f_\star(z)\mathrm{d}z + \int_{\mathcal{Z}}(1-\phi(z))\bar{q}_{\theta_\star,u}(z)\mathrm{d}z\right] \le e^{-\frac{1}{8}\mathsf{r}\left(\frac{jM\epsilon}{2}\right)}. \tag{A.5}$$

We then set

$$\bar{\phi} = \max_{\theta_\star\in\Theta_\star}\ \max_{\delta:\ \|\delta\|_0\le\bar{s}}\ \sup_{j\ge1}\ \max_{\theta\in\mathcal{S}(\theta_\star,\delta,j)}\phi.$$

It then follows that

$$\begin{aligned}\int_{\mathcal{Z}}\bar{\phi}(z)f_\star(z)\mathrm{d}z &\le \sum_{\theta_\star}\sum_{k=0}^{\bar{s}}\ \sum_{\delta:\ \|\delta\|_0=k}\ \sum_{j\ge1}\ \sum_{\theta\in\mathcal{S}(\theta_\star,\delta,j)}\int_{\mathcal{Z}}\phi(z)f_\star(z)\mathrm{d}z\\ &\le |\Theta_\star|\sum_{k=0}^{\bar{s}}\binom{p}{k}9^k\sum_{j\ge1}e^{-\frac{1}{8}\mathsf{r}\left(\frac{jM\epsilon}{2}\right)} \le 2|\Theta_\star|(9p)^{\bar{s}}\sum_{j\ge1}e^{-\frac{1}{8}\mathsf{r}\left(\frac{jM\epsilon}{2}\right)}.\end{aligned}$$

Since $jM\epsilon/2 \geq \epsilon$, we can say that $\mathsf{r}(jM\epsilon/2) \geq 2\bar{\rho}(s_\star + \bar{s})^{1/2}(jM\epsilon/2)$. Hence

$$\sum_{j\geq 1} e^{-\frac{1}{8}\mathsf{r}\left(\frac{jM\epsilon}{2}\right)} \leq \frac{e^{-\frac{M}{8}\bar{\rho}(s_\star+\bar{s})^{1/2}\epsilon}}{1 - e^{-\frac{M}{8}\bar{\rho}(s_\star+\bar{s})^{1/2}\epsilon}}.$$

And if for some δ, such that $\|\delta\|_0 \leq \bar{s}$, some $\theta_\star \in \Theta_\star$, and some $\theta \in \mathbb{R}^p_\delta$ we have $\|\theta - \theta_\star\|_2 > jM\epsilon$, then θ resides within $(iM\epsilon)/2$ of some point $\theta_0 \in \mathcal{S}(\theta_\star, \delta, i)$ for some $i \geq j$. Hence, by (A.5),

$$\int_{\mathcal{Z}} (1 - \bar{\phi}(z))\bar{q}_{\theta_\star,\theta}(z)\mathrm{d}z \leq \int_{\mathcal{Z}} (1 - \phi(z))\bar{q}_{\theta_\star,\theta}(z)\mathrm{d}z \leq e^{-\frac{1}{8}\mathsf{r}\left(\frac{iM\epsilon}{2}\right)} \leq e^{-\frac{1}{8}\mathsf{r}\left(\frac{jM\epsilon}{2}\right)}.$$

This ends the proof. □

A.2 Proof of Posterior Sparsity (Theorem 1)

Let $f :\ \Delta \times \mathbb{R}^p \to [0, \infty)$ be some arbitrary measurable function. Take $\mathcal{E} \subseteq \mathcal{E}_0$. By the control on the normalizing constant obtained in Lemma 10, we have

$$\mathbf{1}_{\mathcal{E}}(z) \int f \mathrm{d}\Pi(\cdot|z) \leq \left(1 + \frac{\bar{\kappa}}{\rho_1}\right)^{\frac{s_\star}{2}} \\ \times \sum_{\delta\in\Delta} \frac{\omega(\delta)}{\omega(\delta_\star)} \left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \mathbf{1}_{\mathcal{E}}(z) \int_{\mathbb{R}^p} f(\delta, u) \frac{e^{\ell(u;z) - \frac{\rho_1}{2}\|u\|_2^2}}{e^{\ell(\theta_\star;z) - \frac{\rho_1}{2}\|\theta_\star\|_2^2}} \mu_\delta(\mathrm{d}u).$$

We write

$$\ell(u; z) - \ell(\theta_\star; z) = \mathcal{L}_{\theta_\star}(u; z) + \langle \nabla\ell(\theta_\star; z), u - \theta_\star \rangle .$$

Therefore, since for $z \in \mathcal{E} \subseteq \mathcal{E}_0$, $\|\nabla \ell(\theta_\star; z)\|_\infty \leq \bar{\rho}/2$, it follows that for $z \in \mathcal{E}$

$$\ell(u; z) - \ell(\theta_\star; z) \leq \mathcal{L}_{\theta_\star}(u; z) + \left(1 - \frac{\rho_1}{\bar{\rho}}\right) \langle \nabla \ell(\theta_\star; z), u - \theta_\star \rangle + \frac{\rho_1}{2}\|u - \theta_\star\|_1.$$

We deduce from the above and Fubini's theorem that

$$\begin{aligned} &\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z) \int f \mathrm{d}\Pi(\cdot|Z)\right] \leq \left(1 + \frac{\bar{\kappa}}{\rho_1}\right)^{\frac{s_\star}{2}} \sum_{\delta \in \Delta} \frac{\omega(\delta)}{\omega(\delta_\star)} \left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \\ &\times \int_{\mathbb{R}^p} f(\delta, u) e^{\frac{\rho_1}{2}\left(\|\theta_\star\|_2^2 - \|u\|_2^2\right) + \frac{\rho_1}{2}\|u - \theta_\star\|_1} \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z) e^{\mathcal{L}(u;Z) + \left(1 - \frac{\rho_1}{\bar{\rho}}\right)\langle \nabla \ell(\theta_\star; Z), u - \theta_\star\rangle}\right] \mu_\delta(\mathrm{d}u). \end{aligned} \tag{A.1}$$

Set $\mathsf{d}(u) \overset{\text{def}}{=} -\rho_1\|u\|_1 + \rho_1\|\theta_\star\|_1 + (\rho_1/2)\|u - \theta_\star\|_1$, $u \in \mathbb{R}^p$. Given (2.2.1), we claim that

$$e^{\mathsf{d}(u)} \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z) e^{\mathcal{L}(u;Z) + \left(1 - \frac{\rho_1}{\bar{\rho}}\right)\langle \nabla \ell(\theta_\star; Z), u - \theta_\star\rangle}\right] \leq e^{\frac{\mathsf{a}_0}{2}} e^{-\frac{\rho_1}{4}\|u - \theta_\star\|_1}, \quad u \in \mathbb{R}^p, \tag{A.2}$$

where $\mathsf{a}_0 = -\min_{x>0}[\mathsf{r}_0 x^2 - 4\rho_1 s_\star^{1/2}]$. The proof of this statement is essentially the same as in Castillo et al. [2015] Theorem 1. We give the details for completeness. Indeed,

$$\begin{aligned} \mathsf{d}(u) &= \frac{\rho_1}{2}\|\delta_\star \cdot (u - \theta_\star)\|_1 + \frac{\rho_1}{2}\|\delta_\star^c \cdot u\|_1 - \rho_1\|\delta_\star \cdot u\|_1 - \rho_1\|\delta_\star^c \cdot u\|_1 + \rho_1\|\theta_\star\|_1 \\ &\leq -\frac{\rho_1}{2}\|\delta_\star^c \cdot (u - \theta_\star)\|_1 + \frac{3\rho_1}{2}\|\delta_\star \cdot (u - \theta_\star)\|_1. \end{aligned}$$

If $\|\delta_\star^c \cdot (u - \theta_\star)\|_1 > 7\|\delta_\star \cdot (u - \theta_\star)\|_1$, we easily deduce that $\mathsf{d}(u) \leq -\frac{\rho_1}{4}\|u - \theta_\star\|_1$. This bound together with (2.2.1) shows that the claim holds true when $\|\delta_\star^c \cdot (u - \theta_\star)\|_1 > 7\|\delta_\star \cdot (u - \theta_\star)\|_1$. If $\|\delta_\star^c \cdot (u - \theta_\star)\|_1 \leq 7\|\delta_\star \cdot (u - \theta_\star)\|_1$, then again by (2.2.1), and the bound on $\mathsf{d}(u)$ obtained above, we deduce that the logarithm of the left-hand side of (A.2) is upper bounded by

$$
\begin{aligned}
-\frac{\rho_1}{2}\|\delta_\star^c\cdot(u-\theta_\star)\|_1 &+ \frac{3\rho_1}{2}\|\delta_\star\cdot(u-\theta_\star)\|_1 - \frac{\mathsf{r}_0}{2}\|\delta_\star\cdot(u-\theta_\star)\|_2^2 \\
&\leq -\frac{\rho_1}{2}\|u-\theta_\star\|_1 + 2\rho_1 s_\star^{1/2}\|\delta_\star\cdot(u-\theta_\star)\|_2 - \frac{\mathsf{r}_0}{2}\|\delta_\star\cdot(u-\theta_\star)\|_2^2 \\
&\leq -\frac{\rho_1}{2}\|u-\theta_\star\|_1 - \frac{1}{2}\left[r_0\|\delta_\star\cdot(u-\theta_\star)\|_2^2 - 4\rho_1 s_\star^{1/2}\|\delta_\star\cdot(u-\theta_\star)\|_2\right] \\
&\leq -\frac{\rho_1}{2}\|u-\theta_\star\|_1 + \frac{\mathsf{a}_0}{2} \leq -\frac{\rho_1}{2}\|u-\theta_\star\|_1 + \frac{2\rho_1^2 s_\star}{\mathsf{r}_0}
\end{aligned}
$$

which also gives the stated claim. Hence (A.1) becomes

$$
\begin{aligned}
\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z)\int f\mathrm{d}\Pi(\cdot|Z)\right] &\leq \left(1+\frac{\bar{\kappa}}{\rho_1}\right)^{\frac{s_\star}{2}} e^{\frac{\mathsf{a}_0}{2}} \sum_{\delta\in\Delta}\frac{\omega(\delta)}{\omega(\delta_\star)}\left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0}{2}} \\
&\times \int_{\mathbb{R}^p} f(\delta,u) e^{\frac{\rho_1}{2}\left(\|\theta_\star\|_2^2-\|u\|_2^2\right)-\rho_1(\|\theta_\star\|_1-\|u\|_1)} e^{-\frac{\rho_1}{4}\|u-\theta_\star\|_1}\mu_\delta(\mathrm{d}u). \quad (A.3)
\end{aligned}
$$

The integral in the last display is bounded from above by

$$
\begin{aligned}
\int_{\mathbb{R}^p} f(\delta,u) & e^{-\frac{\rho_1}{2}\|u-\theta_\star\|_2^2+\rho_1\|\theta_\star\|_2\|u-\theta_\star\|_2+\frac{3\rho_1}{4}\|u-\theta_\star\|_1}\mu_\delta(\mathrm{d}u) \\
&\leq e^{2\rho_1\|\theta_\star\|_2^2} e^{2\rho_1\|\delta\|_0} \int_{\mathbb{R}^p} f(\delta,u) e^{-\frac{\rho_1}{4}\|u-\theta_\star\|_2^2}\mu_\delta(\mathrm{d}u),
\end{aligned}
$$

using some simple algebraic majoration. Then (A.3) becomes

$$
\begin{aligned}
\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z)\int f\mathrm{d}\Pi(\cdot|Z)\right] &\leq \left(1+\frac{\bar{\kappa}}{\rho_1}\right)^{\frac{s_\star}{2}} e^{\frac{\mathsf{a}_0}{2}+2\rho_1\|\theta_\star\|_2^2} \\
&\times \sum_{\delta\in\Delta}\frac{\omega(\delta)}{\omega(\delta_\star)}(\sqrt{2}e^{2\rho_1})^{\|\delta\|_0}\left(\frac{\rho_1}{4\pi}\right)^{\frac{\|\delta\|_0}{2}} \int_{\mathbb{R}^p} f(\delta,u) e^{-\frac{\rho_1}{4}\|u-\theta_\star\|_2^2}\mu_\delta(\mathrm{d}u). \quad (A.4)
\end{aligned}
$$

In the special case where $f(\delta, u) = \mathbf{1}_{\{\|\delta\|_0 \geq s_\star + k\}}$ for some $k \geq 0$, we have

$$\mathbb{E}_\star \left[\mathbf{1}_{\mathcal{E}}(Z)\Pi(\|\delta\|_0 \geq s_\star + k|Z)\right] \leq \left(1 + \frac{\bar{\kappa}}{\rho_1}\right)^{\frac{s_\star}{2}} e^{\frac{\mathsf{a}_0}{2}+2\rho_1\|\theta_\star\|_2^2} \sum_{\delta:\ \|\delta\|_0 \geq s_\star + k} \frac{\omega(\delta)}{\omega_{\delta_\star}} \left(\sqrt{2}e^{2\rho_1}\right)^{\|\delta\|_0}.$$

By H2, we have

$$\begin{aligned}\sum_{\delta:\ \|\delta\|_0 \geq s_\star + k} \frac{\omega(\delta)}{\omega(\delta_\star)} \left(\sqrt{2}e^{2\rho_1}\right)^{\|\delta\|_0} &= \sum_{j=s_\star+k}^{p} \binom{p}{j} \left(\frac{\mathsf{q}}{1-\mathsf{q}}\right)^{j-s_\star} \left(\sqrt{2}e^{2\rho_1}\right)^{j} \\ &\leq \binom{p}{s_\star} \left(\sqrt{2}e^{2\rho_1}\right)^{s_\star} \sum_{j=s_\star+k}^{p} \left(\frac{\sqrt{2}e^{2\rho_1}}{p^u}\right)^{j-s_\star},\end{aligned}$$

using the fact that $\frac{\mathsf{q}}{1-\mathsf{q}} = \frac{1}{p^{u+1}}$, and $\binom{p}{j} \leq p^{j-s_\star}\binom{p}{s_\star}$. Hence for $p^{u/2} \geq 2e^{2\rho_1}$ we get

$$\sum_{\delta:\ \|\delta\|_0 \geq s_\star + k} \frac{\omega(\delta)}{\omega(\delta_\star)} \left(\sqrt{2}e^{2\rho_1}\right)^{\|\delta\|_0} \leq 2\binom{p}{s_\star} \left(\sqrt{2}e^{2\rho_1}\right)^{s_\star} \frac{1}{p^{\frac{uk}{2}}} \leq 2e^{s_\star(\frac{1}{2}+2\rho_1)+s_\star\log(p)-\frac{uk}{2}\log(p)}.$$

Hence we conclude that

$$\begin{aligned}&\mathbb{E}_\star \left[\mathbf{1}_{\mathcal{E}}(Z)\Pi(\|\delta\|_0 \geq s_\star + k|Z)\right] \\ &\qquad \leq 2e^{s_\star\left(\frac{1}{2}+2\rho_1+\log(p)\right)+\frac{s_\star}{2}\log\left(1+\frac{\bar{\kappa}}{\rho_1}\right)} e^{\frac{\mathsf{a}_0}{2}+2\rho_1\|\theta_\star\|_2^2} e^{-\frac{uk}{2}\log(p)} \\ &\qquad \leq 2e^{(1+c_0)s_\star\log(p)} e^{-\frac{uk}{2}\log(p)},\end{aligned}$$

using (2.2.2). Setting $k = (2/u)(1+c_0)s_\star + j$ for some $j \geq 1$ yields the stated result. This completes the proof.

□

A.3 Proof of Posterior Contraction (Theorem 2)

We write $\mathcal{E}_1$ instead of $\mathcal{E}_1(\bar{s})$, and take $\mathcal{E} \subseteq \mathcal{E}_1$. We note that $\mathsf{B}^c = \{\delta \in \Delta : \|\delta\|_0 > \bar{s}\} \cup \mathcal{F}_1 \cup \mathcal{F}_2$, where

$$\mathcal{F}_1 \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_{\bar{s}}} \{\delta\} \times \{\theta \in \mathbb{R}^p : \ \|\theta_\delta - \theta_\star\|_2 > C\epsilon\},$$

and

$$\mathcal{F}_2 \quad \overset{\text{def}}{=} \quad \bigcup_{\delta \in \Delta_{\bar{s}}} \{\delta\} \ \times \ \{\theta \in \mathbb{R}^p : \ \|\theta_\delta - \theta_\star\|_2 \leq C\epsilon, \quad \text{and} \quad \|\theta - \theta_\delta\|_2 > \epsilon_1\},$$

where $\epsilon_1 = \sqrt{(1 + C_1)\rho_0^{-1} p}$. Therefore we have

$$\mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathsf{B}^c|Z) = \mathbf{1}_{\mathcal{E}}(Z)\Pi(\|\delta\|_0 > \bar{s}|Z) + \mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathcal{F}_1|Z) + \mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathcal{F}_2|Z). \tag{A.1}$$

Let ϕ denote the test function asserted by Lemma 12 with $M \leftarrow C$, $\Theta_\star = \{\theta_\star\}$. We can then write

$$\mathbb{E}_\star \left[\mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathcal{F}_1|Z)\right] \leq \mathbb{E}_\star \left(\phi(Z)\right) + \mathbb{E}_\star \left[\mathbf{1}_{\mathcal{E}}(Z) \left(1 - \phi(Z)\right) \Pi(\mathcal{F}_1|Z)\right]. \tag{A.2}$$

Lemma 12 gives

$$\mathbb{E}_\star \left(\phi(Z)\right) \leq \frac{2(9p)^{\bar{s}} e^{-\frac{C}{8}\bar{\rho}_1 (s_\star + \bar{s})^{1/2}\epsilon}}{1 - e^{-\frac{C}{8}\bar{\rho}_1 (s_\star + \bar{s})^{1/2}\epsilon}} \leq 4e^{-\frac{C}{32}\bar{\rho}_1 (s_\star + \bar{s})^{1/2}\epsilon}, \tag{A.3}$$

for $(C/16)\bar{\rho}(\bar{s}+s_\star)^{1/2}\epsilon \geq 2\bar{s}\log(p)$. By Lemma 10, we have

$$\begin{aligned}\mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathcal{F}_1|Z) \;\; &\leq \;\; \mathbf{1}_{\mathcal{E}}(Z)\left(1+\frac{\bar{\kappa}}{\rho_1}\right)^{s_\star/2} \\ &\qquad \times \sum_{\delta\in\Delta_{\bar{s}}} \frac{\omega(\delta)}{\omega(\delta_\star)}\left(\frac{\rho_1}{2\pi}\right)^{\|\delta\|_0/2} \int_{\mathcal{F}_\epsilon^{(\delta)}} \frac{e^{\ell(\theta;Z)-\frac{\rho_1}{2}\|\theta\|_2^2}}{e^{\ell(\theta_\star;Z)-\frac{\rho_1}{2}\|\theta_\star\|_2^2}}\mu_\delta(\mathrm{d}\theta),\end{aligned}$$

where $\mathcal{F}_\epsilon^{(\delta)} \overset{\text{def}}{=} \{\theta\in\mathbb{R}^p:\ \|\theta_\delta-\theta_\star\|_2 > C\epsilon\}$. We use this last display together with Fubini's theorem, to conclude that

$$\begin{aligned}&\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z)\left(1-\phi(Z)\right)\Pi(\mathcal{F}_1|Z)\right] \\ &\qquad\qquad \left(1+\frac{\bar{\kappa}}{\rho_1}\right)^{s_\star/2}\sum_{\delta\in\Delta_{\bar{s}}}\frac{\omega(\delta)}{\omega(\delta_\star)}\left(\frac{\rho_1}{2\pi}\right)^{\|\delta\|_0/2} \\ &\qquad\qquad \times\int_{\mathcal{F}_\epsilon^{(\delta)}}\mathbb{E}_\star\left[(1-\phi(Z))\frac{e^{\ell(\theta;Z)}}{e^{\ell(\theta_\star;Z)}}\mathbf{1}_{\mathcal{E}}(Z)\right]\frac{e^{-\frac{\rho_1}{2}\|\theta\|_2^2}}{e^{-\frac{\rho_1}{2}\|\theta_\star\|_2^2}}\mu_\delta(\mathrm{d}\theta). \quad \text{(A.4)}\end{aligned}$$

We write $\mathcal{F}_\epsilon^{(\delta)} = \cup_{j\geq 1}\mathcal{F}_{j,\epsilon}^{(\delta)}$, where $\mathcal{F}_{j,\epsilon}^{(\delta)} \overset{\text{def}}{=} \{\theta\in\mathbb{R}^p:\ jC\epsilon < \|\theta_\delta-\theta_\star\|_2 \leq (j+1)C\epsilon\}$. Using this and Lemma 12, we have

$$\begin{aligned}&\int_{\mathcal{F}_{j,\epsilon}^{(\delta)}}\mathbb{E}_\star\left[(1-\phi(Z))\frac{e^{\ell(\theta;Z)}}{e^{\ell(\theta_\star;Z)}}\mathbf{1}_{\mathcal{E}}(Z)\right]\frac{e^{-\frac{\rho_1}{2}\|\theta\|_2^2}}{e^{-\frac{\rho_1}{2}\|\theta_\star\|_2^2}}\mu_\delta(\mathrm{d}\theta) \\ &\qquad\qquad\qquad \leq e^{-\frac{1}{8}\mathsf{r}\left(\frac{jC\epsilon}{2}\right)}\int_{\mathcal{F}_{j,\epsilon}^{(\delta)}}\frac{e^{-\frac{\rho_1}{2}\|\theta\|_2^2}}{e^{-\frac{\rho_1}{2}\|\theta_\star\|_2^2}}\mu_\delta(\mathrm{d}\theta). \quad \text{(A.5)}\end{aligned}$$

We note that $\rho_1\|\theta_\star\|_2^2-\rho_1\|\theta\|_2^2 = -\rho_1\|\theta-\theta_\star\|_2^2 - 2\rho_1\left\langle\theta_\star,\theta-\theta_\star\right\rangle \leq -\rho_1\|\theta-\theta_\star\|_2^2 + 2\rho_1\|\theta_\star\|_\infty\|\theta-\theta_\star\|_1$. Therefore, for $\theta\in\mathbb{R}^p_\delta\cap\mathcal{F}_{j,\epsilon}^{(\delta)}$, $\rho_1\|\theta_\star\|_2^2-\rho_1\|\theta\|_2^2 \leq -\rho_1\|\theta-\theta_\star\|_2^2 + 2\rho_1\|\theta_\star\|_\infty(\bar{s}+s_\star)^{1/2}(j+1)C\epsilon$. We deduce that the right-hand size of (A.5) is upper-bounded by

$$e^{-\frac{1}{8}\mathsf{r}\left(\frac{jC\epsilon}{2}\right)}e^{4\rho_1\|\theta_\star\|_\infty(\bar{s}+s_\star)^{1/2}\left(\frac{jC\epsilon}{2}\right)}\left(\frac{2\pi}{\rho_1}\right)^{\|\delta\|_0/2} \leq e^{-\frac{1}{16}\mathsf{r}\left(\frac{jC\epsilon}{2}\right)}\left(\frac{2\pi}{\rho_1}\right)^{\|\delta\|_0/2},$$

using the condition $\bar{\rho} \geq 32\rho\|\theta_\star\|_\infty$. Combined with (A.5) and (A.4) the last inequality implies that

$$\begin{aligned}\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z)\left(1-\phi(Z)\right)\Pi(\mathcal{F}_1|Z)\right] &\leq \left(1+\frac{\bar{\kappa}}{\rho_1}\right)^{s_\star/2}\left(\sum_{\delta\in\Delta_{\bar{s}}}\frac{\omega(\delta)}{\omega(\delta_\star)}\right)\sum_{j\geq 1} e^{-\frac{1}{16}\mathsf{r}\left(\frac{jC\epsilon}{2}\right)} \\ &\leq \left(1+\frac{\bar{\kappa}}{\rho_1}\right)^{s_\star/2}\left(\sum_{\delta\in\Delta_{\bar{s}}}\frac{\omega(\delta)}{\omega(\delta_\star)}\right)\frac{e^{-\frac{C}{16}\bar{\rho}_1(s_\star+\bar{s})^{1/2}\epsilon}}{1-e^{-\frac{C}{16}\bar{\rho}_1(s_\star+\bar{s})^{1/2}\epsilon}}. \quad \text{(A.6)}\end{aligned}$$

We note $\binom{p}{s} \leq p^s$, so that

$$\sum_{\delta\in\Delta_{\bar{s}}}\frac{\omega(\delta)}{\omega(\delta_\star)} = \left(\frac{1-\mathsf{q}}{\mathsf{q}}\right)^{s_\star}\sum_{\delta\in\Delta_{\bar{s}}}\left(\frac{\mathsf{q}}{1-\mathsf{q}}\right)^{\|\delta\|_0} = p^{s_\star(1+u)}\sum_{s=0}^{\bar{s}}\binom{p}{s}\left(\frac{1}{p^{1+u}}\right)^s \leq 2p^{s_\star(1+u)},$$

provided that $p^u \geq 2$. It follows that

$$\begin{aligned}&\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}}(Z)(1-\phi(Z))\Pi(\mathcal{F}_1|Z)\right] \\ &\qquad \leq 2p^{s_\star(1+u)}e^{\frac{s_\star}{2}\log\left(1+\frac{\bar{\kappa}}{\rho_1}\right)}\frac{e^{-\frac{C}{16}\bar{\rho}_1(s_\star+\bar{s})^{1/2}\epsilon}}{1-e^{-\frac{C}{16}\bar{\rho}_1(s_\star+\bar{s})^{1/2}\epsilon}} \leq 4e^{-\frac{C}{32}\bar{\rho}_1(s_\star+\bar{s})^{1/2}\epsilon}, \quad \text{(A.7)}\end{aligned}$$

provided that $(C/32)\bar{\rho}(s_\star+\bar{s})^{1/2}\epsilon \geq s_\star(1+u)\log\left(p+\frac{p\bar{\kappa}}{\rho_1}\right)$.

Let $\mathcal{F}_2^{(\delta)} \overset{\text{def}}{=} \{\theta\in\mathbb{R}^p:\ \|\theta_\delta-\theta_\star\|_2 \leq C\epsilon,\ \text{ and } \|\theta-\theta_\delta\|_2 > \epsilon_1\}$, so that

$$\mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathcal{F}_2|Z) = \mathbf{1}_{\mathcal{E}}(Z)\sum_{\delta\in\Delta_{\bar{s}}}\Pi(\delta|Z)\Pi(\mathcal{F}_2^{(\delta)}|\delta, Z),$$

and $\Pi(\mathcal{F}_2^{(\delta)}|\delta, Z) \leq \mathbb{P}[\|V_\delta\|_2 > \epsilon_1]$, where $V_\delta = (V_1,\ldots,V_{p-\|\delta\|_0}) \overset{i.i.d.}{\sim} \mathbf{N}(0,\rho_0^{-1})$. By Gaussian tails bounds we get $\Pi(\mathcal{F}_2^{(\delta)}|\delta, Z) \leq 2e^{-p}$, for any constant $C_1 \geq 3$. We conclude that

$$\mathbf{1}_{\mathcal{E}}(Z)\Pi(\mathcal{F}_2|Z) \leq \frac{1}{p^{\bar{s}}}, \quad \text{(A.8)}$$

for all p large enough. The theorem follows by collecting the bounds (A.8), (A.7), (A.3), (A.2), and (A.1).

□

A.4 Proof of Selection consistency (Theorem 3)

We write $\mathcal{E}_1$ (resp. $\mathcal{E}_2$) instead of $\mathcal{E}_1(\bar{s})$ (resp. $\mathcal{E}_2(\bar{s})$), and we fix $\mathcal{E} \subseteq \mathcal{E}_2$. First we derive a contraction rate for the frequentist estimator $\hat{\theta}_\delta$. To that end we note that for $\delta \in \mathcal{A}_{\bar{s}}$, and $z \in \mathcal{E}_0$, $\|\nabla \ell^{[\delta]}([\theta_\star]_\delta; z)\|_\infty \leq \bar{\rho}/2$. Furthermore, the curvature assumption on ℓ in $\mathcal{E}_1$ implies that

$$0 \geq -\ell^{([\delta]}(\hat{\theta}_\delta; z) + \ell^{([\delta]}([\theta_\star]_\delta; z) \geq \left\langle -\nabla \ell^{[\delta]}([\theta_\star]_\delta; z), \hat{\theta}_\delta - [\theta_\star]_\delta \right\rangle + \frac{1}{2}\mathsf{r}(\|\hat{\theta}_\delta - [\theta_\star]_\delta\|_2).$$

Using this and the definition of ϵ, it follows that for $\delta \in \mathcal{A}_{\bar{s}}$,

$$\mathbf{1}_{\mathcal{E}_1}(z)\|\hat{\theta}_\delta - [\theta_\star]_\delta\|_2 \leq \epsilon. \tag{A.1}$$

Set $\mathcal{A}_+ \stackrel{\text{def}}{=} \mathcal{A}_{\bar{s}} \setminus \mathcal{A}_{s_\star+j}$, and recall that $\mathsf{B}_j = \cup_{\delta \in \mathcal{A}_{s_\star+j}} \{\delta\} \times \mathsf{B}^{(\delta)}$. Therefore we have

$$\Pi(\mathsf{B}_j|z) + \Pi\left(\cup_{\delta \in \mathcal{A}_+} \{\delta\} \times \mathsf{B}^{(\delta)}|z\right) + \Pi(\mathsf{B}^c|z) = 1,$$

so that

$$\mathbf{1}_{\mathcal{E}}(z)\left(1 - \Pi(\mathsf{B}_j|z)\right) = \mathbf{1}_{\mathcal{E}}(z)\Pi(\mathsf{B}^c|z) + \mathbf{1}_{\mathcal{E}}(z)\Pi\left(\cup_{\delta \in \mathcal{A}_+} \{\delta\} \times \mathsf{B}^{(\delta)}|z\right). \tag{A.2}$$

Hence it remains only to upper bound the last term on the right-hand side of the last display. By definition we have

$$\Pi\left(\cup_{\delta\in\mathcal{A}_+}\{\delta\}\times\mathsf{B}^{(\delta)}|z\right) = \Pi(\delta_\star\times\mathsf{B}^{(\delta_\star)}|z)\sum_{\delta\in\mathcal{A}_+}\frac{\Pi(\delta\times\mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star\times\mathsf{B}^{(\delta_\star)}|z)},$$

and

$$\frac{\Pi(\delta\times\mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star\times\mathsf{B}^{(\delta_\star)}|z)} = \frac{\omega(\delta)}{\omega(\delta_\star)}\left(\frac{\rho_1}{\rho_0}\right)^{\frac{\|\delta\|_0-s_\star}{2}}\frac{\int_{\mathsf{B}^{(\delta)}} e^{\ell(\theta_\delta;z)-\frac{\rho_1}{2}\|\theta_\delta\|_2^2-\frac{\rho_0}{2}\|\theta-\theta_\delta\|_2^2}\mathrm{d}\theta}{\int_{\mathsf{B}^{(\delta_\star)}} e^{\ell(\theta_{\delta_\star};z)-\frac{\rho_1}{2}\|\theta_{\delta_\star}\|_2^2-\frac{\rho_0}{2}\|\theta-\theta_{\delta_\star}\|_2^2}\mathrm{d}\theta}. \tag{A.3}$$

By integrating out the non-selected components $(\theta-\theta_\delta)$, we note that the integral in the numerator of the last display is bounded from above by

$$(2\pi\rho_0^{-1})^{(p-\|\delta\|_0)/2}\int_{\{\theta\in\mathbb{R}^p:\ \|\theta-\theta_\star\|_2\leq C\epsilon\}} e^{\ell(\theta;z)-\frac{\rho_1}{2}\|\theta\|_2^2}\mu_\delta(\mathrm{d}\theta),$$

whereas the integral in the denominator is lower bounded by

$$\begin{aligned}(2\pi\rho_0^{-1})^{(p-s_\star)/2}\mathbb{P}\left(\sqrt{\rho_0^{-1}}\|V\|_2\leq C_1\epsilon_1\right)&\int_{\{\theta\in\mathbb{R}^p:\ \|\theta-\theta_\star\|_2\leq C\epsilon\}} e^{\ell(\theta;z)-\frac{\rho_1}{2}\|\theta\|_2^2}\mu_{\delta_\star}(\mathrm{d}\theta)\\ &\geq\frac{1}{2}(2\pi\rho_0^{-1})^{(p-s_\star)/2}\int_{\{\theta\in\mathbb{R}^p:\ \|\theta-\theta_\star\|_2\leq C\epsilon\}} e^{\ell(\theta;z)-\frac{\rho_1}{2}\|\theta\|_2^2}\mu_{\delta_\star}(\mathrm{d}\theta),\end{aligned}$$

where $V=(V_1,\ldots,V_{p-s_\star})$ is a random vector with i.i.d. standard normal components. These observations together with (A.3) lead to

$$\frac{\Pi(\delta\times\mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star\times\mathsf{B}^{(\delta_\star)}|z)}\leq\frac{2\omega(\delta)}{\omega(\delta_\star)}\left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0-s_\star}{2}}\frac{\int_{\{\theta\in\mathbb{R}^p:\ \|\theta-\theta_\star\|_2\leq C\epsilon\}} e^{\ell(\theta;z)-\frac{\rho_1}{2}\|\theta\|_2^2}\mu_\delta(\mathrm{d}\theta)}{\int_{\{\theta\in\mathbb{R}^p:\ \|\theta-\theta_\star\|_2\leq C\epsilon\}} e^{\ell(\theta;z)-\frac{\rho_1}{2}\|\theta\|_2^2}\mu_{\delta_\star}(\mathrm{d}\theta)}.$$

For $\theta \in \mathbb{R}^p_\delta$, $\delta \in \mathcal{A}_{\bar{s}}$, and $\|\theta - \theta_\star\|_2 \leq C\epsilon$, it is easily checked that

$$-C\|\theta_\star\|_\infty \rho_1 \bar{s}^{1/2}\epsilon \leq \frac{\rho_1}{2}\left(\|\theta_\star\|_2^2 - \|\theta\|_2^2\right) \leq C\|\theta_\star\|_\infty \rho_1 \bar{s}^{1/2}\epsilon,$$

and by the definition of ϖ, and noting from (A.1) that $\|[\theta]_\delta - \hat{\theta}_\delta\|_2 \leq \|[\theta]_\delta - [\theta_\star]_\delta\|_2 + \|\hat{\theta}_\delta - [\theta_\star]_\delta\|_2 \leq (C+1)\epsilon$, we have

$$\left|\ell^{[\delta]}(\theta; z) - \ell^{[\delta]}(\hat{\theta}_\delta; z) - \underbrace{\left\langle \nabla \ell^{[\delta]}(\hat{\theta}_\delta; z), [\theta]_\delta - \hat{\theta}_\delta \right\rangle}_{=0} + \frac{1}{2}([\theta]_\delta - \hat{\theta}_\delta)' \mathcal{I}_\delta ([\theta]_\delta - \hat{\theta}_\delta)\right|$$
$$\leq \frac{\varpi(\delta, (C+1)\epsilon; z)}{6} \bar{s}^{3/2} \|[\theta]_\delta - \hat{\theta}_\delta\|_2^3 \leq \bar{s}^{3/2}\frac{\mathsf{a}_2}{6}((C+1)\epsilon)^3.$$

We conclude that

$$\frac{\Pi(\delta \times \mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star \times \mathsf{B}^{(\delta_\star)}|z)} \leq 2e^{C_0(\rho_1\|\theta_\star\|_\infty \bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}\epsilon^3)}$$
$$\times \frac{\omega(\delta)}{\omega(\delta_\star)} \left(\frac{\rho_1}{2\pi}\right)^{\frac{\|\delta\|_0 - s_\star}{2}} \frac{e^{\ell^{[\delta]}(\hat{\theta}_\delta; z)}}{e^{\ell^{[\delta_\star]}(\hat{\theta}_{\delta_\star}; z)}} \frac{\sqrt{\det\left(2\pi \mathcal{I}_\delta^{-1}\right)}}{\sqrt{\det\left(2\pi \mathcal{I}_{\delta_\star}^{-1}\right)}\mathbf{N}(\hat{\theta}_{\delta_\star}; \mathcal{I}_{\delta_\star}^{-1})(\mathsf{B}_{\delta_\star})},$$

for some absolute constant C_0, where $\mathsf{B}_\delta = \{u \in \mathbb{R}^{\|\delta\|} : \ \|u - [\theta_\star]_\delta\|_2 \leq C\epsilon\}$, and $\mathbf{N}(\hat{\theta}_\delta; \mathcal{I}_\delta^{-1})(A)$ denotes the probability of A under the Gaussian distribution $\mathbf{N}(\hat{\theta}_\delta; \mathcal{I}_\delta^{-1})$. For $z \in \mathcal{E}_1$, using the assumption $(C-1)\epsilon\underline{\kappa}^{1/2} \geq 2(s_\star^{1/2} + 1)$, and for $z \in \mathcal{E}_1$, we have $\mathbf{N}(\hat{\theta}_{\delta_\star}; \mathcal{I}_{\delta_\star}^{-1})(\mathsf{B}_{\delta_\star}) \geq 1/2$. We conclude that

$$\mathbf{1}_{\mathcal{E}_1}(z)\frac{\Pi(\delta \times \mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star \times \mathsf{B}^{(\delta_\star)}|z)} \leq 4e^{C_0(\rho_1\|\theta_\star\|_\infty \bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}\epsilon^3)} \frac{\omega(\delta)}{\omega(\delta_\star)} \left(\rho_1\right)^{\frac{\|\delta\|_0 - s_\star}{2}} \frac{e^{\ell(\hat{\theta}_\delta; z)}}{e^{\ell(\hat{\theta}_{\delta_\star}; z)}} \sqrt{\frac{\det(\mathcal{I}_{\delta_\star})}{\det(\mathcal{I}_\delta)}}. \tag{A.4}$$

For $z \in \mathcal{E}_2$, and $\|\delta\|_0 = s_\star + j$, we have

$$\ell(\hat{\theta}_\delta; z) - \ell(\hat{\theta}_{\delta_\star}; z) \le \frac{ju}{2}\log(p).$$

Recall that $\mathcal{I}_\delta = -\nabla^{(2)}\ell^{[\delta]}(\hat{\theta}_\delta; z)$. Hence we can write

$$\frac{\det(\mathcal{I}_{\delta_\star})}{\det(\mathcal{I}_\delta)} = \frac{\det\left(-\nabla^{(2)}\ell^{[\delta_\star]}(\hat{\theta}_{\delta_\star}; z)\right)}{\det\left(-\nabla^{(2)}\ell^{[\delta]}(\hat{\theta}_{\delta_\star}; z)\right)} \times \frac{\det\left(-\nabla^{(2)}\ell^{[\delta]}(\hat{\theta}_{\delta_\star}; z)\right)}{\det\left(-\nabla^{(2)}\ell^{[\delta]}(\hat{\theta}_{\delta}; z)\right)}.$$

The Cauchy interlacing property (Lemma 18) implies that the first term on the right hand side of the last display is upper bounded by $(1/\underline{\kappa})^j$. To bound the second term, we first note that by convexity of the function $-\log\det$, for any pair of symmetric positive definite matrices A, B of same size, it holds $|\log\det(A) - \log\det(B)| \le \max(\|A^{-1}\|_{\mathsf{F}}, \|B^{-1}\|_{\mathsf{F}})\|A - B\|_{\mathsf{F}}$, where $\|M\|_{\mathsf{F}}$ denotes the Frobenius norm of M. Hence, if a symmetric positive definite matrix $A(\theta)$ depends smoothly on a parameter θ, then we have $|\log\det(A(\theta)) - \log\det(A(\theta_0))| \le \sup_{u\in\Theta}\|A(u)^{-1}\|_{\mathsf{F}}\ \|\nabla A(\bar{\theta})\cdot(\theta - \theta_0)\|_{\mathsf{F}}$, for some $\bar{\theta}$ on the segment between θ and θ_0. We use this together with the definition of a_2, to conclude that the second term on the right hand of the last equation is upper bounded by $e^{\frac{2\mathsf{a}_2\bar{s}^3\epsilon}{\underline{\kappa}}}$. Hence

$$\frac{\det(\mathcal{I}_{\delta_\star})}{\det(\mathcal{I}_\delta)} \le \left(\frac{1}{\underline{\kappa}}\right)^j e^{\frac{2\mathsf{a}_2\bar{s}^3\epsilon}{\underline{\kappa}}}.$$

Using these bounds, we obtain from (A.4),

$$\mathbf{1}_{\mathcal{E}}(z)\frac{\Pi(\delta\times\mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star\times\mathsf{B}^{(\delta_\star)}|z)} \le 4e^{C_0\left(\rho_1\|\theta_\star\|_\infty\bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}\left(\epsilon^3 + \frac{\bar{s}^{1/2}\epsilon}{\underline{\kappa}}\right)\right)}\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{1+\frac{u}{2}}}\right)^j. \tag{A.5}$$

Using (A.5) and summing over $\delta \in \mathcal{A}_+$, it follows that

$$\begin{aligned}
&\mathbf{1}_{\mathcal{E}}(z)\Pi\left(\cup_{\delta\in\mathcal{A}_+}\{\delta\} \times \mathsf{B}^{(\delta)}|z\right) \\
&\quad \leq 4e^{C_0(\rho_1\|\theta_\star\|_\infty \bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}(\epsilon^3+\frac{\bar{s}^{1/2}\epsilon}{\underline{\kappa}}))} \sum_{j=k+1}^{\bar{s}-s_\star} \sum_{\delta\supseteq\delta_\star,\ \|\delta\|_0=s_\star+j} \left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{1+\frac{u}{2}}}\right)^j, \\
&\qquad\qquad \leq 8e^{C_0(\rho_1\|\theta_\star\|_\infty \bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}(\epsilon^3+\frac{\bar{s}^{1/2}\epsilon}{\underline{\kappa}}))} \left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{\frac{u}{2}}}\right)^{k+1},
\end{aligned}$$

provided that $p^{u/2}\sqrt{\underline{\kappa}/\rho_1} \geq 2$. This bound and (A.2) yields the stated bound.

Remark A.1. By tracing the steps in the proof of (A.5), it can be checked that the following lower bound also holds.

$$\mathbf{1}_{\mathcal{E}_1}(z)\frac{\Pi(\delta \times \mathsf{B}^{(\delta)}|z)}{\Pi(\delta_\star \times \mathsf{B}^{(\delta_\star)}|z)} \geq \frac{1}{4}e^{-C_0(\rho_1\|\theta_\star\|_\infty \bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}(\epsilon^3+\frac{\bar{s}^{1/2}\epsilon}{\underline{\kappa}}))} \left(\sqrt{\frac{\rho_1}{\bar{\kappa}}}\frac{1}{p^{u+1}}\right)^j. \qquad \text{(A.6)}$$

□

A.5 Proof of Bernstein Von Mises phenomenon (Theorem 4)

We start with the following general observation. Let π, q, and μ be three probability measures on some measurable space such that $\mu(\mathrm{d}x) = \frac{e^{f(x)}\pi(\mathrm{d}x)\mathbf{1}_A(x)}{\int_A e^{f(u)}\pi(\mathrm{d}u)}$ for some measurable $\mathbb{R}$-valued function f, and a measurable set A such that $\pi(A) \geq 1/2$. Furthermore, suppose that the support of q is A. Then

$$\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\mathrm{d}q = \int_A f\mathrm{d}q - \log\left(\int_A e^f\mathrm{d}\pi\right).$$

By Jensen's inequality we have

$$-\log\left(\int_A e^f \mathrm{d}\pi\right) \leq -\log(\pi(A)) - \int_A f \frac{\mathrm{d}\pi}{\pi(A)}.$$

Since $-\log(1-x) \leq 2x$ for $x \in [0, 1/2]$, we have $-\log(\pi(A)) \leq 2\pi(A^c)$, and we conclude that

$$\begin{aligned}
\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) \mathrm{d}q &\leq \left|\int_A f \mathrm{d}q - \int_A f \mathrm{d}\pi\right| + 2\pi(A^c)\left(1 + \int_A |f| \mathrm{d}\pi\right) \\
&\leq \int_A |f| \mathrm{d}q + 2\int_A |f| \mathrm{d}\pi + 2\pi(A^c). \qquad \text{(A.1)}
\end{aligned}$$

When $q = \mu$, (A.1) writes

$$\mathsf{KL}\,(\mu|\pi) \leq \int_A |f| \mathrm{d}\mu + 2\int_A |f| \mathrm{d}\pi + 2\pi(A^c). \qquad \text{(A.2)}$$

Let us now apply (A.1) and (A.2). Fix $z \in \mathcal{E}$. In order to use these bounds, we first note that the density of $\Pi_\star^{(\infty)}$ with respect to Π that can be written as

$$\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}(\delta, \theta|z) = \frac{e^{-R(\delta,\theta;z)} \mathbf{1}_{\{\delta_\star\}\times\mathbb{R}^p}(\delta, \theta)}{\int_{\{\delta_\star\}\times\mathbb{R}^p} e^{-R(\delta,\theta;z)} \Pi(\mathrm{d}\delta, \mathrm{d}\theta|z)}, \qquad \text{(A.3)}$$

where

$$\begin{aligned}
R(\delta, \theta; z) &\stackrel{\text{def}}{=} \ell(\theta_\delta; z) - \frac{\rho_1}{2}\|\theta_\delta\|_2^2 - \ell(\hat{\theta}_\delta; z) + \frac{\rho_1}{2}\|\hat{\theta}_\delta\|_2^2 + \frac{1}{2}([\theta]_\delta - \hat{\theta}_\delta)' \mathcal{I}_\delta([\theta]_\delta - \hat{\theta}_\delta), \\
&= -\frac{\rho_1}{2}\|\theta_\delta\|_2^2 + \frac{\rho_1}{2}\|\hat{\theta}_\delta\|_2^2 + \frac{1}{6}\nabla^{(3)}\ell^{[\delta]}(\bar{\theta}_\delta; z) \cdot \left([\theta]_\delta - \hat{\theta}_\delta, [\theta]_\delta - \hat{\theta}_\delta, [\theta]_\delta - \hat{\theta}_\delta\right),
\end{aligned}$$

for some element $\bar{\theta}_\delta$ on the segment between $[\theta]_\delta$ and $\hat{\theta}_\delta$. The second equality follows from Taylor expansion and $\nabla \ell^{[\delta]}(\hat{\theta}_\delta; z) = 0$. That second expression of R shows that for $z \in \mathcal{E}$, $\delta \in \mathcal{A}_{\bar{s}}$, and $\theta \in \mathsf{B}^{(\delta)}$,

$$|R(\delta, \theta)| \leq C_0 \rho_1 \bar{s}^{1/2} \epsilon + C_0 \mathsf{a}_2 \bar{s}^{3/2} \epsilon^3, \tag{A.4}$$

for some absolute constant C_0. However, in general when $\theta \notin \mathsf{B}^{(\delta)}$, $R(\delta, \theta)$ is quadratic in θ under the assumptions of the theorem. Indeed, using $\nabla \ell^{[\delta]}(\hat{\theta}_\delta; z) = 0$, we can write that $\ell(\theta_\delta; z) - \ell^{[\delta]}(\hat{\theta}_\delta; z) = -(1/2)([\theta]_\delta - \hat{\theta}_\delta)'[-\nabla^{(2)} \ell^{[\delta]}(\bar{\theta}_\delta; z)]([\theta]_\delta - \hat{\theta}_\delta)$, for some element $\bar{\theta}_\delta$ on the segment between $[\theta]_\delta$ and $\hat{\theta}_\delta$. Hence, for $\theta \in \mathbb{R}^p$

$$\begin{aligned}
|R(\delta, \theta)| \leq \frac{\rho_1}{2} &\left| \|\theta_\delta\|_2^2 - \|\hat{\theta}_\delta\|_2^2 \right| \\
&+ \frac{1}{2} \left| ([\theta]_\delta - \hat{\theta}_\delta)'[-\nabla^{(2)} \ell^{[\delta]}(\bar{\theta}_\delta; z)([\theta]_\delta - \hat{\theta}_\delta) - ([\theta]_\delta - \hat{\theta}_\delta)' \mathcal{I}_\delta([\theta]_\delta - \hat{\theta}_\delta) \right| \\
&\qquad \leq \frac{\rho_1 + \bar{\kappa}}{2} \|[\theta]_\delta - \hat{\theta}_\delta\|_2^2 + \rho_1 \|\hat{\theta}_\delta\|_2 \|[\theta]_\delta - \hat{\theta}_\delta\|_2 \\
&\qquad\qquad \leq (\rho_1 + \bar{\kappa}) \|[\theta]_\delta - \hat{\theta}_\delta\|_2^2 + \frac{\rho_1^2 (\epsilon + \|\theta_\star\|_2)^2}{2(\rho_1 + \bar{\kappa})},
\end{aligned} \tag{A.5}$$

where the second inequality uses (2.3.9), and the third inequality follows from some basic algebra, and (A.1).

Let R be some arbitrary probability measure on $\Delta \times \mathbb{R}^p$ with support $\{\delta_\star\} \times \mathbb{R}^p$. We make use of (A.1) with $q = R$, $\mu = \Pi_\star^{(\infty)}$, $\pi = \Pi$, and $A = \{\delta_\star\} \times \mathbb{R}^p$. We then split the integrals over $\{\delta_\star\} \times \mathbb{R}^p$ into $\{\delta_\star\} \times \mathsf{B}^{(\delta_\star)}$ and $\{\delta_\star\} \times (\mathbb{R}^p \setminus \mathsf{B}^{(\delta_\star)})$, together with (A.4) and (A.5) to get

$$\mathbf{1}_{\mathcal{E}}(z)\int \log\left(\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}\right)\mathrm{d}R \leq 2\mathbf{1}_{\mathcal{E}}(z)\left(1-\Pi(\delta_\star|z)\right) \\ + C_0\left(\rho_1\bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}\epsilon^3\right) + \frac{3\rho_1^2(\epsilon+\|\theta_\star\|_2)^2}{2(\rho_1+\bar{\kappa})} \\ + (\rho_1+\bar{\kappa})\mathbf{1}_{\mathcal{E}}(z)\int_{\{\delta_\star\}\times\mathbb{R}^p\setminus\mathsf{B}^{(\delta_\star)}} \|[\theta]_\delta-\hat{\theta}_\delta\|_2^2 R(\mathrm{d}\delta,\mathrm{d}\theta) \\ + 2(\rho_1+\bar{\kappa})\mathbf{1}_{\mathcal{E}}(z)\int_{\{\delta_\star\}\times\mathbb{R}^p\setminus\mathsf{B}^{(\delta_\star)}} \|[\theta]_\delta-\hat{\theta}_\delta\|_2^2 \Pi(\mathrm{d}\delta,\mathrm{d}\theta|Z). \quad \text{(A.6)}$$

By (2.4.2), (2.3.9) and Lemma 15, the last integral in the last display is bounded from above by

$$(C-1)^2\epsilon^2\left(\frac{\rho_1+\bar{\kappa}}{\rho_1+\underline{\kappa}}\right)^{\frac{s_\star}{2}} e^{-\frac{(C-1)^2\epsilon^2\underline{\kappa}}{32}} + 2e^{-p},$$

provided that $\underline{\kappa}(C-1)\epsilon \geq 4\max(\sqrt{s_\star\underline{\kappa}}, \rho_1(\epsilon + s_\star^{1/2}\|\theta_\star\|_\infty))$. We conclude that

$$\mathbf{1}_{\mathcal{E}}(z)\int \log\left(\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}\right)\mathrm{d}R \leq C_0\left(\rho_1\bar{s}^{1/2}\epsilon + \mathsf{a}_2\bar{s}^{3/2}\epsilon^3\right) + \frac{3\rho^2(\epsilon+\|\theta_\star\|_2)^2}{2(\rho_1+\bar{\kappa})} \\ + C_0(\rho_1+\bar{\kappa})\epsilon^2\left(\frac{\rho_1+\bar{\kappa}}{\rho_1+\underline{\kappa}}\right)^{\frac{s_\star}{2}} e^{-\frac{(C-1)^2\epsilon^2\underline{\kappa}}{32}} + 2(\rho_1+\bar{\kappa})e^{-p} + 2\mathbf{1}_{\mathcal{E}}(z)(1-\Pi(\delta_\star|z)) \\ + (\rho_1+\bar{\kappa})\mathbf{1}_{\mathcal{E}}(z)\int_{\{\delta_\star\}\times\mathbb{R}^p\setminus\mathsf{B}^{(\delta_\star)}} \|[\theta]_\delta-\hat{\theta}_\delta\|_2^2 R(\mathrm{d}\delta,\mathrm{d}\theta). \quad \text{(A.7)}$$

In the particular case where $R = \Pi_\star^{(\infty)}$, Lemma 15 gives

$$\int_{\{\delta_\star\}\times\mathbb{R}^p\setminus\mathsf{B}^{(\delta_\star)}} \|[\theta]_\delta-\hat{\theta}_\delta\|_2^2 R(\mathrm{d}\delta,\mathrm{d}\theta) \leq (C-1)^2\epsilon^2\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^{\frac{s_\star}{2}} e^{-\frac{(C-1)^2\epsilon^2\underline{\kappa}}{32}}. \quad \text{(A.8)}$$

The result follows by plugging the last inequality in (A.7). We note that the last display also holds true if $R = \tilde{\Pi}_\star^{(\infty)}$.

□

A.6 Proof of Variational Approximation (Theorem 5)

We introduce

$$\tilde{Q}(\delta, \mathrm{d}\theta) \propto \tilde{Q}(\delta) e^{-\frac{1}{2}(\theta-\hat{\theta}_\star)'\left(\mathcal{S}\cdot\bar{\mathcal{I}}\right)(\theta-\hat{\theta}_\star)}\mathrm{d}\theta,$$

for some arbitrary distribution $\tilde{Q}$ on Δ of the form $\tilde{Q}(\delta) = \prod_{j=1}^{p} \alpha_j^{\delta_j}(1-\alpha_j)^{1-\delta_j}$, where $\alpha_j = \alpha$ if $\delta_{\star j} = 1$, and $\alpha_j = 1-\alpha$ otherwise, for some $\alpha \in (0,1)$. Note that $\tilde{Q} \in \mathcal{Q}$, and $\|\tilde{Q} - \tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}} \to 0$, as $\alpha \to 1$.

The strong convexity of the KL-divergence (Lemma 16) allows us to write, for any $t \in (0,1)$,

$$t\mathsf{KL}\left(Q|\Pi\right) + (1-t)\mathsf{KL}\left(\tilde{Q}|\Pi\right) \geq \mathsf{KL}\left(tQ + (1-t)\tilde{Q}|\Pi\right) + \frac{t(1-t)}{2}\|\tilde{Q} - Q\|_{\mathrm{tv}}^2.$$

This implies that

$$\frac{t(1-t)}{2}\|\tilde{Q} - Q\|_{\mathrm{tv}}^2 \leq \mathsf{KL}\left(\tilde{Q}|\Pi\right) + t\left(\mathsf{KL}\left(Q|\Pi\right) - \mathsf{KL}\left(\tilde{Q}|\Pi\right)\right) \leq \mathsf{KL}\left(\tilde{Q}|\Pi\right),$$

where the second inequality uses the fact that $\tilde{Q} \in \mathcal{Q}$, and Q is the minimizer of the KL-divergence over that family. Hence with $t = 1/2$ we have

$$\begin{aligned} \|Q - \tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2 &\leq 2\|Q - \tilde{Q}\|_{\mathrm{tv}}^2 + 2\|\tilde{Q} - \tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2 \\ &\leq 16\mathsf{KL}\left(\tilde{Q}|\Pi\right) + 2\|\tilde{Q} - \tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2, \end{aligned}$$

where the second inequality uses the bound on $\|\tilde{Q}-Q\|_{\mathrm{tv}}^2$ obtained above.

$$\mathsf{KL}\left(\tilde{Q}|\Pi\right) = \int \log\left(\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{Q}$$

$$= \int_{(\delta_\star\times\mathbb{R}^p)^c} \log\left(\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{Q} + \int_{\delta_\star\times\mathbb{R}^p} \log\left(\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{Q}.$$

We note that $\tilde{\Pi}_\star^{(\infty)}$ is precisely the restriction of $\tilde{Q}$ on $\{\delta_\star\}\times\mathbb{R}^p$. Therefore, on $\{\delta_\star\}\times\mathbb{R}^p$, the density $\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi}$ can be written as

$$\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi} = \tilde{Q}(\{\delta_\star\}\times\mathbb{R}^p)\frac{\mathrm{d}\tilde{\Pi}_\star^{(\infty)}}{\mathrm{d}\Pi_\star^{(\infty)}}\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}.$$

Hence

$$\int_{\delta_\star\times\mathbb{R}^p} \log\left(\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{Q} \;\le\; \mathsf{KL}\left(\tilde{\Pi}_\star^{(\infty)}|\Pi_\star^{(\infty)}\right) \;+\; \tilde{Q}(\delta_\star)\int_{\delta_\star\times\mathbb{R}^p} \log\left(\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{\Pi}_\star^{(\infty)}.$$

On the other hand,

$$\begin{aligned}
&\int_{(\delta_\star\times\mathbb{R}^p)^c} \log\left(\frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{Q}\\
&\qquad= \sum_{\delta\neq\delta_\star}\tilde{Q}(\delta)\left[\log\left(\frac{\tilde{Q}(\delta)}{\Pi(\delta|z)}\right) + \int\log\left(\frac{\tilde{Q}(\theta)}{\Pi(\theta|\delta,z)}\right)\tilde{Q}(\theta)\mathrm{d}\theta\right]\\
&\qquad\le \left(1-\tilde{Q}(\delta_\star)\right)\max_{\delta\in\Delta}\left[-\log(\Pi(\delta|z)) + \int\log\left(\frac{\tilde{Q}(\theta)}{\Pi(\theta|\delta,z)}\right)\tilde{Q}(\theta)\mathrm{d}\theta\right]. \quad (A.1)
\end{aligned}$$

Collecting all the terms we obtain

$$\begin{aligned}\|Q-\tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2 \leq 16\mathsf{KL}\left(\tilde{\Pi}_\star^{(\infty)}|\Pi_\star^{(\infty)}\right) + 2\|\tilde{Q}-\tilde{\Pi}_\star^{(\infty)}\|_{\mathrm{tv}}^2 \\ + 16\tilde{Q}(\delta_\star)\int_{\delta_\star\times\mathbb{R}^p}\log\left(\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{\Pi}_\star^{(\infty)} \\ + 16\left(1-\tilde{Q}(\delta_\star)\right)\max_{\delta\in\Delta}\left[-\log(\Pi(\delta|z)) + \int\log\left(\frac{\tilde{Q}(\theta)}{\Pi(\theta|\delta,z)}\right)\tilde{Q}(\theta)\mathrm{d}\theta\right].\end{aligned}$$

Letting $\alpha\to 1$ on both sides yields

$$\|Q-\Pi_\star^{(\infty)}\|_{\mathrm{tv}}^2 \leq 16\mathsf{KL}\left(\tilde{\Pi}_\star^{(\infty)}|\Pi_\star^{(\infty)}\right) + 16\int_{\delta_\star\times\mathbb{R}^p}\log\left(\frac{\mathrm{d}\Pi_\star^{(\infty)}}{\mathrm{d}\Pi}\right)\mathrm{d}\tilde{\Pi}_\star^{(\infty)}.$$

Using Lemma 14, we have

$$\mathsf{KL}\left(\tilde{\Pi}_\star^{(\infty)}|\Pi_\star^{(\infty)}\right) = \frac{\zeta}{2},$$

where $\zeta = \log\left(\frac{\det(\bar{\mathcal{I}})}{\det(\mathcal{S}\cdot\bar{\mathcal{I}})}\right) + \mathsf{Tr}\left(\bar{\mathcal{I}}^{-1}(\mathcal{S}\cdot\bar{\mathcal{I}})\right) - p$. Hence the theorem.

□

A.7 Proof of Corollary 6

We will assume that n satisfies

$$n \geq \left[2s_\star\left(1+\frac{6}{u}\right)+\frac{4}{u}\right]\log(p). \tag{A.1}$$

Problem set up and posterior sparsity We set Z as $Z=[Y,X]$, and under A1, the likelihood is given by $\ell(u;z) = (1/2\sigma^2)\|Y-Xu\|_2^2$. The resulting posterior distribution $\Pi(\cdot|Z)$ on $\Delta\times\mathbb{R}^p$ fits squarely in the framework developed in the dissertation, and we will successively apply to it the different general theorems obtained above. From the expression of the likelihood, we have

$$\nabla\ell(\theta_\star; Z) = \frac{1}{\sigma^2} X'(Y - X\theta_\star),$$

and

$$\mathcal{L}_{\theta_\star}(u; Z) = -\frac{n}{2\sigma^2}(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star), \quad u \in \mathbb{R}^p,$$

which does not depend on Y. Let us first apply Theorem 1. We set

$$\mathcal{G} \overset{\text{def}}{=} \left\{ Z \in \mathbb{R}^{n\times(p+1)} : \max_{1\leq k\leq p,\ k\neq j} |\langle X_k, Y - X\theta_\star\rangle| \leq \sigma\tau\sqrt{4n\log(p)} \right\}.$$

We set

$$\bar{\rho} = 4\frac{\tau}{\sigma}\sqrt{n\log(p)}, \quad \bar{\kappa} = (n/\sigma^2)s_\star\tau^2.$$

From the expressions of $\nabla\ell(\theta_\star; z)$, and $\mathcal{L}_{\theta_\star}(\theta; z)$, it is straightforward to check that $\mathcal{G} \subseteq \mathcal{E}_0$ if we define $\mathcal{E}_0$ in H1 by taking $\bar{\rho}$ and $\bar{\kappa}$ as above. We also note that by the choice of ρ_1 and the conditions $\|\theta_\star\|_\infty = O(1)$, we have $32\|\theta_\star\|_\infty\rho_1 \leq \bar{\rho}$ for all p large enough. To apply Theorem 1, it only remains to check (2.2.1). With $\mathcal{G}_1$ and $\mathcal{L}_{\theta_\star}$ as defined above, we have

$$\begin{aligned}
\mathbb{E}_\star & \left[\mathbf{1}_{\mathcal{G}_1}(Z) e^{\mathcal{L}_{\theta_\star}(u;Z)+\left(1-\frac{\rho_1}{\bar{\rho}}\right)\langle\nabla\ell(\theta_\star;Z),u-\theta_\star\rangle}\right] \\
& \leq e^{-\frac{n}{2\sigma^2}(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star)} \mathbb{E}_\star\left(e^{\frac{1}{\sigma^2}\left(1-\frac{\rho_1}{\bar{\rho}}\right)(Y-X\theta_\star)'X(u-\theta_\star)}|X\right) \\
& = e^{-\frac{n}{2\sigma^2}\left(1-\left(1-\frac{\rho_1}{\bar{\rho}}\right)^2\right)(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star)}, \quad \text{(A.2)}
\end{aligned}$$

where the equality uses the moment generating function of the conditionally Gaussian random variable V. For $u \in \mathbb{R}^p$ such that $\|\delta_\star^c \cdot (u-\theta_\star)\|_1 \leq 7\|\delta_\star \cdot (u-\theta_\star)\|_1$, we have

$$(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star) \geq \underline{\nu}\|\delta_\star\cdot(u-\theta_\star)\|_2^2,$$

Therefore, we conclude from (A.2) that (2.2.1) holds with

$$r_0(x) = \frac{n\underline{\nu}}{\sigma^2}\left(1-\left(1-\frac{\rho_1}{\bar{\rho}}\right)^2\right)x^2 \geq \frac{n\underline{\nu}}{\sigma^2}\frac{\rho_1}{\bar{\rho}}x^2,$$

and hence

$$\mathsf{a}_0 = \frac{4s_\star\sigma^2\rho_1\bar{\rho}}{n\underline{\nu}} \leq C_0,$$

for some absolute constant C_0, as $p\to\infty$, given the choice of n, ρ_1 and $\bar{\rho}$. The condition (2.2.2) is easily seen to hold for $c_0 = 2$. Theorem 1 then gives

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_1}(Z)\Pi\left(\|\delta\|_0 > s_\star\left(1+\frac{6}{u}\right)+\frac{4}{u}|Z\right)\right] \leq \frac{2}{p^2}. \tag{A.3}$$

By a standard union bound argument, and Gaussian tail bounds we can also show

$$\mathbb{P}(Z\notin\mathcal{G}|X) = \mathbb{P}\left(\max_{1\leq k\leq p+1,\ k\neq j}|\langle X_k, V\rangle| > 2\sigma\tau\sqrt{n\log(p)}\;|X\right) \leq \frac{2}{p^2}.$$

Therefore, (A.3) becomes

$$\mathbb{E}_\star\left[\Pi\left(\|\delta\|_0 > s_\star\left(1+\frac{6}{u}\right)+\frac{4}{u}|Z\right)\right] \leq \frac{4}{p^2}. \tag{A.4}$$

Contraction and rate   Set $\bar{s} = s_\star\left(1+\frac{6}{u}\right)+\frac{4}{u}$. We now apply Theorem 2 to Π. With similar calculations as above, for $\|\delta\|_0 \leq \bar{s}$, and $u\in\mathbb{R}^p_\delta$,

$$\mathcal{L}_{\theta_\star}(u;z) \leq -\frac{n\underline{\nu}(s_\star+\bar{s})}{2\sigma^2}\|u-\theta_\star\|_2^2,$$

provided that the sample size n satisfies (A.1) which shows that $\mathcal{G} \subseteq \mathcal{E}_1(\bar{s})$ with the rate function $\mathsf{r}(x) = x^2 n\underline{\nu}(s_\star + \bar{s})/(\sigma^2)$. The contraction rate ϵ then becomes

$$\epsilon = \frac{\sigma^2 \bar{\rho}(\bar{s} + s_\star)^{1/2}}{n\underline{\nu}(s_\star + \bar{s})} = \frac{4\tau\sigma}{\underline{\nu}(s_\star + \bar{s})}\sqrt{\frac{(\bar{s} + s_\star)\log(p)}{n}}.$$

The condition (2.3.4) holds by choosing the absolute constant $C \geq 3$ large enough so that $C\tau^2 \geq 2(1+u)\underline{\nu}(s_\star + \bar{s})$. Theorem 2 then gives

$$\mathbb{E}_\star\left[\Pi\left(\mathsf{B}^c|Z\right)\right] \leq \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(Z)\Pi\left(\mathsf{B}^c|Z\right)\right] + \mathbb{P}(Z \notin \mathcal{G}_1|X) \leq \frac{C_0}{p^2}. \tag{A.5}$$

<u>`Model selection consistency`</u> We now apply Theorem 3 to Π. We set

$$\mathcal{G}_1 \overset{\text{def}}{=} \mathcal{G} \bigcap_{k=1}^{\bar{s}-s_\star} \left\{ Z = [Y, X] \in \mathbb{R}^{n\times(p+1)} : \right.$$
$$\left. \max_{\delta \supseteq \delta_\star,\ \|\delta\|_0 = s_\star + k} (Y - X\theta_\star)'\mathcal{P}_{\delta\backslash\delta_\star}(Y - X\theta_\star) \leq \sigma^2 ku\log(p) \right\},$$

where for $\delta \supseteq \delta_\star$, $\mathcal{P}_{\delta\backslash\delta_\star}$ is the orthogonal projector on the sub-space of $\mathsf{span}(X_\delta)$ that is orthogonal to $\mathsf{span}(X_{\delta_\star})$, where the notation $\mathsf{span}(X_\delta)$ denotes the linear space spanned by the columns of X_δ. Indeed, for $\delta \in \mathcal{A}_{\bar{s}}$, the matrix X_δ is full-rank column. Hence if $X_\delta = Q_{(\delta)}R_{(\delta)}$ is the QR decomposition of X_δ, then

$$\ell^{[\delta]}(\hat{\theta}_\delta; Z) - \ell^{[\delta_\star]}(\hat{\theta}_\star; Z) = \frac{1}{2\sigma^2}\|Q'_{(\delta\backslash\delta_\star)}(Y - X\theta_\star)\|_2^2 = \frac{1}{2\sigma^2}(Y - X\theta_\star)'\mathcal{P}_{\delta\backslash\delta_\star}(Y - X\theta_\star).$$

It then follows that $\mathcal{G}_2 \subseteq \mathcal{E}_2(\bar{s})$. Furthermore, since ℓ is quadratic, (2.3.8) holds with $\underline{\kappa} = n\underline{\nu}(\bar{s})/(\sigma^2)$, and (2.3.9) holds with $\bar{\kappa} = (n/\sigma^2)s_\star\tau^2$. Theorem 3 (applied

$\mathsf{a}_2 = 0$), and (A.5) give for all $k \geq 0$,

$$\begin{aligned}\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_1}(Z)\Pi\left(\mathsf{B}_k^c|Z\right)\right] \leq C_0\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{u/2}}\right)^{k+1} &+ \mathbb{E}_\star\left[1_{\mathcal{G}}(Z)\Pi(\mathsf{B}^c|Z)\right] \\ &\leq C_0\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{u/2}}\right)^{k+1} + \frac{C_0}{p^2}. \quad \text{(A.6)}\end{aligned}$$

Hence we write

$$\mathbb{E}_\star\left[\Pi\left(\mathsf{B}_k^c|Z\right)\right] \leq \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_1}(Z)\Pi\left(\mathsf{B}_k^c|Z\right)\right] + \mathbb{P}_\star\left[Z \notin \mathcal{G}_1|X\right].$$

Given $\delta \in \mathcal{A}_{s_\star+k}$, by the Hanson-Wright inequality (Lemma 17),

$$\begin{aligned}&\mathbb{P}\left((Y - X\theta_\star)'\mathcal{P}_{\delta\setminus\delta_\star}(Y - X\theta_\star) > \sigma^2 ku\log(p)|X\right) \\ &\qquad\qquad = \mathbb{P}\left(V'\mathcal{P}_{\delta\setminus\delta_\star}V > ku\log(p)|X\right) \leq \frac{1}{p^{\frac{uk}{4}}},\end{aligned}$$

for all p large enough. Hence by union bound, for $u \geq 8$,

$$\mathbb{P}(Z \notin \mathcal{G}_2|X) \leq \mathbb{P}(Z \notin \mathcal{G}_1|X) + \sum_{k\geq 1}\frac{1}{p^{\frac{uk}{4}}} \leq \frac{4}{p^2}.$$

We conclude that for all $k \geq 0$,

$$\mathbb{E}_\star\left[Pi\left(\mathsf{B}_k^c|Z\right)\right] \leq C_0\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{u/2}}\right)^{k+1} + \frac{C_0}{p^2}. \quad \text{(A.7)}$$

Bernstein-von Mises approximation and variational approximations Taking $k = 0$ in (A.7) together with Theorem 4 gives

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(Z)\mathsf{KL}\left(\Pi_\star^{(\infty)}|\Pi\right)\right] \leq C_0(\bar{s} + s_\star)\frac{\log(p)}{n} + \frac{C_0}{p^{\frac{u}{2}-1}} + \frac{C_0}{p},$$

for some absolute constant C_0, assuming that $u > 2$. Finally we apply (2.4.7) and (A.8) applied with $R = \tilde{\Pi}_\star^{(\infty)}$ to get the stated controls on the variational approximations. This ends the proof. □

A.8 Proof of Corollary 7

<u>`On the event` $\mathcal{G}$</u> We first constructed the event $\mathcal{G}$. Let $\tau_\Sigma \overset{\text{def}}{=} \max_j \Sigma_{jj}$. For $c_1 = 5$, $c_2 = 1/4$, and $c_3 = 9$, for $j = 1, \ldots, p+1$, we set $\mathcal{G} \overset{\text{def}}{=} \bigcap_{j=1}^{p+1} \mathcal{H}^{(j)}$, where

$$\mathcal{H}^{(j)} \overset{\text{def}}{=} \Bigg\{ Z \in \mathbb{R}^{n\times(p+1)} : \max_{1\leq k\leq p,\ k\neq j} \left| \frac{\|Z_k\|_2^2}{n} - \Sigma_{jj} \right| \leq c_1 \tau_\Sigma$$

$$\text{for all } v \in \mathbb{R}^p : \frac{\|X^{(j)} v\|_2}{\sqrt{n}} \geq c_2 \|\Sigma^{1/2} v\|_2 - c_3 \tau_\Sigma \sqrt{\frac{\log(p)}{n}} \|v\|_1 \Bigg\}.$$

When B1 holds, by Theorem 1 of Raskutti et al. [2010] and Lemma 1 of Ravikumar et al. [2011] there exist absolute positive constant c_4, c_5 such that

$$\mathbb{P}(Z \notin \mathcal{G}) \leq 4(p+1)e^{-n/128} + c_4(p+1)e^{-c_5 n} \to 0,$$

as $p \to \infty$, provided that $n \geq (256/\min(1, 128c_5))\log(p)$. In what follows we will assume that n satisfies

$$n \geq \frac{256}{\min(1, 128c_5)} \log(p), \quad \text{and } n \geq \left(\frac{16 c_3 \tau_\Sigma}{c_2 \lambda_{\mathsf{min}}^{1/2}(\Sigma)} \right)^2 \left[\max_j 2 s_\star^{(j)} \left(1 + \frac{6}{u}\right) + \frac{4}{u} \right] \log(p). \tag{A.1}$$

`Problem set up and posterior sparsity` For any j we can partition Z as $Z = [Y^{(j)}, X^{(j)}]$, and under B1,

$$Y^{(j)} = X^{(j)}\theta_\star^{(j)} + \frac{1}{\sqrt{[\vartheta_\star]_{jj}}} V^{(j)}, \quad \text{where } V^{(j)}|X^{(j)} \sim \mathbf{N}_n(0, I_n). \tag{A.2}$$

The quasi-likelihood of the j-th regression is $\ell^{(j)}(u; z) = (1/2\sigma_j^2)\|Y^{(j)} - X^{(j)}u\|_2^2$. Again, the resulting quasi-posterior distribution $\Pi^{(j)}(\cdot|Z)$ on $\Delta \times \mathbb{R}^p$ fits squarely in the framework developed in the dissertation , and we proceed to successively apply to it the different general theorems obtained above. However to keep the notation simple, and when there is no risk of confusion, we shall omit the index j from the various quantities. For instance we will Y instead of $Y^{(j)}$, X instead of $X^{(j)}$, etc. From the expression of the quasi-likelihood, we have

$$\nabla \ell(\theta_\star; Z) = \frac{1}{\sigma^2} X'(Y - X\theta_\star),$$

and

$$\mathcal{L}_{\theta_\star}(u; Z) = -\frac{n}{2\sigma^2}(u - \theta_\star)' \left(\frac{X'X}{n}\right)(u - \theta_\star), \;\; u \in \mathbb{R}^p,$$

which does not depend on Y. Let us first apply Theorem 1. We set

$$\mathcal{G}_1 \stackrel{\text{def}}{=} \mathcal{G} \bigcap \big\{Z = [Y^{(j)}, X^{(j)}] \in \mathbb{R}^{n\times(p+1)} : \\ \max_{1\leq k\leq p,\, k\neq j} \left|\left\langle X_k, Y^{(j)} - X^{(j)}\theta_\star^{(j)}\right\rangle\right| \leq \sqrt{\frac{6\tau_\Sigma}{[\vartheta_\star]_{jj}}(1 + c_1)n\log(p)} \Bigg\}.$$

We set

$$\bar{\rho} = \frac{2}{\sigma_j^2}\sqrt{\frac{6\tau_\Sigma}{[\vartheta_\star]_{jj}}(1 + c_1)n\log(p)}, \quad \bar{\kappa} = (n/\sigma^2)(1 + c_1)s_\star^{(j)}\tau_\Sigma.$$

We stress again that these quantities and events are specific to the j-th regression. From the expressions of $\nabla\ell(\theta_\star; z)$, and $\mathcal{L}_{\theta_\star}(\theta; z)$, it is straightforward to check that $\mathcal{G}_1 \subseteq \mathcal{E}_0$ if we define $\mathcal{E}_0$ in H1 by taking $\bar{\rho}$ and $\bar{\kappa}$ as above. We also note that by the choice of ρ_1 and the conditions $\|\theta_\star\|_\infty = O(1)$, we have $32\|\theta_\star\|_\infty\rho_1 \leq \bar{\rho}$ for all p large enough. To apply Theorem 1, it only remains to check (2.2.1). With $\mathcal{G}_1$ and $\mathcal{L}_{\theta_\star}$ as defined above, we have

$$\begin{aligned}
&\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_1}(Z)e^{\mathcal{L}_{\theta_\star}(u;Z)+\left(1-\frac{\rho_1}{\bar{\rho}}\right)\langle\nabla\ell(\theta_\star;Z),u-\theta_\star\rangle}\right] \\
&\quad\leq \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(X)e^{-\frac{n}{2\sigma^2}(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star)}\mathbb{E}_\star\left(e^{\frac{1}{\sigma^2}\left(1-\frac{\rho_1}{\bar{\rho}}\right)(Y-X\theta_\star)'X(u-\theta_\star)}|X\right)\right] \\
&\qquad\qquad = \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(X)e^{-\frac{n}{2\sigma^2}\left(1-\frac{\left(1-\frac{\rho_1}{\bar{\rho}}\right)^2}{\sigma^2\vartheta_{\star,11}}\right)(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star)}\right], \quad \text{(A.3)}
\end{aligned}$$

where the equality uses the moment generating function of the conditionally Gaussian random variable V. For $u \in \mathbb{R}^p$ such that $\|\delta_\star^c \cdot (u-\theta_\star)\|_1 \leq 7\|\delta_\star \cdot (u-\theta_\star)\|_1$, and for $Z \in \mathcal{G}$, we have

$$\frac{1}{\sqrt{n}}\|X(u-\theta_\star)\|_2 \geq c_2\lambda_{\mathsf{min}}(\Sigma)^{1/2}\|u-\theta_\star\|_2 - 8c_3 s_\star^{1/2}\tau_\Sigma\sqrt{\frac{\log(p)}{n}}\|(\delta_\star \cdot (u-\theta_\star)\|_2.$$

It follows that

$$(u-\theta_\star)'\left(\frac{X'X}{n}\right)(u-\theta_\star) \geq \frac{c_2^2}{4}\lambda_{\mathsf{min}}(\Sigma)\|\delta_\star \cdot (u-\theta_\star)\|_2^2,$$

if the sample size n satisfies

$$n \geq \left(\frac{16c_3\tau_\Sigma}{c_2\lambda_{\mathsf{min}}^{1/2}(\Sigma)}\right)^2 s_\star \log(p).$$

Therefore, Since $\sigma^2[\vartheta_\star]_{jj} \geq 1$, we conclude from (A.3) that (2.2.1) holds with

$$r_0(x) = \frac{nc_2^2\lambda_{\mathsf{min}}(\Sigma)}{4\sigma^2}\left(1 - \left(1 - \frac{\rho_1}{\bar{\rho}}\right)^2\right)x^2 \geq \frac{nc_2^2\lambda_{\mathsf{min}}(\Sigma)}{4\sigma^2}\frac{\rho_1}{\bar{\rho}}x^2,$$

and hence

$$\mathsf{a}_0 = \frac{16 s_\star \sigma^2 \rho_1 \bar{\rho}}{nc_2^2\lambda_{\mathsf{min}}(\Sigma)} \leq C_0,$$

for some absolute constant C_0, as $p \to \infty$, given the choice of n, ρ_1 and $\bar{\rho}$. The condition (2.2.2) is easily seen to hold for $c_0 = 2$. Theorem 1 then gives

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_1}(Z)\Pi\left(\|\delta\|_0 > s_\star\left(1 + \frac{6}{u}\right) + \frac{4}{u}|Z\right)\right] \leq \frac{2}{p^2}. \tag{A.4}$$

Since $Y = X\theta_\star + \frac{1}{\sqrt{[\vartheta_\star]_{jj}}}V$, where $V|X \sim \mathbf{N}(0, I_n)$, by a standard union bound argument, and Gaussian tail bounds

$$\begin{aligned}&\mathbf{1}_{\mathcal{G}}(X)\mathbb{P}(Z \notin \mathcal{G}_1|X)\\ &\qquad = \mathbf{1}_{\mathcal{G}}(X)\mathbb{P}\left(\max_{1\leq k\leq p+1,\, k\neq j} |\langle X_k, V\rangle| > \sqrt{6\tau_\Sigma(1+c_1)n\log(p)}\; |X\right) \leq \frac{2}{p^2}.\end{aligned}$$

Therefore, (A.4) becomes

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{H}}(X)\Pi\left(\|\delta\|_0 > s_\star\left(1 + \frac{6}{u}\right) + \frac{4}{u}|Z\right)\right] \leq \frac{4}{p^2}. \tag{A.5}$$

<u>`Contraction and rate`</u> Set $\bar{s} = s_\star\left(1 + \frac{6}{u}\right) + \frac{4}{u}$. We now apply Theorem 2 to $\Pi^{(j)}$. With similar calculations as above, for $\|\delta\|_0 \leq \bar{s}$, and $u \in \mathbb{R}^p_\delta$,

$$\mathcal{L}_{\theta_\star}(u; z) \leq -\frac{nc_2^2\lambda_{\mathsf{min}}(\Sigma)}{8\sigma^2}\|u - \theta_\star\|_2^2,$$

provided that the sample size n satisfies (A.1) which shows that $\mathcal{G}_1 \subseteq \mathcal{E}_1(\bar{s})$ with the rate function $\mathsf{r}(x) = x^2 n c_2^2 \lambda_{\mathsf{min}}(\Sigma)/(4\sigma^2)$. The contraction rate ϵ then becomes

$$\epsilon = \frac{4\sigma^2 \bar{\rho}(\bar{s}+s_\star)^{1/2}}{nc_2^2\lambda_{\mathsf{min}}(\Sigma)} = \frac{8\sqrt{2(1+c_1)}}{c_2^2}\frac{\tau_\Sigma^{1/2}}{\lambda_{\mathsf{min}}(\Sigma)[\vartheta_\star]_{jj}^{1/2}}\sqrt{\frac{(\bar{s}+s_\star)\log(p)}{n}}.$$

The condition (2.3.4) holds by choosing the absolute constant $C \geq 3$ large enough so that $C(1+c_1)\tau_\Sigma \geq (1+u)c_2^2\lambda_{\mathsf{min}}(\Sigma)\sigma^2[\vartheta_\star]_{jj}$. Theorem 2 then gives

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(X)\Pi\left(\mathsf{B}^c|Z\right)\right] \leq \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_1}(Z)\Pi\left(\mathsf{B}^c|Z\right)\right] + \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(X)\mathbb{P}(Z \notin \mathcal{G}_1|X)\right] \leq \frac{C_0}{p^2}. \quad \text{(A.6)}$$

<u>`Model selection consistency`</u> We now apply Theorem 3 to $\Pi^{(j)}$ With $\bar{s} = \bar{s}^{(j)}$ as above, set

$$\mathcal{G}_2 \overset{\text{def}}{=} \mathcal{G}_1 \bigcap_{k=1}^{\bar{s}-s_\star} \Big\{ Z = [Y, X] \in \mathbb{R}^{n\times(p+1)} : \\ \max_{\delta \supseteq \delta_\star,\ \|\delta\|_0 = s_\star + k} (Y - X\theta_\star)'\mathcal{P}_{\delta\setminus\delta_\star}(Y - X\theta_\star) \leq \sigma^2 k u \log(p) \Big\},$$

where for $\delta \supseteq \delta_\star$, $\mathcal{P}_{\delta\setminus\delta_\star}$ is the orthogonal projector on the sub-space of $\mathsf{span}(X_\delta)$ that is orthogonal to $\mathsf{span}(X_{\delta_\star})$, where the notation $\mathsf{span}(X_\delta)$ denotes the linear space spanned by the columns of X_δ. We note that $\mathcal{G}_2 \subseteq \mathcal{E}_2(\bar{s})$. Indeed, for $\delta \in \mathcal{A}_{\bar{s}}$, and $X \in \mathcal{G}$, the matrix X_δ is full-rank column. Hence if $X_\delta = Q_{(\delta)}R_{(\delta)}$ is the QR decomposition of X_δ, then

$$\ell^{[\delta]}(\hat{\theta}_\delta; Z) - \ell^{[\delta_\star]}(\hat{\theta}_\star; Z) = \frac{1}{2\sigma^2}\|Q'_{(\delta\setminus\delta_\star)}(Y - X\theta_\star)\|_2^2 = \frac{1}{2\sigma^2}(Y - X\theta_\star)'\mathcal{P}_{\delta\setminus\delta_\star}(Y - X\theta_\star).$$

It then follows that $\mathcal{G}_2 \subseteq \mathcal{E}_2(\bar{s})$. Furthermore, since ℓ is quadratic, (2.3.8) holds with $\underline{\kappa} = nc_2^2\lambda_{\mathsf{min}}(\Sigma)/(4\sigma^2)$, and (2.3.9) holds with $\bar{\kappa} = (n/\sigma^2)(1+c_1)s_\star^{(j)}\tau_\Sigma$, provided that the sample size condition (A.1) holds. Theorem 3 (applied $\mathsf{a}_2 = 0$), and (A.5) give for all $k \geq 0$,

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_2}(Z)\Pi\left(\mathsf{B}_k^c|Z\right)\right] \leq C_0\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{u/2}}\right)^{k+1} + \mathbb{E}_\star\left[1_{\mathcal{G}_1}(Z)\Pi(\mathsf{B}^c|Z)\right]$$
$$\leq C_0\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{u/2}}\right)^{k+1} + \frac{C_0}{p^2}. \quad \text{(A.7)}$$

To replace $\mathcal{G}_2$ by $\mathcal{G}$, we write

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(X)\Pi\left(\mathsf{B}_k^c|Z\right)\right] \leq \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_2}(Z)\Pi\left(\mathsf{B}_k^c|Z\right)\right] + \mathbb{P}_\star\left[X \in \mathcal{G}, Z \notin \mathcal{G}_2\right].$$

Given $\delta \in \mathcal{A}_{s_\star+k}$, by the Hanson-Wright inequality (Lemma 17),

$$\mathbf{1}_{\mathcal{G}}(X)\mathbb{P}\left((Y - X\theta_\star)'\mathcal{P}_{\delta\setminus\delta_\star}(Y - X\theta_\star) > \sigma^2 ku\log(p)|X\right)$$
$$= \mathbf{1}_{\mathcal{G}}(X)\mathbb{P}\left(V'\mathcal{P}_{\delta\setminus\delta_\star}V > \sigma^2[\vartheta_\star]_{jj}ku\log(p)|X\right) \leq \frac{1}{p^{\frac{\sigma^2[\vartheta_\star]_{jj}uk}{4}}},$$

for all p large enough. Hence by union bound, for $\sigma^2[\vartheta_\star]_{jj}u \geq 8$,

$$\mathbf{1}_{\mathcal{G}}(X)\mathbb{P}(Z \notin \mathcal{G}_2|X) \leq \mathbf{1}_{\mathcal{G}}(X)\mathbb{P}(Z \notin \mathcal{G}_1|X) + \sum_{k\geq 1}\frac{1}{p^{\frac{\sigma^2[\vartheta_\star]_{jj}uk}{4}}} \leq \frac{4}{p^2}.$$

We conclude that for all $k \geq 0$,

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(X)\Pi\left(\mathsf{B}_k^c|Z\right)\right] \leq C_0\left(\sqrt{\frac{\rho_1}{\underline{\kappa}}}\frac{1}{p^{u/2}}\right)^{k+1} + \frac{C_0}{p^2}. \quad \text{(A.8)}$$

Bernstein-von Mises approximation and variational approximations Taking $k = 0$ in (A.8) together with Theorem 4 gives

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}}(Z)\max_{1\leq j\leq p+1}\mathsf{KL}\left(\Pi_\star^{(j,\infty)}|\Pi^{(j)}\right)\right] \leq \frac{C_0\max_j(\bar{s}^{(j)}+s_\star^{(j)})}{\min_j[\vartheta_\star]_{jj}}\frac{\log(p)}{n}+\frac{C_0}{p^{\frac{u}{2}-1}}+\frac{C_0}{p},$$

for some absolute constant C_0, assuming that $\sigma^2[\vartheta_\star]_{jj}u \geq 16$, and $u > 2$. Finally we apply (2.4.7) and (A.8) applied with $R = \tilde{\Pi}_\star^{(\infty)}$ to get the stated controls on the variational approximations. This ends the proof.

□

A.9 Proof of Corollary 8

The theoretical properties discussed in Chapter 2 are based on nice curvature of the Bregmann divergence which is given as

$$\mathcal{L}_{\theta_\star}(u;z) = -\sum_{i=1}^n\left[g(\langle x_i,u\rangle) - g(\langle x_i,\theta_\star\rangle) - g^{(1)}(\langle x_i,\theta_\star\rangle)\langle x_i,u-\theta_\star\rangle\right]$$

We start by verifying Assumption H1. Using Taylor's expansion, $g^{(2)}(x) \leq 1/4$, $\forall x \in \mathbb{R}$, and $u \in \mathbb{R}^p_{\delta_\star}$ we have,

$$\mathcal{L}_{\theta_\star}(u;z) \geq -\frac{n}{8}(u-\theta_\star)'\frac{X'X}{n}(u-\theta_\star) \geq -\frac{n\bar{v}(s_\star)}{8}\|u-\theta_\star\|_2^2$$

Contraction rate We introduce the set

$$\mathcal{G}_0 \overset{\text{def}}{=} \{Y \in \{0,1\}^n : \max_{1\leq j\leq p}|\sum_{i=1}^n\left(x_{ji}y_i - x_i'g^{(1)}(\langle x_i,\theta_\star\rangle)\right)| \leq \frac{\bar{\rho}}{2}\}. \qquad \text{(A.1)}$$

For $\|X\|_\infty \leq \mathbf{b}$, $\bar{\kappa} = \frac{n\bar{v}(s_\star)}{8}$ and $\bar{\rho} = 4\mathbf{b}\sqrt{n\log(p)}$, we can show using sub-gaussian tail bounds that

$$\mathbb{P}_\star(Y \notin \mathcal{G}_0|X) \leq 2\exp\left[\log(p) - \frac{\bar{\rho}^2}{8n\|X\|_\infty^2}\right] \leq \frac{2}{p}$$

We use the proof of Theorem 4 in supplementary of Atchade [2017] to show that for all $\delta \in \Delta(\bar{s}), u \in \mathbb{R}^p_\delta$

$$\begin{aligned}\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{G}_0}(Z)e^{\mathcal{L}_{\theta_\star}(u;Z)}|X\right] &\leq \exp[-\frac{n\underline{v}(s_\star+\bar{s})\|u-\theta_\star\|_2^2}{2+\sqrt{s_\star+\bar{s}}\mathbf{b}\|u-\theta_\star\|_2}] \\ &\leq e^{-\frac{1}{2}r(|u-\theta_\star\|_2)},\end{aligned}$$

where $r(x) = \frac{n\underline{v}(s_\star+\bar{s})x^2}{1+\sqrt{s_\star+\mathbf{b}\bar{s}}x/2}$.
If $n\underline{v}(s_\star+\bar{s}) > (s_\star+\bar{s})\bar{\rho}\mathbf{b}$, then $\epsilon = \frac{\sqrt{s_\star+\bar{s}}\bar{\rho}}{n\underline{v}(s_\star+\bar{s})-(s_\star+\bar{s})\bar{\rho}\mathbf{b}}$

For $n\underline{v}(s_\star+\bar{s}) \geq (s_\star+\bar{s})\bar{\rho}\mathbf{b}$, we can show

$$\epsilon \leq \frac{16\mathbf{b}}{\underline{v}(s_\star+\bar{s})}\sqrt{\frac{(s_\star+\bar{s})\log(p)}{n}} \leq \infty$$

The sample size condition translates as

$$\sqrt{n} \geq (16/3)\mathbf{b}^2\frac{(s_\star+\bar{s})}{\underline{v}(s_\star+\bar{s})}\sqrt{\log(p)}$$

Given choice of $\bar{\rho}$, we have

$$\epsilon \geq \frac{\sqrt{s_\star+\bar{s}}\bar{\rho}}{n\underline{v}(s_\star+\bar{s})} = \frac{4\mathbf{b}}{\underline{v}(s_\star+\bar{s})}\sqrt{\frac{(s_\star+\bar{s})\log(p)}{n}}$$

For $C > 3$, $\rho_1 \sim \sqrt{\log(p)/n}$ and $\bar{\rho} = 4\mathbf{b}\sqrt{n\log(p)}$, we satisfy condition 2.3.4

and can proceed to apply Theorem 2 giving the bound

$$
\begin{aligned}
\mathbb{E}_\star[\Pi(B^c|Z)] &\leq \mathbb{E}_\star[\mathbf{1}_{\mathcal{G}_0}(Z)\Pi(B^c|Z)] + \mathbb{P}_\star(Z \notin \mathcal{G}_0)] \\
&\leq 8\exp[-C\frac{s_\star + \bar{s}}{2\underline{v}(s_\star + \bar{s})}\mathbf{b}^2 \log(p)] + \frac{2}{p} + 2e^{-p} \quad \text{(A.2)}
\end{aligned}
$$

Remark A.2. Note that we can drop the term $\Pi(\|\delta\|_0 \geq \bar{s}|Z)$ in the application of Theorem 2 by virtue of the hard sparsity induced by the prior in C2.

<u>`Model Selection Consistency`</u> To show model selection consistency we need to construct a set analogous to $\mathcal{E}_2(\bar{s})$ defined for Theorem 3 in Chapter 2. For a fixed $\bar{s}$, we define

$$
\begin{aligned}
\mathcal{G}_1 = \mathcal{G}_0 \cap \bigcap_{k=k_\star}^{\bar{s}-s_\star} \Bigg\{ Z : \max_{\delta \in \mathcal{A}:\, \|\delta\|_0 = s_\star + k} \Bigg(\sum_{i=1}^n y_i \langle x_{\delta i}, \hat{\theta}_\delta - \hat{\theta}_{\delta_\star} \rangle \\
- \Big(g(\langle x_{\delta i}, \hat{\theta}_\delta \rangle) - g(\langle x_{\delta i}, \hat{\theta}_{\delta_\star} \rangle) \Big) \Bigg) \leq \log(p)\frac{uk}{2} \Bigg\} \quad \text{(A.3)}
\end{aligned}
$$

The growth condition of $\mathcal{G}_1$ is hard to verify in case of non-gaussian models. However under certain assumptions we use Lemma 13 to show that $\mathcal{G}_1$ occurs with high probability under the true model.

Under Assumption C1-3, a direct application of Lemma 13 yields, for some absolute Constant $C_0 > 0$,

$$
\begin{aligned}
P_\star(Z \notin \mathcal{G}_1 | \mathcal{G}_0, X) &\leq \sum_{k \geq 1} 2\exp\left[-\log(p)\frac{ku}{C_0} \right] \\
&\leq 4\exp\left[-\log(p)\frac{u}{C_0} \right] \leq \frac{4}{p^{C_1 u}},
\end{aligned}
$$

Hence $P_\star(Z \notin \mathcal{G}_1 | \mathcal{G}_0, X) \to 0$ as $p \to \infty$

The operator norm a_2 introduced in Theorem (3) can be computed as follows

$$\varpi(\delta, (C+1)\epsilon; z) = \sup_{u \in \mathbb{R}^{\|\delta\|_0} : \|u - \hat{\theta}_\delta\|_2 \leq (C+1)\epsilon} \max_{1 \leq i,j,k \leq \|\delta\|_0} \left| -\sum_{t=1}^{n} x_{it} x_{jt} x_{kt} p_t(u)(1 - p_t(u))(1 - 2p_t(u)) \right|$$

where $p_t(u) = \frac{\exp(\langle x_t, u \rangle)}{1+\exp(\langle x_t, u \rangle)}$. Therefore by Assumption C1

$$\begin{aligned} \varpi(\delta, (C+1)\epsilon; z) &\leq \max_{1 \leq i,j,k \leq \|\delta\|_0} \sum_{t=1}^{n} |x_{it} x_{jt} x_{kt}| \\ &\leq \max_{1 \leq i,j,k \leq \|\delta\|_0} \|X_i\|_2 \|X_j\|_4 \|X_k\|_4 \leq n\mathbf{b}^3 \end{aligned}$$

From the above expression it can be deduced that $\mathsf{a}_2 = \sup_{\delta \in \mathcal{A}_{\bar{s}}} \varpi(\delta, (C+1)\epsilon; z) \leq n\mathbf{b}^3$ Further it can be shown that in the context of Theorem 3 $\underline{\kappa} = n\underline{v}_1(\bar{s})$ and $\bar{\kappa} = n\bar{v}(s_\star)$ Hence for $k > 1$

$$\mathbb{E}_\star[1_{\mathcal{G}_1} \Pi(B_k^c)|X] \leq \mathbb{E}_\star[1_{\mathcal{G}_0} \Pi(B^c)|X] + C_0 e^{\frac{1}{\underline{v}_1(\bar{s})}\sqrt{\frac{(\bar{s}\log(p))^3}{n}}} \left(\sqrt{\frac{\rho_1}{n\underline{v}_1(\bar{s})}} \frac{1}{p^{u/2}} \right)^{k+1}$$

Hence for $n\underline{v}_1(\bar{s})^2 > (\bar{s}\log(p))^3$,

$$\begin{aligned} \mathbb{E}_\star[\Pi(B_k^c|Z)|X] &\leq \mathbb{E}_\star[1_{\mathcal{G}_0} \Pi(B_k^c|Z)|X] + \mathbb{P}_\star(Z \notin \mathcal{G}_0|X) \\ &\leq \mathbb{E}_\star[1_{\mathcal{G}_1} \Pi(B_k^c|Z)|X] + \mathbb{P}_\star(Z \notin \mathcal{G}_1|\mathcal{G}_0, X) + \mathbb{P}_\star(Z \notin \mathcal{G}_0|X) \\ &\leq C_0 \left(\sqrt{\frac{\rho_1}{n\underline{v}_1(\bar{s})}} \frac{1}{p^{u/2}} \right)^{k+1} + \mathbb{E}_\star[1_{\mathcal{G}_0} \Pi(B^c)|X] + \frac{4}{p^{C_1 u}} + \frac{C_2}{p} \end{aligned}$$

`Bernstein Von Mises phenomenon` In the context of logistic regression the application of Theorem4 would give the following bound.

$$\mathbf{1}_{\mathcal{G}_2}(z)\mathsf{KL}\left(\Pi_\star^{(\infty)}|\Pi\right) \le C_0\left(\rho_1\bar{s}^{1/2}\epsilon + n\mathsf{b}^3\bar{s}^{3/2}\epsilon^3\right) + \frac{3\rho_1^2(\epsilon+\|\theta_\star\|_2)^2}{2(\rho_1+\bar{\kappa})}$$
$$+ C_0(\rho_1+\bar{\kappa})\epsilon^2\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^{\frac{s_\star}{2}} e^{-\frac{(C-1)^2\epsilon^2\underline{\kappa}}{32}} + C_0(\rho_1+\bar{\kappa})e^{-p} + 2\mathbf{1}_{\mathcal{E}}(z)(1-\Pi(\delta_\star|z)), \quad (A.4)$$

Note that $\bar{\kappa} = n\bar{v}(s_\star)/8$ and $\underline{\kappa} = n\underline{v}_2(\bar{s})$. assuming $\frac{\bar{v}(s_\star)}{\underline{v}_2(\bar{s})} \sim O(1)$, for given choice of ϵ and ρ_1, the dominating term in the bound appears from the first terms. Hence,

$$\mathbb{E}_\star(1_{\mathcal{G}_2}KL(\Pi_\star^{(\infty)}|\Pi)) \le C_0\frac{(s_\star+\bar{s})\log(p)}{n\underline{v}(s_\star+\bar{s})} + n\mathbf{b}^3(\bar{s}+s_\star)^3\left(\frac{\log(p)}{n}\right)^{(3/2)}$$

thus ending the proof of Corollary 8

A.10 Proof of Corollary 9

The proof follows the same steps as in the proof of Theorem 2. Let

$$\bar{\rho} = \frac{8C_0\vartheta}{\sigma^2}\sqrt{n\left(\frac{p}{\vartheta}+\log(p)\right)}, \quad \bar{\kappa} = \frac{c_1 n}{\sigma^2}, \quad \mathsf{r}(x) = \frac{c_2 n}{\sigma^2}x^2,$$
$$\text{and} \quad \epsilon = \frac{8C_0\vartheta}{c_2}\sqrt{\frac{\frac{p}{\vartheta}+\log(p)}{n}}(\bar{s}+s_\star),$$

for some absolute constants C_0, c_1, c_2, that we specify later. For $\theta_0 \in \{\theta_\star, -\theta_\star\}$, let B_{θ_0} be the set B defined in (2.3.2) but with $\theta_\star$ replaced by θ_0, ϵ as above, and for some absolute constant C, C_1. Similarly let $\mathcal{E}_{0,\theta_0}$ (resp. $\mathcal{E}_{1,\theta_0}(\bar{s})$) be the set $\mathcal{E}_0$ (resp. $\mathcal{E}_1(\bar{s})$) but with $\theta_\star$ replaced by θ_0, and $\bar{\kappa}, \bar{\rho}$ as above and the rate function r as above.

Also for absolute constant $C \geq 3$, set

$$\mathcal{F}_{1,\theta_0} \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_{\bar{s}}} \{\delta\} \times \{\theta \in \mathbb{R}^p : \ \|\theta_\delta - \theta_0\|_2 > C\epsilon\},$$

$$\mathcal{F}_{2,\theta_0} \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_{\bar{s}}} \{\delta\} \times \{\theta \in \mathbb{R}^p : \ \|\theta_\delta - \theta_0\|_2 \leq C\epsilon, \quad \text{and} \quad \|\theta - \theta_\delta\|_2 > \epsilon_1\}.$$

From the definitions we can write $\Delta \times \mathbb{R}^p = \{\delta : \ \|\delta\|_0 > \bar{s}\} \cup \mathcal{F}_{1,\theta_0} \cup \mathcal{F}_{2,\theta_0} \cup \mathsf{B}_{\theta_0}$. Using this and $\Pi(\|\delta\|_0 > \bar{s}|X) = 0$, it follows that

$$\Pi\left(\mathsf{B}_{\theta_0}|X\right) = 1 - \Pi\left(\mathcal{F}_{1,\theta_0}|X\right) - \Pi\left(\mathcal{F}_{2,\theta_0}|X\right).$$

Hence it suffices to show that for $\varepsilon \in \{-1, 1\}$,

$$\lim_{p\to\infty} \mathbb{E}_\star\left[\mathbf{1}_{\{\mathsf{sign}(\langle V_1,\theta_\star\rangle)=\varepsilon\}}\left(\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right) + \Pi\left(\mathcal{F}_{2,\varepsilon\theta_\star}|X\right)\right)\right] = 0.$$

We have

$$\begin{aligned}\mathbb{E}_\star&\left[\mathbf{1}_{\{\mathsf{sign}(\langle V_1,\theta_\star\rangle)=\varepsilon\}}\left(\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right) + \Pi\left(\mathcal{F}_{2,\varepsilon\theta_\star}|X\right)\right)\right]\\ &\leq \mathbb{P}_\star\left(X \notin \mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s}), \mathsf{sign}(\langle V_1,\theta_\star\rangle) = \varepsilon\right)\\ &\quad + \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})}(X)\left(\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right) + \Pi\left(\mathcal{F}_{2,\varepsilon\theta_\star}|X\right)\right)\right]. \quad \text{(A.1)}\end{aligned}$$

With the same argument as in the proof of Theorem 2, we have

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})}(X)\Pi\left(\mathcal{F}_{2,\varepsilon\theta_\star}|X\right)\right] \leq 4e^{-p}.$$

We use the test constructed in Lemma 12 with $\Theta_\star = \{\theta_\star, -\theta_\star\}$, and $M = C$ to write

$$\mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})}(X)\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right)\right] \leq \mathbb{E}_\star[\phi(X)] \\ + \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})}(X)\left(1-\phi(X)\right)\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right)\right],$$

and

$$\mathbb{E}_\star[\phi(X)] \leq \frac{4(9p)^{\bar{s}} e^{-\frac{C}{8}\bar{\rho}_1(\bar{s}+s_\star)^{1/2}\epsilon}}{1-e^{-\frac{C}{8}\bar{\rho}_1(\bar{s}+s_\star)^{1/2}\epsilon}} \to 0,$$

as $p \to \infty$, by appropriately choosing the absolute constant C. The same argument leading to (A.7) applies to the second term on the right hand side of the last display, and we deduce that

$$\lim_{p\to\infty} \mathbb{E}_\star\left[\mathbf{1}_{\mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})}(X)\left(1-\phi(X)\right)\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right)\right] = 0.$$

Collecting these limiting behaviors we conclude from (A.1) that

$$\lim_{p\to\infty} \mathbb{E}_\star\left[\mathbf{1}_{\{\mathsf{sign}(\langle V_1,\theta_\star\rangle)=\varepsilon\}}\left(\Pi\left(\mathcal{F}_{1,\varepsilon\theta_\star}|X\right)+\Pi\left(\mathcal{F}_{2,\varepsilon\theta_\star}|X\right)\right)\right] \\ \leq \lim_{p\to\infty} \mathbb{P}_\star\left(X \notin \mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s}), \mathsf{sign}(\langle V_1,\theta_\star\rangle)=\varepsilon\right).$$

Hence it suffices to show that with $\bar{\kappa}$, $\bar{\rho}$, and the rate function r as above we have $\mathbb{P}_\star\left(X \notin \mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})|\mathsf{sign}(\langle V_1,\theta_\star\rangle)=\varepsilon\right) \to 0$, as $p \to \infty$.

For $\theta_0 \in \{\theta_\star, -\theta_\star\}$, and $\theta \in \mathbb{R}^p_\delta$, for any $\delta \in \Delta_{\bar{s}}$,

$$\mathcal{L}_{\theta_0}(\theta; X) = -\frac{n}{\sigma^2}(\theta-\theta_0)'\left(\frac{X'X}{n}\right)(\theta-\theta_0).$$

Lemma 1 of Ravikumar et al. [2011], and Theorem 1 of Raskutti et al. [2010] then show that the function $\theta \mapsto \mathcal{L}_{\theta_0}(\theta; X)$ satisfies the requirements of $\mathcal{E}_{1,\varepsilon\theta_\star}(\bar{s})$ with high probability, provided that the sample size n satisfies $n \geq C_0(\bar{s} + s_\star)\log(p)$, for some absolute constant C_0. Hence it remains only to show that

$$\lim_{p\to\infty} \mathbb{P}_\star\left(\|\nabla\ell(\varepsilon\theta_\star; X)\|_\infty > \frac{\bar{\rho}}{2}, \mathsf{sign}(\langle V_1, \theta_\star\rangle) = \varepsilon\right) = 0, \tag{A.2}$$

where $\bar{\rho}$ is as defined at the beginning of the proof. The largest eigenvalue of Σ is $1 + \vartheta$ with corresponding eigenvector $\theta_\star$. Hence, by the Davis-Kahan's theorem (Corollary 1 Yu et al. [2014]), on $\{\mathsf{sign}(\langle V_1, \theta_\star\rangle) = \varepsilon\}$,

$$\|V_1 - \varepsilon\theta_\star\|_2 \leq \frac{4}{\vartheta}\left\|\frac{X'X}{n} - \Sigma\right\|_2. \tag{A.3}$$

Noting that $y = \Lambda_{11}U_1 = XV_1$, we have for $\theta_0 \in \{\theta_\star, -\theta_\star\}$,

$$\begin{aligned}\nabla\ell(\theta_0; X) = \frac{1}{\sigma^2}X'(y - X\theta_0) = \frac{1}{\sigma^2}X'X(V_1 - \theta_0)& \\ = \frac{1}{\sigma^2}(X'X - n\Sigma)(V_1 - \theta_0) + \frac{n}{\sigma^2}\Sigma(V_1 - \theta_0).&\end{aligned}$$

Hence

$$\|\nabla\ell(\theta_0; X)\|_\infty \leq \frac{n}{\sigma^2}\left(\left\|\frac{X'X}{n} - \Sigma\right\|_2 + (1 + \|\theta_\star\|_\infty\vartheta)\right)\|V_1 - \theta_0\|_2.$$

This bound together with the Davis-Kahan's theorem (A.3) yields that on $\{\mathsf{sign}(\langle V_1, \theta_\star\rangle) = \varepsilon\}$, we have

$$\|\nabla\ell(\varepsilon\theta_\star; X)\|_\infty \leq \frac{4n}{\sigma^2\vartheta}\left[\left\|\frac{X'X}{n} - \Sigma\right\|_2 + (1 + \|\theta_\star\|_\infty\vartheta)\right]\left\|\frac{X'X}{n} - \Sigma\right\|_2. \tag{A.4}$$

Note then that if the covariance $X'X/n$ satisfies

$$\left\|\frac{X'X}{n} - \Sigma\right\|_2 \leq C_0 \left[\sqrt{\frac{\frac{p}{\vartheta} + \log(p)}{n}} + \frac{\frac{p}{\vartheta} + \log(p)}{n}\right](\vartheta + 1), \tag{A.5}$$

for some absolute constant C_0, then for $n \geq C_0(\frac{p}{\vartheta} + \log(p))$, we get $\|(X'X)/n - \Sigma\|_2 \leq C_0\vartheta$, and in that case (A.4) gives

$$\|\nabla \ell(\varepsilon\theta_\star; X)\|_\infty \leq \frac{4nC_0}{\sigma^2}\left\|\frac{X'X}{n} - \Sigma\right\|_2 \leq \frac{4C_0\vartheta}{\sigma^2}\sqrt{n\left(\frac{p}{\vartheta} + \log(p)\right)} = \frac{\bar{\rho}_1}{2},$$

for some absolute constant C_0. This means that the probability on the right hand side of (A.2) is upper bounded by the probability that (A.5) fails. The matrix Σ has the property that $\mathsf{Tr}(\Sigma)/\|\Sigma\|_2 = (p + \vartheta)/(1 + \vartheta) \leq 1 + (p/\vartheta)$. Using this and by deviation bound for Gaussian distribution with covariance matrix with low intrinsic dimension (see e.g. Vershynin [2018] Theorem 9.2.4), (A.5) holds that with probability at least $1 - 1/p$. Hence the results.

□

A.11 Deviation bound for quasi-likelihood ratio

Lemma 13. *Suppose that $\nabla \ell(\theta_\star; Z) = X'\epsilon$, and $\nabla^{(2)}\ell(\theta_\star; Z) = -X'WX$ for some random matrix $X \in \mathbb{R}^{n \times p}$, and a random diagonal matrix $W \in \mathbb{R}^{n \times n}$ of the form $W = H(X)$ for some measurable function $H : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times n}$. Furthermore assume that $\mathbb{E}(\epsilon|X) = 0$, and there exists $\sigma > 0$ such that $\mathbb{P}(|\langle u, \epsilon\rangle| > t|X) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$,* *for all unit-vector $u \in \mathbb{R}^n$, and all $t > 0$. For integer $s \geq 1$, we set*

$$\underline{w}(s) = \inf\left\{\frac{u'(X'WX)u}{n\|u\|_2^2},\ u \neq 0,\ \|u\|_0 \leq s\right\},$$

$$\text{and}\quad \bar{v}(s) \stackrel{\text{def}}{=} \sup\left\{\frac{u'(X'X)u}{n\|u\|_2^2},\ u \neq 0,\ \|u\|_0 \leq s\right\}.$$

We further define

$$C = \frac{\max_{i\neq j}\langle X_i, WX_j\rangle}{n}$$

and assume

$$\frac{s_0 C}{\underline{w}(s_0)} \leq 1/2$$

We can find an absolute constant C_0 such that

$$\boldsymbol{1}_{\{\underline{w}(s_\star+j)>0\}}(X)\mathbb{P}\left[\max_{\delta\in\mathcal{A}:\ \|\delta\|_0=s_\star+j} \ell^{([\delta]}(\hat{\theta}_\delta; Z) - \ell^{([\delta_\star]}(\hat{\theta}_{\delta_\star}; Z) > C_0\log(p)\frac{\sigma^2 j\bar{v}(s_\star+j)}{\underline{w}(s_\star+j)}|X\right] \leq \frac{2}{p^j}.$$

Remark A.3. The sub-Gaussian tail tail bound in the assumption implies that $\mathbb{E}_\star(\langle uV\rangle^2) \leq 4\sigma^2$, and the Orlicz norm $\|\langle uV\rangle\|_{\psi_2} \stackrel{\text{def}}{=} \inf\{t > 0 :\ \mathbb{E}(e^{\langle uV\rangle^2/t^2}) \leq 2\} \leq \sqrt{6}\sigma$. See for instance Vershynin [2018] Section 2.5.1 for details.

Proof. Fix a model δ that contains the true model $\delta_\star$, and $\|\delta\|_0 = s_0 = s_\star + j$. We will abuse notations and identify $\theta_\star$ with the element $[\theta_\star]_{\delta_\star} \in \mathbb{R}^{s_\star}$, as well as with $[\theta_\star]_\delta \in \mathbb{R}^{s_\star+j}$. By the optimality condition on $\hat{\theta}_\delta$, and Taylor expansion, we have

$$0 = \nabla\ell^{[\delta]}(\hat{\theta}_\delta; Z) = \nabla\ell^{[\delta]}(\theta_\star; Z) + \nabla^{(2)}\ell^{[\delta]}(\theta_\star; Z)(\hat{\theta}_\delta - \theta_\star) + T_1,$$

where using the quantity a_2, we have

$$\|T_1\|_\infty \leq C_0\mathsf{a}_2 s_0\epsilon^2,$$

for some absolute constant C_0. This implies that

$$(\hat{\theta}_\delta-\theta_\star)' \left[\nabla^{(2)}\ell^{[\delta]}(\theta_\star;Z)\right] (\hat{\theta}_\delta-\theta_\star) = \nabla\ell^{[\delta]}(\theta_\star;Z)' \left[\nabla^{(2)}\ell^{[\delta]}(\theta_\star;Z)\right]^{-1} \nabla\ell^{[\delta]}(\theta_\star;Z)+T_1',$$

where

$$\begin{aligned}|T_1'| = T_1 \left[-\nabla^{(2)}\ell^{[\delta]}(\theta_\star;Z)\right]^{-1} T_1 + 2\left|T_1 \left[\nabla^{(2)}\ell^{[\delta]}(\theta_\star;Z)\right]^{-1} \nabla\ell^{[\delta]}(\theta_\star;Z)\right| \\ \leq \frac{C_0}{n\underline{w}(s_0)}\mathsf{a}_2 s_0^2\epsilon^2(\mathsf{a}_2 s_0\epsilon^2+\bar{\rho}).\end{aligned}$$

On the other hand we also have

$$\ell^{([\delta]}(\theta_\star;Z) = \ell^{([\delta]}(\hat{\theta}_\delta;Z) + \frac{1}{2}(\hat{\theta}_\delta-\theta_\star)' \left[\nabla^{(2)}\ell^{[\delta]}(\theta_\star;Z)\right] (\hat{\theta}_\delta-\theta_\star) + T_2,$$

where

$$|T_2| \leq C_0\mathsf{a}_2 s_0^{3/2}\epsilon^3.$$

We conclude that

$$\ell^{[\delta]}(\hat{\theta}_\delta;Z) - \ell^{[\delta]}(\theta_\star;Z) = \frac{1}{2}\nabla\ell^{[\delta]}(\theta_\star;Z)' \left[-\nabla^{(2)}\ell^{[\delta]}(\theta_\star;Z)\right]^{-1} \nabla\ell^{[\delta]}(\theta_\star;Z) + T_3,$$

where

$$|T_3| \leq C_0\left(\mathsf{a}_2 s_0^{3/2}\epsilon^3 + \frac{\mathsf{a}_2 s_0^2\epsilon^2(\mathsf{a}_2 s_0\epsilon^2+\bar{\rho})}{n\underline{w}(s_0)}\right).$$

The same development hold for $\delta = \delta_\star$. Put together both identities yield

$$\begin{aligned}
2&\left[\ell^{[\delta]}(\hat{\theta}_\delta; Z) - \ell^{[\delta_\star]}(\hat{\theta}_{\delta_\star}; Z)\right] \\
&= \nabla \ell^{[\delta]}(\theta_\star; Z)' \left[-\nabla^{(2)}\ell^{[\delta]}(\theta_\star; Z)\right]^{-1} \nabla \ell^{[\delta]}(\theta_\star; Z) \\
&\quad - \nabla \ell^{[\delta_\star]}(\theta_\star; Z)' \left[-\nabla^{(2)}\ell^{[\delta_\star]}(\theta_\star; Z)\right]^{-1} \nabla \ell^{[\delta_\star]}(\theta_\star; Z) + T_4,
\end{aligned}$$

where

$$|T_4| \leq C_0 \left(\mathsf{a}_2 s_0^{3/2}\epsilon^3 + \frac{\mathsf{a}_2 s_0^2 \epsilon^2 (\mathsf{a}_2 s_0 \epsilon^2 + \bar{\rho})}{n\underline{w}(s_0)}\right).$$

. When $\bar{\rho} \sim \sqrt{n\log(p)}T_4 \to 0$ for $\frac{log(p)^3}{n} \to 0$ We partition $\nabla^{(2)}\ell^{[\delta]}(\theta_\star; Z)$ as $\begin{bmatrix} A & B \\ B' & C \end{bmatrix}$, where $A = \nabla^{(2)}\ell^{[\delta_\star]}(\theta_\star; Z)$. We make use of the following well-known block-matrix inversion (see e.g. Horn and Johnson [2012] Section 0.7.3). If $M = \begin{bmatrix} A & B \\ B' & C \end{bmatrix}$ is invertible and A is invertible, then with $S = C - B'A^{-1}B$,

$$\begin{pmatrix} u \\ v \end{pmatrix}' M^{-1} \begin{pmatrix} u \\ v \end{pmatrix} - u'A^{-1}u = \|S^{-1/2}\left(B'A^{-1}u - v\right)\|_2^2.$$

Hence

$$2\left[\ell^{([\delta]}(\hat{\theta}_\delta; Z) - \ell^{([\delta_\star]}(\hat{\theta}_{\delta_\star}; Z)\right] = V'S^{-1}V + T_4,$$

where

$$S \overset{\text{def}}{=} [-\nabla^{(2)}\ell(\theta_\star; Z)]_{\delta-\delta_\star,\delta-\delta_\star} - [\nabla^{(2)}\ell(\theta_\star; Z)]_{\delta-\delta_\star,\delta_\star}\left(-[\nabla^{(2)}\ell(\theta_\star; Z)]_{\delta_\star,\delta_\star}\right)^{-1}[\nabla^{(2)}\ell(\theta_\star; Z)]_{\delta_\star,\delta-\delta_\star},$$

and

$$V = [\nabla^{(2)}\ell(\theta_\star; Z)]_{\delta-\delta_\star,\delta_\star} \left(-[\nabla^{(2)}\ell(\theta_\star; Z)]_{\delta_\star,\delta_\star}\right)^{-1} [\nabla\ell(\theta_\star; Z)]_{\delta_\star} - [\nabla\ell(\theta_\star; Z)]_{\delta-\delta_\star}.$$

Since $[\nabla\ell(\theta_\star; Z)]_\delta = X'_\delta\epsilon$, the tail-bound assumption on ϵ implies that the conditional Orlicz norm of V is upper-bounded by

$$\sqrt{6}\sigma\bar{v}(s_0)^{1/2}\sqrt{n}\left(1 + \frac{s_\star C}{\underline{w}(s_\star)}\right) \leq \sigma\sqrt{24n}\bar{v}(s_0)^{1/2}.$$

Furthermore the smallest eigenvalue of S is bounded from below by

$$n\underline{w}(s_0) - \frac{ns_\star jC^2}{\underline{w}(s_\star)} \geq \frac{n\underline{w}(s_0)}{2}.$$

Under the assumption that $\frac{\bar{v}(s_0)}{\underline{w}(s_0)} \sim O(1)$, we then apply HW inequality (B.1) to conclude that for some absolute constant C_0

$$\mathbb{P}\left[\ell^{[\delta]}(\hat{\theta}_\delta; Z) - \ell^{[\delta_\star]}(\hat{\theta}_{\delta_\star}; Z) > C_0\log(p)\frac{\sigma^2 j\bar{v}(s_0)}{\underline{w}(s_0)}\right] \leq 2e^{-2j\log(p)},$$

and by union bound

$$\mathbb{P}\left[\max_{\delta\in\mathcal{A}:\ \|\delta\|_0 = s_\star + j} \ell^{([\delta]}(\hat{\theta}_\delta; Z) - \ell^{([\delta_\star]}(\hat{\theta}_{\delta_\star}; Z) > C_0\log(p)\frac{\sigma^2 j\bar{v}(s_\star + j)}{\underline{w}(s_\star + j)}\right] \leq \frac{2}{p^j}.$$

□

□

APPENDIX B

Some technical results

B.1 KL-divergence of Gaussian distributions

We make use of the following expression of the KL-divergence between two Gaussian distributions.

Lemma 14. *For $i = 1, 2$ let π_i denote the probability distribution of the Gaussian distribution $\boldsymbol{N}(\mu_i, \Sigma_i)$. We have*

$$\mathsf{KL}\left(\pi_1|\pi_2\right) = \frac{1}{2}(\mu_2 - \mu_1)'\Sigma_2^{-1}(\mu_2 - \mu_1) + \frac{1}{2}\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) + \frac{1}{2}\mathit{Tr}(\Sigma_2^{-1}\Sigma_1) - \frac{p}{2}.$$

B.2 Gaussian deviation bounds

The following lemma follows readily from standard Gaussian deviation bounds. We omit the details.

Lemma 15. *Suppose that a $\mathbb{R}^p$-valued random variable X has density $f(x) \propto e^{-\ell(x)-\rho\|x\|_2^2/2}$, for a twice differentiable function ℓ such that $mI_p \preceq \nabla^{(2)}\ell \preceq MI_p$,* *for some constants* $0 < m \leq M$*, and* $\rho > 0$*. Let* μ *denote the mode of* ℓ*. For all* $t \geq 4\max\left(\frac{\rho}{\rho+m}\|\mu\|_2, \sqrt{\frac{p}{\rho+m}}\right)$ *we have*

$$\mathbb{P}\left(\|X-\mu\|_2 > t\right) \leq \left(\frac{M+\rho}{m+\rho}\right)^{\frac{p}{2}} e^{-\frac{t^2(m+\rho)}{16}},$$

$$\textit{and} \quad \mathbb{E}\left(\|X-\mu\|_2^2 \boldsymbol{1}_{\{\|X-\mu\|_2>t\}}\right) \leq t^2\left(\frac{M+\rho}{m+\rho}\right)^{\frac{p}{2}} e^{-\frac{t^2(m+\rho)}{32}}.$$

Proof. By Taylor expansion of ℓ around μ:

$$-\frac{M}{2}\|x-\mu\|_2^2 - \frac{\rho}{2}\|x\|_2^2 \leq \ell(\mu) - \ell(x) - \frac{\rho}{2}\|x\|_2^2 \leq -\frac{m}{2}\|x-\mu\|_2^2 - \frac{\rho}{2}\|x\|_2^2, \quad x \in \mathbb{R}^p.$$

This implies that

$$\int_{\mathbb{R}^p} e^{\ell(\mu)-\ell(x)-\frac{\rho}{2}\|x\|_2^2}\mathrm{d}x \geq e^{-\frac{M\rho}{2(M+\rho)}\|\mu\|_2^2}\left(\frac{2\pi}{\rho+M}\right)^{p/2}.$$

Therefore, for any $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\|X-\mu\|_2 > t\right) &\leq e^{\frac{M\rho}{2(M+\rho)}\|\mu\|_2^2}\left(\frac{\rho+M}{\rho+m}\right)^{p/2}\mathbb{P}\left(\left\|\frac{Z}{\sqrt{\rho+m}} - \frac{\rho\mu}{\rho+m}\right\|_2 > t\right), \\ &\leq e^{\frac{\rho}{2}\|\mu\|_2^2}\left(\frac{\rho+M}{\rho+m}\right)^{p/2} e^{-\frac{1}{2}\left(t\sqrt{m+\rho}-\frac{\rho\|\mu\|_2}{\sqrt{m+\rho}}-\sqrt{p}\right)^2}.\end{aligned}$$

where $Z \sim \mathbf{N}_p(0, I_p)$. For $t \geq 4\max(\rho\|\mu\|_2/(\rho+m), \sqrt{\frac{p}{m+\rho}})$, this yields

$$\mathbb{P}\left(\|X-\mu\|_2 > t\right) \leq \left(\frac{\rho+M}{\rho+m}\right)^{p/2} e^{-\frac{t^2(m+\rho)}{16}}.$$

By Holder's inequality

$$\mathbb{E}\left(\|X-\mu\|_2^2\mathbf{1}_{\{\|X-\mu\|_2>t\}}\right) \leq \mathbb{E}^{1/2}(\|X-\mu\|_2^4)\mathbb{P}^{1/2}\left(\|X-\mu\|_2 > t\right).$$

With the same calculations as above,

$$\begin{aligned}
\mathbb{E}(\|X-\mu\|_2^4) &\leq e^{\frac{\rho}{2}\|\mu\|_2^2}\left(\frac{\rho+M}{\rho+m}\right)^{p/2}\mathbb{E}\left(\left\|\frac{Z}{\sqrt{\rho+m}}-\frac{\rho\mu}{\rho+m}\right\|_2^4\right), \\
&\leq 8e^{\frac{\rho}{2}\|\mu\|_2^2}\left(\frac{\rho+M}{\rho+m}\right)^{p/2}\left(\frac{3p^2}{(m+\rho)^2}+\frac{\rho^4\|\mu\|_2^4}{(m+\rho)^4}\right) \\
&\leq e^{\frac{\rho}{2}\|\mu\|_2^2}\left(\frac{\rho+M}{\rho+m}\right)^{p/2}\frac{t^4}{8},
\end{aligned}$$

using the assumption $t \geq 4\max(\frac{\rho}{\rho+m}\|\mu\|_2, \sqrt{\frac{p}{m+\rho}})$, which implies the second inequality. □

B.3 Strong convexity of KL-divergence

The next results establishes the strong convexity of the KL divergence. The proof is due to I. Pinelis (Pinelis [2018]). We reproduce it here for completeness.

Lemma 16. *Let P_0, P_1 be two probability measures that are absolutely continuous with respect to a probability measure Q, on some measure space $\mathcal{X}$. For any $t \in (0,1)$, we have*

$$t\mathsf{KL}\left(P_1|Q\right)+(1-t)\mathsf{KL}\left(P_0|Q\right) \geq \mathsf{KL}\left(tP_1+(1-t)P_0|Q\right)+\frac{t(1-t)}{2}\|P_1-P_0\|_{\text{tv}}^2.$$

Proof. For $j=0,1$, set $f_j = \mathrm{d}P_j/\mathrm{d}Q$. For $t\in[0,1]$, set $f_t = tf_1+(1-t)f_0$, and $P_t(\mathrm{d}u) = f_t(u)Q(\mathrm{d}u)$. Set $h(x) = x\log(x)$, $x \geq 0$. By Taylor expansion with integral remainder, for $j \in \{0,1\}$, $t \in [0,1]$, and $x \in \mathcal{X}$, we have

$$\begin{aligned} h(f_j(u)) &= h(f_t(u)) + (f_j(u) - f_t(u))\, h'(f_t(u)) \\ &\qquad + (f_j(u) - f_t(u))^2 \int_0^1 h''\left((1-\alpha)f_t(u) + \alpha f_j(u)\right)(1-\alpha)\mathrm{d}\alpha. \end{aligned}$$

$h'(x) = \log(x) - 1$, and $h''(x) = 1/x$, so that

$$\begin{aligned} th(f_1(u)) + (1-t)h(f_0(u)) - h(f_t(u) &= t(1-t)\left(f_1(u) - f_0(u)\right)^2 \\ \times \int_0^1 \left[\frac{t}{(1-\alpha)f_t(u) + \alpha f_0(u)} + \frac{1-t}{(1-\alpha)f_t(u) + \alpha f_1(u)}\right](1-\alpha)\mathrm{d}\alpha. \quad \text{(B.1)} \end{aligned}$$

We can write $(1-\alpha)f_t(u) + \alpha f_0(u) = f_{s_0(\alpha,t)}(u)$, where $s_0(\alpha,t) = (1-\alpha)t$. Similarly, $(1-\alpha)f_t(u) + \alpha f_1(u) = f_{s_1(\alpha,t)}$, where $s_1(\alpha,t) = \alpha + t(1-\alpha)$. Using these expressions, and integrating both sides of (B.1) gives

$$\begin{aligned} & t\mathsf{KL}\left(P_1|Q\right) + (1-t)\mathsf{KL}\left(P_0|Q\right) - \mathsf{KL}\left(P_t|Q\right) \\ &= t(1-t)\int_0^1 (1-\alpha)\left[t\int \frac{(f_1(u) - f_0(u))^2}{f_{s_0(\alpha,t)}(u)}Q(\mathrm{d}u) + (1-t)\int \frac{(f_1(u) - f_0(u))^2}{f_{s_1(\alpha,t)}(u)}Q(\mathrm{d}u)\right]\mathrm{d}\alpha. \end{aligned}$$

For any $s \in (0,1)$,

$$\begin{aligned} \int \frac{(f_1(u) - f_0(u))^2}{f_s(u)}Q(\mathrm{d}u) &= \frac{1}{(1-s)^2}\int \frac{(f_1(u) - f_s(u))^2}{f_s(u)}Q(\mathrm{d}u) \\ = \frac{1}{(1-s)^2}\int \left(\frac{f_1(u)}{f_s(u)} - 1\right)^2 f_s(u)Q(\mathrm{d}u) &\geq \frac{1}{(1-s)^2}\left[\int \left|\frac{f_1(u)}{f_s(u)} - 1\right| Q_s(\mathrm{d}u)\right]^2 \\ &= \frac{1}{(1-s)^2}\|P_s - P_1\|_{\mathrm{tv}}^2 = \|P_1 - P_0\|_{\mathrm{tv}}^2. \end{aligned}$$

We conclude that

$$t\mathsf{KL}\left(P_1|Q\right) + (1-t)\mathsf{KL}\left(P_0|Q\right) - \mathsf{KL}\left(P_t|Q\right)$$
$$\geq t(1-t)\|P_1 - P_0\|_{\mathrm{tv}}^2 \int_0^1 \alpha(1-\alpha)\mathrm{d}\alpha = \frac{t(1-t)}{2}\|P_1 - P_0\|_{\mathrm{tv}}^2,$$

as claimed. □

B.4 Hanson-Wright inequality

The following deviation bound is known as the Hanson-Wright inequality. This version is taken from (Vershynin [2018]).

Lemma 17. *Let $X = (X_1, \ldots, X_n)$ be a random vector with independent mean zero components. Suppose that there exists $\sigma > 0$ such that for all unit-vector $u \in \mathbb{R}^n$, and all $t \geq 0$, $\mathbb{P}(|\langle u, X\rangle| > t) \leq 2e^{-t^2/(2\sigma^2)}$. Then for all $t \geq 6$, it holds*

$$\mathbb{P}\left[X'AX > (4+t)\sigma^2 r\lambda_{\mathit{max}}(A)\right] \leq e^{-\frac{ctr}{6}}, \tag{B.1}$$

for some absolute constant c where $r = rank(A)$. In the particular case where $X \sim \boldsymbol{N}_n(0, I_n)$, $\sigma = 1$, and we can take $c = 3$.

B.5 Relation between determinants of sub-matrices

We will also need the following lemma on determinants of sub-matrices.

Lemma 18. *If symmetric positive definite matrices A, M and $D \in \mathbb{R}^{q\times q}$ are such* *that* $M = \begin{pmatrix} A & B \\ B' & D \end{pmatrix}$, *then*

$$\det(A)\lambda_{\mathsf{min}}(M)^q \leq \det(M) \leq \det(A)\lambda_{\mathsf{max}}(M)^q.$$

Proof. This follows from Cauchy's interlacing property for eigenvalues. See for instance Horn and Johnson [2012] Theorem 4.3.17. □

BIBLIOGRAPHY

BIBLIOGRAPHY

F. Abramovich and V. Grinshtein. High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079, 2019.

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv e-prints*, art. arXiv:1706.09293, Jun 2017.

Y. A. Atchade. On the contraction properties of some high-dimensional quasi-posterior distributions. *Ann. Statist.*, 45(5):2248–2273, 10 2017.

Y. F. Atchadé. An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodol Comput Appl Probab*, 8:235–254, 2006.

Y. F. Atchadé. Quasi-bayesian estimation of large gaussian graphical models. *Journal of Multivariate Analysis*, 173:656 – 671, 2019.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

S. Banerjee and S. Ghosal. Posterior convergence rates for estimating large precision matrices using graphical models. *ArXiv e-prints*, Feb. 2013.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, 36:192–236, 1974. ISSN 0035-9246. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

A. Bhattacharya, D. Pati, N. Pillai, and D. Dunson. Dirichlet-laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110 (512):1479–1490, 2015. doi: 10.1080/01621459.2014.960967. URL https://doi.org/10.1080/01621459.2014.960967. PMID: 27019543.

A. Bhattacharya, A. Chakraborty, and B. Mallick. Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985 – 991, 2016.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112 (518):859–877, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Springer Series in Statistics. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5.

B. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Roy. Stat. Soc. Series B (Methodological)*, 57(3):473–484, 1995.

I. Castillo and A. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101, 08 2012.

I. Castillo, J. Schmidt-Hieber, and A. van der Vaart. Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018, 10 2015. doi: 10.1214/15-AOS1334.

H. Cattell and A. Mead. *Personality Measurement and Testing*, volume 2. SAGE, 2008. URL http://dx.doi.org/10.4135/9781849200479.

V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *J. Econometrics*, 115(2):293–346, 2003.

V. Chernozhukov, G. W. Imbens, and W. K. Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4 – 14, 2007.

A. Dobra, A. Lenkoski, and A. Rodriguez. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.*, 106(496):1418–1433, 2011.

M. Ekeberg, C. Lövkvist, Y. Lan, and E. A. M. Weigt. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. *Physical Review*, 87, 2013.

S. Epskamp, J. Kruis, and M. Marsman. Estimating psychopathological networks: Be careful what you wish for. *PLoS ONE*, 12, 2017. URL https://doi.org/10.1371/journal.pone.0179891.

S. Epskamp, D. Borsboom, and E. Fried. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods*, 50, 2018. URL https://doi.org/10.3758/s13428-017-0862-1.

C. Gao and H. H. Zhou. Rate-optimal posterior contraction for sparse pca. *Ann. Statist.*, 43(2):785–818, 04 2015. doi: 10.1214/14-AOS1268.

E. I. George and R. E. McCulloch. Approaches to bayesian variable selection. *Statist. Sinica*, 7(1):339–373, 1997.

H.-O. Georgii. Gibbs measures and phase transitions. *de Gruyter Studies in Mathematics*, 9, 1988.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

A. Greenfield, C. Hafemeister, and Bonneau.R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29:1060–1067, 2013.

J. Guo, J. Cheng, E. Levina, G. Michailidis, and J. Zhu. Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *Ann. Appl. Stat.*, 9(2):821–848, 06 2015. doi: 10.1214/13-AOAS700. URL https://doi.org/10.1214/13-AOAS700.

X. Guyon. *Random fields on a network*. Probability and its Applications (New York). Springer-Verlag, New York, 1995. ISBN 0-387-94428-1. Modeling, statistics, and applications, Translated from the 1992 French original by Carenne Ludeña.

H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.*, 10:883–906, 2009.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 2nd edition, 2012. ISBN 0521548233, 9780521548236.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71 – 120, 1993.

E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift Für Physik A Hadrons and Nuclei*, 31:253–258, 1925.

C. Ji and S. Schmidler. Adaptive markov chain monte carlo for bayesian variable selection. *Journal of Computational and Graphical Statistics*, 22(3):708–728, 2013.

W. Jiang and M. A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.*, 36(5):2207–2231, 2008.

I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649, 08 2004.

I. Jolliffe. *Principal Component Analysis.* Springer Verlag, 1986.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

K. Kato. Quasi-Bayesian analysis of nonparametric instrumental variables models. *Ann. Statist.*, 41(5):2359–2390, 2013.

Z. S. Khondker, H. Zhu, H. Chu, W. Lin, and J. G. Ibrahim. The Bayesian covariance lasso. *Stat. Interface*, 6(2):243–259, 2013.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877, 2006.

J. Lei and V. Q. Vu. Sparsistency and agnostic inference in sparse pca. *Ann. Statist.*, 43(1):299–322, 02 2015. doi: 10.1214/14-AOS1273.

A. Lenkoski and A. Dobra. Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graph. Statist.*, 20(1):140–157, 2011. Supplementary material available online.

C. Li and W. Jiang. Model Selection for Likelihood-free Bayesian Methods Based on Moment Conditions: Theory and Numerical Examples. *ArXiv e-prints*, May 2014.

F. Li and N. R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.*, 105:1202–1214, 2010.

Y. Liao and W. Jiang. Posterior consistency of nonparametric conditional moment restricted models. *Ann. Statist.*, 39(6):3003–3031, 12 2011.

J. Liu. Monte carlo strategies in scientific computing. *Springer*, 2001.

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statist. Sci.*, 30(4):443–467, 11 2015.

R. Martin, R. Mess, and S. G. Walker. Empirical bayes posterior concentration in sparse high dimensional linear models. *Bernoulli*, 23:1822–1847, 2017.

N. Meinshausen and P. Buhlmann. High-dimensional graphs with the lasso. *Annals of Stat.*, 34:1436–1462, 2006.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *JASA*, 83(404):1023–1032, 1988.

M. Moores, G. Nicholls, A. Pettitt, and K. Mengersen. Scalable bayesian inference for the inverse temperature of a hidden potts model. *Bayesian Analysis*, 15:1–27, 2020.

E. Moreno, J. Girón, and G. Casella. Posterior model consistency in variable selection as the model dimension grows. *Statistical Science*, 30:228–241, 2015.

K. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012.

I. Murray, Z. Ghahramani, and D. MacKay. MCMC for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence UAI06*, pages 359–366, 2006.

N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42(2):789–817, 04 2014.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012. doi: 10.1214/12-STS400.

D. J. Nott and R. Kohn. Adaptive sampling for bayesian variable selection. *Biometrica*, 92(4):747–763, 2005.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1, 2013.

D. Pati, A. Bhattacharya, N. S. Pillai, and D. Dunson. Posterior contraction in sparse bayesian factor models for massive covariance matrices. *Ann. Statist.*, 42 (3):1102–1130, 06 2014. doi: 10.1214/14-AOS1215. URL https://doi.org/10.1214/14-AOS1215.

B. Peng, D. Zhu, B. P. Ander, X. Zhang, F. Xue, F. Sharp, and X. Yang. An integrative framework for bayesian variable selection with informative priors for identifying genes and pathways. *PLoS ONE*, 8(7), 2013.

C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

I. Pinelis. Is KL-divergence $D(P||Q)$ strongly convex over P in infinite dimension?, 2018. URL https://mathoverflow.net/q/307251.

N. Polson and J. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, 9:501–539, 2010.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *ArXiv e-prints*, May 2012.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3): 1287–1319, 2010.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

K. Ray and B. Szabo. Variational bayes for high-dimensional linear regression with sparse priors, 2019.

C. P. Robert and G. Casella. Monte carlo statistical methods. *Springer Texts in Statistics, Springer-Verlag, New York.*, 2004a.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004b. ISBN 0-387-21239-6.

V. Rockova and E. I. George. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

R. Rosu, J. Giovannelli, A. Giremus, and C. Vacar. Potts model parameter estimation in bayesian segmentation of piecewise constant images. pages 4080–4084, 2015.

S. Roy, Y. Atchadé, and G. Michailidis. Change point estimation in high dimensional markov random field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1187–1206, 2017.

H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015 – 1034, 2008.

K. Sjöstrand, L. Clemmensen, R. Larsen, G. Einarsson, and B. ErsbÃc ll. Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software, Articles*, 84(10):1–37, 2018. doi: 10.18637/jss.v084.i10.

M. Studham, A. Tjarnberg, T. Nordling, S. Nelander, and E. Sonnhammer. Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, 30:i130–i138, 2014.

P. Sur and E. Candés. A modern maximum-likelihood theory for high-dimensional logistic regression. *PNAS*, 116(29):14516–14525, 2019.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Y. Wang and D. M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 0(0):1–15, 2018.

F. Xie, Y. Xu, C. E. Priebe, and J. Cape. Bayesian Estimation of Sparse Spiked Covariance Matrices in High Dimensions. *arXiv e-prints*, art. arXiv:1808.07433, Aug 2018.

W. Yang and X. He. Bayesian empirical likelihood for quantile regression. *Ann. Statist.*, 40(2):1102–1131, 2012.

Y. Yang, M. J. Wainwright, and M. I. Jordan. On the computational complexity of high-dimensional bayesian variable selection. *Ann. Statist.*, 44(6):2497–2532, 12 2016.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 04 2014.

X. Zhou and S. Schmidler. Bayesian parameter estimation in ising and potts models: A comparative study with applications to protein modeling. *Technical Report*, 2009.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.