January 2016

# Efficient Sparse Bayesian Learning using Spike-and-Slab Priors

Syed Abbas Zilqurnain Naqvi Zilqurnain Naqvi
*Purdue University*

*Graduate School Form*
*30 Updated 12/26/2015*

**PURDUE UNIVERSITY**
**GRADUATE SCHOOL**
**Thesis/Dissertation Acceptance**

This is to certify that the thesis/dissertation prepared

By Syed Abbas Zilqurnain Naqvi

Entitled

EFFICIENT SPARSE BAYESIAN LEARNING USING SPIKE-AND-SLAB PRIORS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

| | |
|---|---|
| Yuan Qi<br>Chair | Ninghui Li |
| Charles A. Bouman | |
| Jennifer Neville | |
| David F. Gleich | |

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Yuan Qi

Approved by: William J. Gorman 4/21/2016

Head of the Departmental Graduate Program Date

# EFFICIENT SPARSE BAYESIAN LEARNING USING SPIKE-AND-SLAB PRIORS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Syed Abbas Zilqurnain, Naqvi

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2016

Purdue University

West Lafayette, Indiana

To my parents and wife

# ACKNOWLEDGMENTS

A number of people have contributed towards my academic development, without whom I would not have been able to complete my academic journey at Purdue. First, I would like to acknowledge the invaluable moral support given to me by my parents and my wife. If not for their moral backing, I would not have been able to survive the test and pressures of doing PhD.

I would especially like to thank my PhD advisor Professor Alan Qi for his invaluable academic support and guidance throughout my PhD. Prof Qi contributed significantly towards developing my scientific writing and speaking skills, and his mentorship made tremendous influence on me. Professor Qi also played a major role in developing good work ethics in me, and I owe to him a lot for this.

Shandian Zhe also played a significant role in my academic development, and I want to express my sincere appreciation towards him. I collaborated with him on two projects. He is a trustworthy and capable person. I wish to continue collaborating with him in the future.

I would also like to thank my committee members Prof Charles Bouman, Prof David Gliech, and Professor Jennifer Neville. I would especially like to show my sincere gratitude to Professor Neville, who was always there to guide me regarding my final defense related issues.

Some part of introduction, the whole of chapter 2, and some portion of summary section of this dissertation is a pre-copy edited, author-produced PDF of an article accepted for publication in Bioinformatics: [Oxford journal] following peer review. The version of record [Zhe S, Naqvi SA, Yang Y, Qi Y. Joint network and node selection for pathway-based genomic data analysis. Bioinformatics. 2013 Jun 8:btt335]. is available online at: [http://bioinformatics.oxfordjournals.org/content/29/16/1987.abstract].

TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

Naqvi, Syed Abbas Zilqurnain PhD, Purdue University, May 2016. Efficient Sparse Bayesian Learning using Spike-and-Slab Priors. Major Professor: Yuan Qi.

In the context of statistical machine learning, sparse learning is a procedure that seeks a reconciliation between two competing aspects of a statistical model: good predictive power and interpretability. In a Bayesian setting, sparse learning methods invoke sparsity inducing priors to explicitly encode this tradeoff in a principled manner. Recently, spike-and-slab priors have been very popular in the sparse machine learning community. This popularity stems from the selective shrinkage property of the priors: irrelevant variables are shrunk aggressively, but relevant variables are regularized mildly. However, classical formulation of the spike-and-slab priors does not explicitly incorporate information about the correlation structure between the variables which is available in various domains, and could be useful for revealing the sparsity structure. In this dissertation we focus on supervised parametric linear models, and propose a generalized formulation of the spike-and-slab priors that tries to achieve optimal model complexity by exploiting this domain based correlation structure information, and hence seeks to improve the predictive power and interpretability of the results. Bayesian learning through spike-and-slab priors, though attractive, is not free of challenges. One huge bottleneck associated with current Bayesian inference methodologies is the high computational cost at high dimensions. In this dissertation we also propose scalable Bayesian inference strategies for classical spike-and-slab models. First, we present a new sparse Bayesian approach, called Network and Node Selection (NaNOS), for joint group and feature selection. NaNOS extends the classical spike-and-slab prior for group selection by presenting a generalized formulation of the prior that incorporates correlation structure information provided by the domain for each group, and allows our model to induce structured sparsity, guided by domain knowledge, within the

selected groups. NaNOS also provides a principled framework for jointly selecting relevant groups as well as relevant features within the selected groups. Simulation and real data results demonstrate improved predictive performance and selection accuracy of our method over alternative methods. Second, we propose a scalable approximate Bayesian inference algorithm based on Laplace's method for the classical spike-and-slab models. Our method can be seen as a hybrid of Bayesian and frequentist treatments taking benefits from both worlds. From a frequentist perspective, our approach is computationally efficient, and possesses asymptotic consistency properties; and from a Bayesian point of view, our method performs posterior inference better than or comparable to existing approximate inference techniques. Experimental results show improved performance of our approach compared to alternative approximate inference methods, but with computational efficiency comparable to frequentist $l_1$ approaches.

# 1 INTRODUCTION

## 1.1 Principle of sparsity

One of the main objectives of scientific research is to provide appropriate explanations for observed phenomena and processes. In its most abstract form, the principle of sparsity or parsimony states that the best explanation is the simplest one. The principle traces back its origin to the theory formulated by a philosopher and theologian William of Ockham in the 14th century [1]. According to this theory, nature favors simplicity over complexity, and the apparent complexity of any phenomena can be reduced to very simple rules. In the context of statistical modelling, Wrinch and Jeffreys [2] defined simplicity in terms of number of parameters appearing in the model. Following the framework put forward by Wrinch and Jeffreys [2], subsequent statistical research has focused on building statistical models with good generalization performance in terms of prediction. Out of multiple options, simpler models are preferred over complex ones to explain observed phenomena. This process of selecting an appropriate model is known as model selection, and the number of parameters is used as a guide to perform this task [1, 3, 4, 5, 6, 7, 8, 9, 10].

The principle of sparsity also finds its applications in signal processing. Here, signal is the observed data, and modelling of this data allows the signal to be processed in different ways: restoration, compression, and also for handling related inverse problems [1]. Signal processing research focuses on sparse linear combination of basic elements called dictionary elements, leading to a very simple model [11, 12, 13, 14, 15, 16, 17]. The parsimony principle has also been utilized in some other fields. For example, Markowitz [18] exploited it for portfolio selection in finance, in geophysics [19, 20], and pioneering work by Olshausen and Field [21, 22] in neuroscience. Olshausen and Field [21, 22] work was subsequently exploited in numerous applications for image and audio processing [23, 24, 25].

## 1.2 Sparse learning

In its most general definition, sparse learning refers to a collection of procedures that try to find the simplest and best explanation for an observed phenomena. In the context of statistical modelling, sparse learning seeks to encode the tradeoff between good predictive power and sparsity of the result, the latter enhances the interpretability [26]. Broadly speaking, there are two approaches for sparse learning in statistical modeling: frequentist and Bayesian approaches. We will briefly discuss the frequentist and Bayesian paradigms in the next few sections.

### 1.2.1 Frequentist sparse learning approaches

In a frequentist setting, problems related to sparse learning are treated as constraint optimization problems. The constraints provide a principled way of incorporating the sparsity conditions as part of the optimization process. The sparsity controlling parameters or the regularization parameters serve as a knob to adjust the balance between prediction and sparsity. Hence, the optimization process directly generates more interpretable results without requiring any after the fact analysis [26]. This is in contrast to thresholded PCA [27] and related thresholding techniques that do not incorporate sparsity requirements as part of the learning algorithm [26]. Most of the frequentist statistical procedures can be formulated as optimizing an objective function $\mathcal{F}(\mathbf{w})$ which can be written as:

$$\mathcal{F}(\mathbf{w}) = L(\mathbf{w}) + \lambda R(\mathbf{w})$$

where $\mathbf{w}$ is the model parameter vector, $L(\mathbf{w})$ is the loss function, $R(\mathbf{w})$ is the regularization penalty and $\lambda$ is the regularization coefficient. $R(\mathbf{w})$ enforces the sparsity conditions on the model and $\lambda$ adjusts the balance between prediction and sparsification. When $R(\mathbf{w}) = 0$, the problem reduces to the ordinary least square problem. One drawback of ordinary least square is that it does not enforce any sparsity, and hence generates dense solutions. In order to improve ordinary least squares, Mallows [3, 4] developed the $C_p$

statistics. Akaike [5] and Schwarz [9] later proposed generalizations with AIC and BIC respectively. For all these methods, $R(\mathbf{w}) = ||\mathbf{w}||_0 = \sum_{i=1}^{p} \mathbf{I}(w_i \neq 0)$, and hence these methods seek a best subset of parameters that optimizes the objective function. A similar problem arises in signal processing where a discrete signal $x$ in $\mathcal{R}^p$ space has to be approximated by a sparse linear combination of wavelet elements. The problem can be formulated as finding a sparse vector $\alpha$ with $k$ non zero elements, that minimizes [1]:

$$\frac{1}{2}||x - \mathbf{D}\alpha||_2^2 \; s.t. \; ||\alpha||_0 \leq k$$

where $\mathbf{D} = [\mathrm{d}_1, \cdots, \mathrm{d}_p]$ is an orthogonal wavelet basis set satisfying $\mathbf{D}^\top \mathbf{D} = \mathbf{I}$ with $\mathbf{I}$ being an identity matrix. The above optimization problem is a special case of a more general formulation in statistics.

It can be shown that the best subset selection procedure is optimal in terms of prediction error, and in terms of achieving the trade off between sparsity and prediction, but there are certain drawbacks: since the procedure is inherently discrete, the optimization procedure is combinatorial in nature, which leads to an exponential increase in the computational complexity with the number of parameters. Secondly, due to the discreteness of the process, the learning algorithm is unstable: a small change in the data could lead to a significant change in the outcome of the process.

In order to address the above mentioned issues with subset selection approaches, $l_1$ norm based approaches were proposed. [19, 20] were pioneering works in geophysics; Tibshirani [10] introduced them as a lasso estimator in statistics, and Chen [15] proposed basis pursuit formulations in signal processing. $l_1$ optimization corresponds to the case where $R(\mathbf{w}) = ||\mathbf{w}||_1 = \sum_{i=1}^{p} |w_i|$ which is a continuous convex surrogate of the $l_0$ penalty, and hence it overcomes some of the issues with $l_0$ optimization: first, since the penalty is convex continuous, we no longer have to deal with combinatorial optimization procedures. Instead, we can use well developed convex optimization tools to significantly improve computational efficiency. Secondly, since the process is continues, it imparts stability to the learning algorithms. Despite these benefits, there are certain drawback: $l_1$

approaches suffer from over sparsification. If there is high correlation between variables, $l_1$ methods only select one or few from the set of these variables, and do not care which one to chose. This leads to performance degradation as they might select irrelevant or redundant variables. Due to the same reason, $l_1$ estimators are inconsistent under high correlation settings. Secondly, $l_1$ methods do not have a grouping effect. They select only one variable from a group of variables [28]. Some variants of $l_1$ approaches have been proposed to overcome these limitations: Elastic net [28] was introduced to induce grouping effect, and adaptive techniques [29, 30] have been proposed to impart consistency to the $l_1$ based estimators.

### 1.2.2 Bayesian sparse learning

In general, Bayesian learning can be represented by the following relation:

$$P(\mathbf{w}|D) = P(D|\mathbf{w}) \times P(\mathbf{w})/Z$$

where $P(D|\mathbf{w})$ is the likelihood, $Z$ is a partition function, $P(\mathbf{w})$ is a prior, and $P(\mathbf{w}|D)$ is the posterior. In the context of sparse learning, $P(\mathbf{w})$ are the sparsity inducing priors, and they are the Bayesian counter part to $R(\mathbf{w})$ in the frequentist settings. We can establish a one to one correspondence between frequentist regularization penalties and sparse priors by transforming the priors in the log domain. For example:

- $l_1 \leftrightarrow$ Laplace prior
- $l_2 \leftrightarrow$ Gaussian priors.

In general, frequentist regularized optimization has correspondence with the optimization of $-log(P(\mathbf{w}|D))$.

Background on sparse Priors

Most sparse Bayesian learning approaches assume independence in the components ($w_i$) of the model parameter vector $\mathbf{w}$, and employ a sparse prior on each component separately which can be represented as a scale mixture of normals [31, 32, 33]:

$$p(w_i) = \int \mathcal{N}(w_i; 0, \sigma_0 \gamma_i) f(\gamma_i) \mathrm{d}\gamma_i \tag{1.1}$$

where $\sigma_0$ can be assumed to be fixed, or can be applied a hyperprior. $\gamma_i$ can be thought of as the relevance measure for the component $w_i$. If $\gamma_i$ is large it encourages larger values for $w_i$, smaller values lead to shrinkage of $w_i$ [34]. Different choices for the prior on $\gamma_i$ define a whole spectrum of sparse priors on $w_i$. For example, assuming an inverse gamma prior on $\gamma_i$ leads to a student t prior on $w_i$ [34], and defining $f(\gamma_i) \propto \gamma_i^{-1}$ recovers the ARD prior [35, 36].

A unified perspective on sparse priors can be achieved by employing a generalized beta distribution of the second kind for $\gamma_i$ [34]:

$$p(\gamma_i) = \gamma_i^{a-1}(1 + \gamma_i/d)^{-a-b} d^{-a} / B(a, b) \tag{1.2}$$

where $B(a, b)$ is a the beta function. A horse shoe prior [37] corresponds to the case where $a = b = 1/2$ [34]. The normal exponential gamma prior [38] corresponds to case $a = 1$ [34] whereas normal gamma prior [38] is obtained by setting $b = d = \infty$ [34]. When the above two conditions are met at the same time, we recover the prior for Bayesian lasso [39].

### 1.2.3 Group sparsity and structured sparsity

In many applications, one needs to select groups of variables instead of individual variables. For example, in signal processing, a group could be defined according to neighbourhood relationship of wavelet coefficients [1]. Then, if it is known that the data could be explained by a few groups of variables, selecting those relevant groups enhances the pre-

dictive performance and interpretability of the output [1, 40, 41, 42, 43]. Group lasso [44] is a popular approach for inducing group level sparsity, and is defined as:

$$R(\mathbf{w}) = \sum_{g \in \mathcal{G}} ||\mathbf{w}_g||_q$$

where $|| \cdot ||_q$ is either $l_2$ or $l_\infty$ norm. The penalty corresponding to $q = 2$ was introduced in [45, 46], and for $q = \infty$ in [40]. The Group lasso penalty can be seen as an $l_1$ penalty on the vector $[||\mathbf{w}_g||_q]_{g \in \mathcal{G}}$ of size $|\mathcal{G}|$ [1]. Hence, it induces sparsity at the group level. There is also a strong link between this penalty and the group thresholding approach for wavelets in signal processing [1].

In a Bayesian framework, sparse learning techniques employ special priors that enforce sparsity on groups. [47] provides a unified perspectives on some of these sparse priors by presenting the idea of scale mixtures of multivariate Normals [34]:

$$p(\mathbf{w}_g) = \int \mathcal{N}(\mathbf{w}_g; \mathbf{0}, \Lambda_g u_g) f(u_g) \mathrm{d}u_g \tag{1.3}$$

where $\mathbf{w}_g$ consist of variables in group $g$, $u_g$ is the group level scalar parameter of relevance, and $\Lambda_g$ is a diagonal matrix of variances [34]. By defining $f(u_g)$ appropriately, one can recover group horseshoe, group ARD [48] and Bayesian group lasso [49] from this formulation[34].

Group sparsity is the simplest case of a more general notion called structured sparsity [50, 51, 52, 53]. In this setting, regularization functions are specifically designed to enforce sparsity with a particular structure [1]. For example, NP hard combinatorial approaches for overlapping groups were proposed in [53, 54], generating sparse solutions having support as the union of few number of groups [1]. In order to address some of the drawbacks of these discrete approaches, a convex relaxation of the penalties proposed in [53] was proposed [51]. Group lasso penalty has also been applied to cases with overlapping groups [50, 52]. This strategy was exploited in [50] to develop a regularization penalty that encourages sparse solutions with rooted subtree structures [1]. The idea is very similar to the

zero tree coding scheme [55] in signal processing literature as it attaches the relevance of a variable to its parent variable in the subtree [1]. A generalization of the work in [55] was proposed in [52] to address more general group structures [1]. Network constraint regularization approaches [56] have also been proposed to address complex variable correlation structures. These methods are applied under the settings in which each variable is represented by a node in a network, and an edge between two nodes represents some association or correlation between variables. Any kind of network topology can be encoded through, for example, a Laplacian matrix which is subsequently employed to design a network constraint regularization penalty that encourages smooth solutions satisfying the constraints enforced by the topology of the network.

## 1.3 Main goal of the dissertation

In the first part of this dissertation we focus on the issue of structured sparsity, and perform variable selection in a high dimensional structured space [57]. Instead of simply relying on the inherent characteristics of sparsity inducing strategies, we focus on exploiting valuable structure information about the variables that is available in various domains to enhance the structured sparsity inducing effect of a statistical model. We consider data sets in which the number of variables far exceed the number of samples. Below we describe a motivating example to show the application of this issue:

- By capturing various biochemical interactions, biological pathways provide insight into underlying biological processes. Given high-dimensional microarray or RNA-sequencing data, a critical challenge is how to integrate them with rich information from pathway databases to jointly select relevant pathways and genes for phenotype prediction or disease prognosis. Addressing this challenge can help us deepen biological understanding of phenotypes and diseases from a systems perspective. In the context of parametric linear models, features are the genes, and the response is the related phenotype. The high dimensional space of genes is a highly structured feature space consisting of groups of highly correlated genes. These groups of highly

correlated genes behave as one functional group or pathway. The task is to perform good phenotype prediction, and at the same time select important pathways and genes relevant to the phenotype guided by the structural information revealed by the pathways.

In the above example, the known structural information about the variables can be represented by an undirected graph [57]. In this graph, each node represents a gene, and an edge between two nodes represents some association or correlation between the genes. We need to exploit this valuable correlation structure information to develop our sparse modelling framework. With this approach we not only select relevant groups, but induce structured sparsity within selected groups dictated by this structural information. Bayesian paradigm is an appropriate choice to incorporate this prior structural information [57]. Hence, we focus on sparse Bayesian learning in this dissertation. In Bayesian learning, the choice of prior could be crucial in the performance of the learning algorithm. Our choice in this dissertation is the spike-and-slab prior [58, 59]. Our choice for spike-and-slab priors is dictated by the selective shrinkage effect induced by the priors. We will exploit this critical property to perform variable selection in high dimensional spaces [57]. However, since classical formulation of spike-and-slab priors does not explicitly incorporate correlation structure information about the variables which is available from various domains, we propose a generalized formulation of the classical prior for group selection that incorporates this valuable structure information, and allows our model to generate structured sparse results for each selected group. This structural constraint regularization capability, guided by domain knowledge, combined with the selective shrinkage effect makes our proposed prior an attractive tool for sparse Bayesian learning. Details of the spike-and-slab prior and selective shrinkage effect are given in the later sections of this introduction.

Sparse Bayesian learning using spike-and-slab priors, though attractive, is not free of challenges. A major computational bottleneck is the Bayesian inference of intractable posterior distribution. Due to the intractability of the posterior, approximate inference techniques have to be employed. Stochastic approximate inference methods (MCMC) [59], though provide convergence guarantees, are extremely slow to converge at high dimen-

sions. Deterministic approximate inference approaches, (Expectation propagation (EP), Variational Bayes (VB)) [60, 61] have to impose factorization constraints on the posterior to improve computational efficiency, but at the cost of reduced approximation quality. The second part of this dissertation focuses on developing fast posterior inference strategies for sparse Bayesian classical spike-and-slab models. The goal is to perform fast posterior inference, but without compromising the quality of the posterior approximation significantly, and by imposing no factorization constraints on the joint posterior. The details will be given in the related chapter.

## 1.4 Organization of the dissertation

Section 1.3 gives a background on spike-and-slab priors. Contributions for this dissertation are briefly described in 1.4. Chapter 2 and 3 discuss the technical details of the contributions. Specific terminologies and mathematical symbols are explained in each chapter. Related work is also provided in each chapter based on the context of each chapter. Finally, Chapter 4 summarizes the dissertation and lists possible directions for future work.

### 1.4.1 Spike-and-slab priors

Recently, spike-and-slab priors have been very popular in the sparse machine learning community. This popularity stems from the selective shrinkage property of the spike-and-slab priors: unlike the $L_1$ penalization, (i.e., equivalently, the Laplace prior) which shrinks all features—regardless of relevance or not—in the same way, the spike-and-slab prior is a mixture of two components: one component regularizes relevant variables mildly while the other one shrinks irrelevant variables aggressively. Let us consider an example to further illustrate the concept of selective shrinkage. In this example, we will going to compare the regularization penalty of lasso with adaptive lasso [29]:

- Lasso: $R(\mathbf{w}) = \lambda \sum_{i=1}^{p} |w_i|$
- Adaptive Lasso: $R(\mathbf{w}) = \sum_{i=1}^{p} \lambda_i |w_i|$

By comparing the two penalties above it can be seen that while in lasso there is only one regularization coefficient $\lambda$ for all components, adaptive lasso contains a separate regularization coefficient $\lambda_i$ for the $i_{th}$ component, and it is assigned a value $\lambda_i = 1/w_{OLSi}$. Here, $w_{OLSi}$ is the solution of the ordinary least squares problem, and it is a measure of relevance for the $i_{th}$ feature. This is unlike lasso where $\lambda$ is not affected by the relevance of the features. If the value of $w_{OLSi}$ is small, it indicates the irrelevance of the feature, and $\lambda_i$, in adaptive lasso, is set to a large value which shrinks the $i_{th}$ component to $0$. On the other hand, larger values of $w_{OLSi}$ indicate the importance of the $i_{th}$ feature to the response variable, $\lambda_i$ is now set to a small value which encourages the $i_{th}$ component to take bigger values. In this scheme, the output of ordinary least squares algorithm is used as a measure of relevance to decide whether to shrink a variable , or not. This is the basic idea behind selective shrinkage. In adaptive lasso, however, the selective shrinkage effect is not induced in a principled manner. The formulation of the adaptive lasso penalty does not create this effect inherently. It requires a preprocessing step to induce such an effect. Unlike adaptive lasso, spike-and-slab priors do not require any preprocessing step, the formulation of the priors inherently generates this affect.

A pioneering work on spike-and-slab priors was done by Lampers [58] and Mitchels and Baeuchamp [59]. They proposed a two point mixture distribution for the regression weight vector $\mathbf{w}$ which consisted of uniform flat distribution (slab) and a degenerate distribution at zero(spike) [62]. In this dissertation, we adopt the prior formulation proposed by [62]. Their model formulation is significantly different from the classical one proposed by [58], but it essentially creates the same effect as the original model. In their model, $\mathbf{w}$ is assigned a multivariate Normal scale mixture distribution specified through the prior on hyper variances $\gamma$ [62]:

$$(\mathbf{w}|\gamma) \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \Gamma) \tag{1.4}$$
$$\gamma \sim \pi(\mathrm{d}\gamma)$$

where $\mathbf{0}$ is a $p$ dimensional zero vector, $\Gamma$ is a $p \times p$ diagonal matrix $\mathrm{diag}(\gamma_1, ..., \gamma_p)$, and $\pi$ is a prior measure for $\gamma = (\gamma_1, ..., \gamma_p)^\top$ [62]. In this setting, sparsity is achieved through manipulating the values of the hyper variances of the Normal distributions. Smaller values of the hyper variances force the coefficients to zero while larger values inflate coefficients [62]. In this dissertation, we will more specifically focus on spike-and-slab prior formulations introduced by George and McCulloch [63] which can be considered as a special case of the prior introduced by [62]. In their prior settings, a two point discrete prior is assigned on each hyper variance $\gamma_j$; the complete hierarchical prior for each $w_j$ is as follows:

$$
\begin{aligned}
(w_j|\gamma_j) &\overset{ind}{\sim} \mathcal{N}(w_j|0, \gamma_j) \\
(\gamma_j|z_j) &\overset{ind}{\sim} (1 - z_j)\delta_{r_0}() + z_j\delta_{r_1}() \\
(z_j|s_j) &\overset{ind}{\sim} (1 - s_j)\delta_0() + s_j\delta_1()
\end{aligned}
$$

where $z_j$ is a binary indicator variable for feature selection, and $s_j \in [0, 1]$ represents the selection probability for the $j$-feature, $j$ varies from 1 to $p$. By marginalizing out $\gamma_j$, the hierarchical prior for $w_j$ reduces to:

$$
\begin{aligned}
(w_j|z_j) &\overset{ind}{\sim} \mathcal{N}(w_j|0, r_0)^{(1-z_j)}\mathcal{N}(w_j|0, r_1)^{z_j} \\
(z_j|s_j) &\overset{ind}{\sim} (1 - s_j)\delta_0() + s_j\delta_1()
\end{aligned}
\tag{1.5}
$$

where $r_0$ and $r_1$ are the variances of the two Gaussian components. To provide the required selective shrinkage, the spike component is assigned a very small variance, hence it is concentred around zero and favors less significant variables, while the slab component is assigned a large variance which leads to a mild regularization of relevant variables. This prior formulation switches its regularization level based on the indicator variable $z_j$, which is either 0 or 1. Therefore, complete separation of spike and slab regularizations is achieved, and hence it creates a very strong selective shrinkage effect.

If $z_j$ is marginalized, the prior becomes:

$$(w_j|s_j) \overset{ind}{\sim} \mathcal{N}(w_j|0, r_0) \times (1 - s_j) + \mathcal{N}(w_j|0, r_1) \times (s_j) \tag{1.6}$$

This prior setting switches its regularization level based on the value of $s_j$. Since $s_j$ is a continuous variable taking a range of values from the set $[0, 1]$, complete separation of spike and slab penalizations is not possible, and hence selective shrinkage is not as strong as the previous formulation.

Assuming a beta prior for $s_j$ with parameters $a_0 = b_0 = 1$ and marginalizing it we get:

$$p(w_j) = \frac{1}{2}\mathcal{N}(w_j|0, r_1) + \frac{1}{2}\mathcal{N}(w_j|0, r_0) \tag{1.7}$$

In this setting, the prior always has a mixture of two penalizers. Since the mixture weights are not influenced by features, this prior formulation has the least selective shrinkage effect.

## 1.5 Contributions of the dissertation

The contribution of this dissertation is two fold: we propose a new sparse Bayesian approach, called Network and Node Selection (NaNOS), for joint group and feature selection; and a scalable Laplace approximation for sparse Bayesian spike-and-slab models (FLAS).

### 1.5.1 Efficient spike-and-slab models for joint group and feature selection

Various domains provide correlation structure information about the variables which could be helpful for sparse learning in revealing the underlying sparsity pattern. Classical formulation of spike-and-slab priors is not designed to incorporate this correlation structure information. The embedding of this critical information into the classical formulation could greatly enhance the modelling power of the priors. To accomplish this task, we present a novel sparse Bayesian model for joint network (group) and node (features) selection (NaNOS). Our model includes a generalized formulation of the classical spike-

and-slab prior for group selection that incorporates domain based correlation structure information for each group. This generalization step allows the prior to induce a structural constraint regularization effect on the selected groups. Combined with the selective shrinkage effect, this generalized prior can serve as an attractive tool for sparse Bayesian learning. Secondly, our model provides a principled framework for exploiting this correlation structure information in the joint selection of relevant groups as well as relevant features within the selected groups. Specifically, our model is a combination of conditional and generative components: the conditional component includes the generalized spike-and slab prior that induces network level sparsity via the selective shrinkage effect, and imposes structural constraints, guided by domain knowledge, on each network through the use of graph Laplacian matrices. The generative component imposes node level sparsity, within a network, through the application of standard spike-and-slab prior on the network nodes. This modelling approach will find its application in genomic data analysis where there is a need to jointly discover pathways and genes that are relevant for phenotype prediction or disease prognosis.

### 1.5.2 Scalable sparse Bayesian learning for spike-and-slab models

We consider the application of Bayesian spike-and-slab models in high-dimensional feature selection problems. To do so, we propose simple yet effective fast approximate Bayesian inference algorithms based on Laplace's method (FLAS,FLAS*,FLAS**). We exploit two efficient optimization methods, GIST [64] and L-BFGS [65], to obtain the mode of the posterior distribution. Then we propose an ensemble Nyström approach to calculate the diagonal of the inverse Hessian over the mode to obtain the approximate posterior marginals in $O(knp)$ time, $k << p$. The theoretical analysis of the ensemble method is also provided. With the posterior marginals of model weights, we use quadrature integration to estimate the marginal posteriors of selection probabilities and indicator variables for all features, which quantify the selection uncertainty. Unlike existing approximate inference methods, our approach does not require any factorization constraints on the posterior

to enhance its computational efficiency. Our method can be seen as a hybrid of Bayesian and frequentist treatments taking benefits from both worlds. From a frequentist perspective, our approach is computationally efficient, and possesses asymptotic consistency properties; and from a Bayesian point of view, it performs posterior inference better than or comparable to existing approximate inference methods. On simulated data, our methods perform feature selection better than or comparable to the alternative approximate methods, with less running time, and provide higher prediction accuracy than various sparse methods including VB, EP, automatic relevance determination, lasso, elastic net and a capped-$L_1$ method. On large real benchmark datasets, our methods often achieve improved prediction accuracy compared to alternative methods, but with a convergence time comparable to frequentist $l_1$ methods. Finally, application on Region-of-Interest (ROI) analysis of high dimensional brain image data shows interesting discoveries, many of which are supported by existing literature.

### 1.5.3 Significance of the contributions

In the final part of this chapter we present two tables highlighting the significance of our contributions compared to the existing trends.

As can be seen from table 1.1, our method, NaNOS, not only performs group and feature selection, but exploits the domain based correlation structure information provided for each group to induce structured sparsity within each selected group. Since our model builds upon spike-and-slab formulation, it also has the advantage of selective shrinkage effect. In brief, our method combines the strong points of existing approaches in a principled manner, and show improved performance both in terms of prediction and selection results.

Table 1.2 compares our approaches (FLAS, FLAS*,FLAS**) with popular approximate inference approaches. The table clearly shows the benefit of our approaches compared to others. While Expectation propagation (EP) and Variational Bayes (VB) require structural or factorization constraints in the joint posterior to achieve linear time complexity, our methods do not require such constraints, and still achieve linear time complexity. The fac-

torization constraints might lead to performance degradation especially when the variables are correlated. Since we do not impose such constraints, our approaches perform better than or comparable to the existing approximate inference methods both in terms of prediction and selection results. Lastly, although, MCMC based posterior inference techniques provide convergence guarantees, they are extremely slow to converge at high dimensions, and hence not scalable.

Table 1.1.: Sparsity inducing methods. EN: elastic net; $ss$: classical spike-and-slab; NCR: network constraint regularization

| Approaches | $l_1$ | Group $l_1$ | EN | $ss$ | Group $ss$ | NCR | NaNOS |
|---|---|---|---|---|---|---|---|
| Feature sparsity | Yes | No | Yes | Yes | No | Yes | Yes |
| Group sparsity | No | Yes | No | No | Yes | No | Yes |
| Selective shrinkage | No | No | No | Yes | Yes | No | Yes |
| Structured sparsity | No | No | No | No | No | Yes | Yes |
| Sparsity within groups | No | No | No | No | No | No | Yes |

Table 1.2.: Posterior inference methods for spike-and-slab models. EP: Expectation propagation; VB: Variational Bayes; MCMC: Markov Chain Monte Carlo. $n$ is the number of samples and $p$ is the number of dimensions.

| methods | VB | EP | MCMC | FLAS,FLAS*,FLAS** |
|---|---|---|---|---|
| Factorization constraints | Yes | Yes | No | No |
| Computational complexity | $O(np)$ | $O(np)$ | very high | $O(np)$ |

# 2 EFFICIENT SPIKE-AND-SLAB MODELS FOR JOINT GROUP AND FEATURE SELECTION

## 2.1 Motivation

A major objective of sparse learning is to strike a balance between predictive power and generating results that are more interpretable. The task becomes more complicated if there is high correlation between variables. Various frequentist approaches try to address this issue by proposing smoothing penalties or incorporating sparsity constraints. They recast a sparse learning problem into an optimization problem, and embed sparsity into the optimization procedure. In a Bayesian setting, this regularization effect is achieve through the application of sparsity inducing priors. Sparse priors essentially encode our prior beliefs about the sparsity pattern of the parameters. Recently, spike-and-slab priors have been very popular in the sparse machine learning community. This popularity stems from the selective shrinkage property of the spike-and-slab priors: unlike most of the frequentist approaches, these priors selectively shrink irrelevant variables, and mildly regularize relevant ones. However, classical formulation of spike-and-slab priors is not designed to incorporate correlation structure information about the variables which is provided by various domains, and could be helpful in revealing the sparsity pattern. The embedding of this critical information into the classical formulation could greatly enhance the modelling power of the spike-and-slab priors.

To accomplish this task, we present a novel sparse Bayesian model for joint network (group) and node (features) selection. Specifically, our model is a combination of conditional and generative components: the conditional component includes the generalized spike-and slab prior that induces network level sparsity via the selective shrinkage effect, and imposes structural constraints, guided by domain knowledge, on each network through the use of graph Laplacian matrices, details will be given later. The generative component

imposes node level sparsity, within a network, through the application of standard spike-and-slab prior on the network nodes. The integration of these two components provides a principled framework for joint selection of networks and relevant correlated nodes in the selected networks, guided by the domain based correlation structure information. To make the selection process efficient, we employ a variational Bayes procedure for the Bayesian inference.

In order to demonstrate the predictive power and selection accuracy of our model, we conduct extensive simulation experiments. The simulation results clearly reveal the advantage our method has over other alternative approaches. We also apply our method for genomic data analysis. We use three expression datasets for cancer study and the KEGG pathway database. The pathways and genes selected by our method are shown to be quite relevant to the cancer growth. Some of the pathways and genes are also supported by existing biological literature.

## 2.2 Model

This section discusses the specific details of our hybrid sparse Bayesian model, NaNOS, for network and node selection. Assume $n$ independent and identically distributed samples $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$, where $\mathbf{x}_i$ is the $p$ dimensional node vector of the $i$-th sample, and $t_i$ is its response. Our aim is to predict the response vector $\mathbf{t} = [t_1, \ldots, t_n]^\top$ based on the design matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ and selecting a small number of networks as well as nodes within selected networks relevant to the prediction. For real-world scenarios, we have $n \ll p$ in many cases, and hence the selection task becomes challenging.

To perform variable selection efficiently, we can exploit the valuable correlation structure information of the nodes encoded in the networks. For example, biological pathways consist of a set of highly correlated genes acting together to perform certain biological functions. Hence, representing various gene interactions. Assume that we have $M$ networks, we organize the node vector $\mathbf{x}_i$ into $M$ subparts, each part corresponds to one of

the $M$ networks. If a node appears in multiple networks, we distribute its value across all participating networks.

Following the general framework proposed by [66], we formulate our model as a hybrid of conditional and generative formulations: the conditional component induces network level sparsity, and selects relevant networks; the generative component enforces node level sparsity within selected networks; and the two models are linked through a joint prior distribution. In this modelling framework, both the conditional and generative components influence and help each other to facilitate the joint selection of network and nodes.

For the conditional component, we use Gaussian likelihood function for regression analysis:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \tau) = \prod_{i=1}^{n} \mathcal{N}(t_i|\mathbf{x}_i^\top \mathbf{w}, \tau^{-1}). \tag{2.1}$$

where $\mathbf{w}$ are regression weights, and $\tau$ is the precision parameter. We employ a diffuse Gamma prior, $\mathrm{Gam}(\tau|g, h)$ with $g = h = 10^{-6}$ for $\tau$.

For classification, we use a logistic likelihood

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \sigma(\mathbf{x}_i^\top \mathbf{w})^{t_i}[1 - \sigma(\mathbf{x}_i^\top \mathbf{w})]^{1-t_i}, \tag{2.2}$$

where $t_i \in \{0, 1\}$, $\mathbf{w}$ are classifier weights, and $\sigma(\cdot)$ is the logistic function (i.e., $\sigma(y) = (1+\exp(-y))^{-1}$). We divide the vector $\mathbf{w}$ into $M$ subparts, each part corresponding to one of the $M$ networks. The partitioned vector $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]^\top$ where $\mathbf{w}_k$ are the weights for the node variables in the $k$-th network.

To exploit the correlation structure information embedded in a network, we compute the normalized Laplacian matrix representation of the network. More Specifically, if we are given the adjacency matrix $\mathbf{G}_k$ of the $k$-th network where each entry of the adjacency matrix represents edges between nodes in the $k$-th network, the normalized Laplacian matrix $\mathbf{L}_k$ is defined as

$$\mathbf{L}_k(i,j) = \begin{cases} 1 & i = j \text{ and } d_i \neq 0 \\ -\frac{1}{\sqrt{d_i d_j}} & i \neq j \text{ and } \mathbf{G}_k(i,j) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{G}_k$ is the adjacency matrix for the $k$-th network, and $d_i = \sum_j \mathbf{G}_k(i,j)$ is the degree of the $i$-th node in the $k$-th network.

Once the graph Laplacian matrices have been computed, we utilize them to formulate a sparse prior over $\mathbf{w}_k$. The prior is essentially a generalization of the classical spike-and-slab prior for sparse group selection:

$$p(\mathbf{w}_k|\alpha_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0}, s_1\mathbf{L}_k^{-1})^{\alpha_k}\mathcal{N}(\mathbf{w}_k|\mathbf{0}, s_2\mathbf{I}_k)^{1-\alpha_k} \tag{2.3}$$

where $\alpha_k$ is a binary variable indicating the selection of the $k$-th network, $s_1 > s_2$, $s_2 \approx 0$, and $\mathbf{I}_k$ is an identity matrix. $s_1$, and $s_2$ are computed based on cross-validation procedure. $\mathbf{L}_k$ is a positive semi definite matrix based on the original definition, and hence using the inverse of $\mathbf{L}_k$ as the covariance matrix of a Gaussian distribution is not justified. Therefore, we deviate slightly from the classical definition, and add a scaled diagonal matrix $10^{-6}\mathbf{I}_k$ to $\mathbf{L}_k$. The diagonal perturbation does not disturb the correlation information of the $k$-th network encoded by $\mathbf{L}_k$, and hence fits well into our modelling framework. The classical spike-and-slab prior for group selection is a special case of our general framework. Indeed, if $\mathbf{L}_k$ is replaced by $\mathbf{I}_k$, the prior (2.3) reduces to the classical prior [67]. We can analyze the regularization effect of the generalized prior by transforming the prior in the log domain. Taking the negative log of the prior gives us the following expression:

$$-log(p(\mathbf{w}_k|\alpha_k)) = \frac{\alpha_k}{2s_1}\mathbf{w}_k^\top\mathbf{L}_k\mathbf{w}_k + \frac{1-\alpha_k}{2s_2}||\mathbf{w}_k||_2^2 \tag{2.4}$$

When the binary indicator variable $\alpha_k = 0$, due to very small value of $s_2$, the regularization effect is similar to the square of the $l_2$ penalty with very large regularization coefficient ($\frac{1}{s_2}$). Consequently, $\mathbf{w}_k$ vector is shrunk towards zero. On the other hand, if $\alpha_k = 1$, the prior has a network constraint regularization effect given by the following expression:

$$\mathbf{w}_k^\top\mathbf{L}_k\mathbf{w}_k = \sum_{(i,j)\in E_k}(\frac{w_i}{\sqrt{d_i}} - \frac{w_j}{\sqrt{d_j}})^2 \tag{2.5}$$

where $E_k$ is the edge set of the $k_{th}$ network. From the expression above it can be seen that correlated nodes with similar degrees, within the $k$-th network, are encouraged to have similar weight values. If two connected nodes have different degrees, the one with a higher degree is given more weigh. This is a desirable property because nodes with higher degrees are expected to be more influential, and hence more significant.

From the discussion above it can be seen that our generalized prior combines the selective shrinkage effect with network regularization ability in a principled manner, and plays a critical role in giving our model the capability to induce the required structured sparsity effect.

To model uncertainty in $\alpha_k$, we assign a Bernoulli prior distribution: $p(\alpha_k) = (u_k)^{\alpha_k}(1-u_k)^{1-\alpha_k}$, $u_k \in [0,1]$ is the selection probability. We assign an uninformative prior over $u_k$: $p(u_k) = 1$ (i.e.,, $p(u_k) = \text{Beta}(u_k; a, b)$ where $a = b = 1$).

For the purposes of selecting relevant nodes within each selected network, for each network $k$, we introduce a latent auxiliary vector $\tilde{\mathbf{w}}_k$ in the generative model. The vector $\tilde{\mathbf{w}}_k$ is tightly linked to the $k$-th network vector $\mathbf{w}_k$ via a special linking prior distribution. The details will be given shortly. To induce sparsity into the vector $\tilde{\mathbf{w}}_k$, we simply apply the classical spike-and-slab prior:

$$\begin{aligned} p(\tilde{\mathbf{w}}_k|\boldsymbol{\beta}_k) &= \prod_{j=1}^{p_k} \mathcal{N}(\tilde{w}_{kj}|0, r_1)^{\beta_{kj}} \mathcal{N}(\tilde{w}_{kj}|0, r_2)^{1-\beta_{kj}} \\ &= \prod_{j=1}^{p_k} \mathcal{N}(0|\tilde{w}_{kj}, r_1)^{\beta_{kj}} \mathcal{N}(0|\tilde{w}_{kj}, r_2)^{1-\beta_{kj}} \\ &= p(\mathbf{0}|\tilde{\mathbf{w}}_k, \boldsymbol{\beta}_k) \end{aligned} \tag{2.6}$$

where $p_k$ is the size of the $k$-th network, $r_2 \approx 0$, and $\beta_{kj}$ is a binary indicator variable for the $j$-th node in the $k$-th network. We assign a Bernoulli prior to $\beta_{kj}$: $p(\beta_{kj}) = (v_{kj})^{\beta_{kj}}(1-v_{kj})^{1-\beta_{kj}}$, and a uniform prior to $v_{kj}$: $p(v_{kj}) = 1$ From the rearrangement shown above, it can be seen that from a modelling perspective, the spike-and-slab prior term $p(\tilde{\mathbf{w}}_k|\boldsymbol{\beta}_k)$ and $p(\mathbf{0}|\tilde{\mathbf{w}}_k, \boldsymbol{\beta}_k)$ will have the same effect on our model. The term $p(\mathbf{0}|\tilde{\mathbf{w}}_k, \boldsymbol{\beta}_k)$ can be considered as a generative model, the observation $\mathbf{0}$ is sampled from $\tilde{\mathbf{w}}_k$. Since a

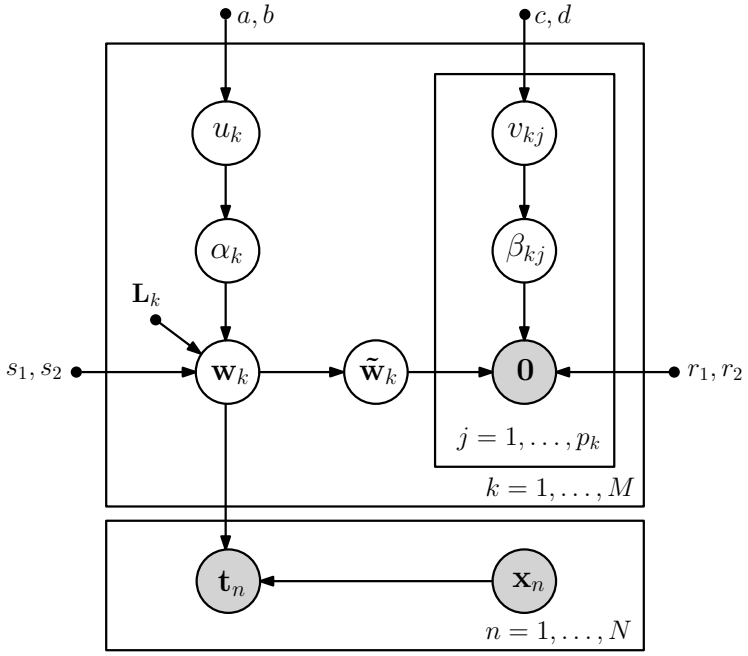



Figure 2.1.: The graphical model representation of NaNOS.

variable can not be sampled twice in a Bayesian network, this rearrangement is essential for our modelling framework as it allows the spike-and-slab prior term $p(\tilde{\mathbf{w}}_k|\boldsymbol{\beta}_k)$ to be incorporated into our Bayesian model. This trick also allows us to integrate the two sparse components of our modelling framework in a principled manner.

As explained earlier, we design a special prior distribution to establish a link between the conditional and generative components. To accomplish this, we propose the following prior on $\tilde{\mathbf{w}}_k$:

$$p(\tilde{\mathbf{w}}_k|\mathbf{w}_k) = \mathcal{N}(\tilde{\mathbf{w}}_k|\mathbf{w}_k, \lambda\mathbf{I}) \tag{2.7}$$

Since lambda is the variance parameter of the Gaussian distribution, it controls the degree to which $\tilde{\mathbf{w}}_k$ is concentrated around $\mathbf{w}_k$. In the limit $\lambda \to 0$, the Gaussian distribution approaches the delta function and with probability 1, $\mathbf{w}_k = \tilde{\mathbf{w}}_k$. We enforce this equality constraint by setting $\lambda = 0$. The constraint allows us to not only influence the $\mathbf{w}_k$ vector as a whole, but also its individual components via the vector $\tilde{\mathbf{w}}_k$. It is essentially this feature that allows sparsity at both the network and node level. Figure 2.1 shows the graphical model diagram for our joint model.

Probabilistically speaking, our modelling framework ensures consistency in the selection of networks and nodes. If a network is discarded, all nodes contained in that network are removed. Our model ensures this by enforcing an equality constraint $\mathbf{w}_k = \tilde{\mathbf{w}}_k$ through the delta prior $\delta(\tilde{\mathbf{w}}_k - \mathbf{w}_k)$. Hence, when $\alpha_k = 0$, $\mathbf{w}_k = \mathbf{0}$ implies $\tilde{\mathbf{w}}_k = \mathbf{0}$. Consequently, the spike component will be dominant for all the nodes in the $k$-th network, and force the weight values of the nodes towards zero. Our modelling framework also ensures that if at least one node in a network is selected, then that network is also selected. One novel feature of our approach is that it does not impose the hard consistency constraint: our model will not select all the networks that share one common selected node. We avoid this constraint by duplicating the value of the common selected node across all the participating networks, and using the duplicated weights as separate model parameters.

## 2.3 Algorithm

In this section we will explain the Bayesian inference algorithm of our model. In order to perform efficient Bayesian inference, we employ variational Bayesian (VB) approach for approximate inference. Specifically, we present variational updates equations to approximate the posteriors of $\mathbf{w}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{v}$, and $\tau$, $\tau$ is required for regression only. Once the posteriors have been computed, network and node selection can be performed based on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

The joint posterior distribution for our regression model is

$$p(\mathbf{w}, \tilde{\mathbf{w}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \tau | \mathbf{t}, \mathbf{X}) = \frac{1}{Z}\mathcal{N}(\mathbf{t}|\mathbf{X}\mathbf{w}, \tau^{-1}\mathbf{I})\text{Gamma}(\tau)\cdot$$

$$\prod_k p(\mathbf{w}_k|\boldsymbol{\alpha}_k)p(\tilde{\mathbf{w}}_k|\mathbf{w}_k)p(\mathbf{0}|\tilde{\mathbf{w}}_k, \boldsymbol{\beta}_k)\text{Bern}(\alpha_k|u_k)\text{Beta}(u_k)\cdot$$

$$\prod_j \text{Bern}(\beta_{kj}|v_{kj})\text{Beta}(v_{kj}) \tag{2.8}$$

where $p(\mathbf{w}_k|\boldsymbol{\alpha}_k)$ and $p(\mathbf{0}|\tilde{\mathbf{w}}_k, \boldsymbol{\beta}_k)$ are defined in (2.3) and (2.6), $p(\tilde{\mathbf{w}}_k|\mathbf{w}_k) = \delta(\tilde{\mathbf{w}}_k - \mathbf{w}_k)$, and $Z$ is the partition function. The expression for the joint posterior in the classification model is similar to (2.8), except that the Gaussian likelihood (2.1) is replaced by the logistic function (2.2), and the prior term for $\tau$ is removed from (2.8).

Stochastic approximate inference techniques such as Classical Markov chain Monte Carlo methods are attractive due to their convergence properties, but they lack scalability with respect to the number of dimensions. Even for moderate dimensions, the MCMC algorithms exhibit slow mixing times. In addition to that, there are no practical tools available to accurately gauge the convergence of MCMC samplers. Thus, we decide to employ computationally efficient variational Bayes procedure for approximate inference of (2.8).

In variational approximate inference, we enforce a factorization constraint on the exact joint posterior distribution, and try to learn this factorized distribution instead of the true posterior. Specifically, for our model, we learn the following factorized distribution as an approximation to (2.8): $Q(\boldsymbol{\theta}) = Q(\mathbf{w})Q(\boldsymbol{\alpha})Q(\boldsymbol{\beta})Q(\mathbf{u})Q(\mathbf{v})Q(\tau)$, where $\boldsymbol{\theta}$ combines all the variables in the distribution on which inference is being performed. It is to be noted that for the classification model, we do not need to do Bayesian inference for $Q_\tau(\tau)$ as it is not part of the joint distribution. Since we have enforced an equality constraint on $\tilde{\mathbf{w}}$ and $\mathbf{w}$, we do not need to separately update a posterior distribution for $\tilde{\mathbf{w}}$.

The variational inference procedure explores the space of factorized distributions of the form $Q(\boldsymbol{\theta})$, and seeks to find the optimal distribution that minimizes the KL divergence between the exact and the approximate posterior of $\boldsymbol{\theta}$:

$$\mathrm{KL}(Q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{t}, \mathbf{X})) = \int Q(\boldsymbol{\theta}) \ln \frac{Q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{t}, \mathbf{X})} \mathrm{d}\boldsymbol{\theta}. \tag{2.9}$$

The variational updates are derived by applying coordinate descent procedure to the KL divergence minimization problem. This leads to efficient update equations for the posterior distributions as explained in the coming sections. The overall procedure is iterative in nature: one posterior distribution is updated at a time by keeping all others fixed. This procedure is repeated until convergence is achieved. The variational updates ensure that

the value of KL divergence decreases with every iteration, and since KL divergence is bounded from below by zero, the procedure is guaranteed to converge [68].

### 2.3.1 Regression

The variational updates for the regression model are as follows:

$$
\begin{aligned}
Q(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma}) && (2.10)\\
Q(\boldsymbol{\alpha}) &= \prod\nolimits_k {\gamma_k}^{\alpha_k}(1-\gamma_k)^{1-\alpha_k} && (2.11)\\
Q(\boldsymbol{\beta}) &= \prod\nolimits_k \prod\nolimits_j (\eta_{kj})^{\beta_{kj}}(1-\eta_{kj})^{1-\beta_{kj}} && (2.12)\\
Q(\mathbf{u}) &\propto \prod\nolimits_k (u_k)^{\tilde{a}_k-1}(1-u_k)^{\tilde{b}_k-1} && (2.13)\\
Q(\mathbf{v}) &\propto \prod\nolimits_k \prod\nolimits_j (v_{kj})^{\tilde{c}_{kj}-1}(1-v_{kj})^{\tilde{d}_{kj}-1} && (2.14)\\
Q(\tau) &= \Gamma(\tau|\tilde{g}, \tilde{h}). && (2.15)
\end{aligned}
$$

The update equations for the parameters of the above posterior distributions are given by:

$$
\begin{aligned}
\mathbf{\Sigma} &= (\mathbf{A} + \langle\tau\rangle \mathbf{X}^\top \mathbf{X})^{-1} \qquad \mathbf{m} = \langle\tau\rangle \mathbf{\Sigma} \mathbf{X}^\top \mathbf{t} && (2.16)\\
\tilde{a}_k &= \gamma_k + a \qquad \tilde{b}_k = 1 - \gamma_k + b && (2.17)\\
\tilde{c}_{kj} &= \eta_{kj} + c \qquad \tilde{d}_{kj} = 1 - \eta_{kj} + d && (2.18)\\
\gamma_k &= 1/(1 + \exp(\langle \ln(1-u_k)\rangle - \langle \ln u_k\rangle + \frac{p_k}{2}\ln\frac{s_1}{s_2}\\
&\quad -\frac{1}{2}\ln|\mathbf{L}_k| + \frac{1}{2}\mathbf{tr}(\langle \mathbf{w}_k {\mathbf{w}_k}^\top\rangle(\frac{1}{s_1}\mathbf{L}_k - \frac{1}{s_2}\mathbf{I}_k)) && (2.19)\\
\eta_{kj} &= 1/(1 + \exp(\langle \ln(1-v_{kj})\rangle - \langle \ln v_{kj}\rangle\\
&\quad +\frac{1}{2}\ln\frac{r_1}{r_2} + \frac{1}{2}\langle (w_{kj})^2\rangle(\frac{1}{r_1} - \frac{1}{r_2}))) && (2.20)\\
\tilde{h} &= h + \frac{1}{2}\mathbf{t}^\top\mathbf{t} - \mathbf{m}^\top\mathbf{X}^\top\mathbf{t} + \frac{1}{2}\sum\nolimits_i \mathbf{x}_i^\top \langle \mathbf{w}\mathbf{w}^\top\rangle \mathbf{x}_i && (2.21)\\
\tilde{g} &= g + \frac{n}{2} && (2.22)
\end{aligned}
$$

where $\mathbf{A} = \frac{1}{s_1}\mathrm{diag}(\{\gamma_k \mathbf{L}_k\}_k) + \frac{1}{s_2}\mathrm{diag}(\{(1-\gamma_k)\mathbf{I}_k\}_k) + \frac{1}{r_1}\mathrm{diag}(\boldsymbol{\eta}) + \frac{1}{r_2}\mathrm{diag}(1-\boldsymbol{\eta})$. $\mathrm{diag}(\{\gamma_k \mathbf{L}_k\}_k)$ is a block-diagonal matrix, and $\langle\cdot\rangle$ denotes expectation with respect to a

posterior distribution. All the moments appearing in the update equations are computed as follows:

$$\langle \mathbf{w}\mathbf{w}^\top \rangle = \mathbf{\Sigma} + \mathbf{m}\mathbf{m}^\top \qquad \langle \tau \rangle = \tilde{g}/\tilde{h}$$
$$\langle \ln u_k \rangle = \psi(\tilde{a}_k) - \psi(\tilde{e}_k) \qquad \langle \ln(1-u_k) \rangle = \psi(\tilde{b}_k) - \psi(\tilde{e}_k)$$
$$\langle \ln v_{kj} \rangle = \psi(\tilde{c}_{kj}) - \psi(\tilde{f}_{kj}) \qquad \langle \ln(1-v_{kj}) \rangle = \psi(\tilde{d}_{kj}) - \psi(\tilde{f}_{kj})$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{dx}} \ln \Gamma(x)$ , $\tilde{e}_k = \tilde{a}_k + \tilde{b}_k$ and $\tilde{f}_{kj} = \tilde{c}_{kj} + \tilde{d}_{kj}$.

From a frequentist perspective the posterior mean of $\mathbf{w}$ can be shown to have estimation consistency for the case when $p = \sum_{k=1}^{M} p_k$ is fixed and $n \to \infty$. Let us assume that $\mathbf{w}_0$ is the true coefficient vector of the regression model. Define $S_0 = \{j : w_{0j} \neq 0\}$. Let $\mathcal{S}$ denote the space in which $S_0$ lies. We will use the following assumptions for the consistency proof:

**Assumption 1 [69].** Let $C_{SS} = n^{-1}(\mathbf{X}_S^\top \mathbf{X}_S)$ for any $S \in \mathcal{S}$. Let $\lambda_i$ be the $i_{th}$ eigenvalue of $C_{SS}$, then the following condition holds:

$$0 < c_1 \leq \lambda_{min}(C_{SS}) \leq \lambda_{max}(C_{SS}) \leq c_2 < \infty \tag{2.23}$$

**Assumption 2.** For parameters $r_1$, $r_2$, $s_1$, $s_2$, and $\tau$, assume that there exist finite positive constants $k_{low}$, $k_{up}$, $\tau_{min}$, and $\tau_{max}$ such that $k_{low} \leq r_1, r_2, s_1, s_2 \leq k_{up}$, and $\tau_{min} \leq \tau \leq \tau_{max}$.

Assumption 1 enforces positive definiteness of the sample covariance matrix. This assumption is reasonable for large sample sizes because the covariance matrix is full rank.

Assumption 2 is mild as it only requires a compact support for the parameters.

The form of the argument presented in the following theorem is very similar to the one given in [69], but it can be applied to prove estimation consistency result for our case.

**Theorem 2.3.1** *Assuming that 1, 2 are satisfied, with $p$ fixed, then*

$$P(||\mathbf{m} - \mathbf{w}_0||^2 > \xi_n) \leq c_0 \exp\{-\log(n\xi_n)\} \tag{2.24}$$

*for some positive finite constant $c_0$ and $\xi_n$. Assume that $\xi_n \propto n^{-\alpha^*}$ for some $\alpha^* > 0$. Then if $0 < \alpha^* < 1$, $P(||\mathbf{m} - \mathbf{w}_0||^2 > \xi_n) \to 0$ as $n \to \infty$.*

**Proof.** Let $k_1 = max(2\gamma_1, ..., 2\gamma_M)$, and $k_2 = max(1-\gamma_1, ..., 1-\gamma_M) < 1$. Since we assume positive definiteness of $\mathbf{X}^\top\mathbf{X}$, the proof is only valid for disjoint groups. In case of overlapping groups, we use duplication procedure that renders the matrix $\mathbf{X}^\top\mathbf{X}$ singular for any value of $n$. Since the maximum eigenvalue of every $\mathbf{L}_k$ is upper bounded by 2, and every element of the vector $\boldsymbol{\eta}$ and every $\gamma_k$ is less than 1, by Weyle's inequality,

$$\begin{aligned}
||\mathbf{A}||_2^2 &= \lambda_{max}^2(\mathbf{A}) \\
&\leq (\frac{1}{s_1}\lambda_{max}(\text{diag}(\{\gamma_k\mathbf{L}_k\}_k)) + \frac{1}{s_2}\lambda_{max}(\text{diag}(\{(1-\gamma_k)\mathbf{I}_k\}_k)) \\
&+ \frac{1}{r_1}\lambda_{max}(\text{diag}(\boldsymbol{\eta})) + \frac{1}{r_2}\lambda_{max}(\text{diag}(1-\boldsymbol{\eta})))^2 \\
&\leq (\frac{1}{s_1}k_1 + \frac{1}{s_2}k_2 + \frac{1}{r_1} + \frac{1}{r_2})^2 \\
&\leq (\frac{1}{k_{low}}(k_1+3))^2 \\
&= \mu_0^2
\end{aligned}$$

Now

$$\begin{aligned}
\mathbf{m} - \mathbf{w}_0 = -(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})^{-1}\mathbf{A}\mathbf{w}_0 + \\
\tau(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^\top\epsilon
\end{aligned}$$

which implies

$$\begin{aligned}
E(||\mathbf{m} - \mathbf{w}_0||_2^2) &\leq 2||(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})^{-1}\mathbf{A}\mathbf{w}_0||_2^2 + \\
&2E(||\tau(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^\top\epsilon||_2^2) \\
&\leq 2||(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})^{-1}||_2^2||\mathbf{A}\mathbf{w}_0||_2^2 + \\
&2\tau^2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})E(\epsilon^\top\mathbf{X}^\top\mathbf{X}\epsilon) \\
&\leq 2\mu_0^2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})||\mathbf{w}_0||_2^2 + \\
&2\tau^2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A})E(\epsilon^\top\mathbf{X}^\top\mathbf{X}\epsilon) \\
&\leq 2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{A}) \\
&(\mu_0^2||\mathbf{w}_0||_2^2 + \tau Tr(\mathbf{X}^\top\mathbf{X}))
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X}+\mathbf{A}) \\
&(\mu_0^2||\mathbf{w}_0||_2^2 + p\lambda_{max}(\mathbf{X}^\top\mathbf{X})\tau) \\
&\leq \frac{2\mu_0^2||\mathbf{w}_0||_2^2 + 2\tau pnc_2}{(\tau nc_1)^2} \\
&\leq \frac{2n^{-1}\mu_0^2||\mathbf{w}_0||_2^2 + 2\tau pc_2}{\tau^2 nc_1^2} \\
&\leq \frac{2n^{-1}\mu_0^2||\mathbf{w}_0||_2^2 + 2\tau_{max} pc_2}{\tau_{min}^2 nc_1^2}
\end{aligned}
$$

Now by using Markov inequality

$$
\begin{aligned}
P(||\mathbf{m}-\mathbf{w}_0||_2^2 > \xi_n) &\leq \frac{2n^{-1}\mu_0^2||\mathbf{w}_0||_2^2 + 2\tau_{max} pc_2}{\tau_{min}^2 nc_1^2\xi_n} \\
&\leq \frac{2\mu_0^2||\mathbf{w}_0||_2^2 + 2\tau_{max} pc_2}{\tau_{min}^2 nc_1^2\xi_n}
\end{aligned}
$$

for $n > 1$. Let $c_0 = \frac{2\mu_0^2||\mathbf{w}_0||_2^2 + 2\tau_{max} pc_2}{\tau_{min}^2 c_1^2}$. Then,

$$
P(||\mathbf{m}-\mathbf{w}_0||^2 > \xi_n) \leq c_0 \exp\{-\log(n\xi_n)\}
$$

If $\xi_n \propto n^{-\alpha^*}$, then $n\xi_n = n^{1-(\alpha^*)}$. Then under the condition $0 < \alpha^* < 1$, the term $n^{1-(\alpha^*)} \to \infty$ as $n \to \infty$. Therefore if $0 < \alpha^* < 1$, $P(||\mathbf{m}-\mathbf{w}_0||^2 > \xi_n)$ will approach $0$, which completes the proof.

### 2.3.2 Classification

Unlike regression, there are no closed form variational updates for Classification. Due to the logistic function (2.2), variational distribution $Q(\mathbf{w})$ can not be computed in a straight forward manner. Therefore, in order to make variational approximation tractable, we employ a lower bound on the logistic function proposed by [70] and replace the logistic function with this approximate expression in the joint distribution:

$$
\begin{aligned}
&\sigma(y)^t\big(1-\sigma(y)\big)^{1-t} \\
&\geq \sigma(\xi)\exp\Big(\frac{(2t-1)y-\xi}{2} - f(\xi)\big((2t-1)^2y^2-\xi^2\big)\Big)
\end{aligned} \tag{2.25}
$$

where $f(\mathbf{x}) = \frac{1}{4\xi}\tanh(\xi/2)$, and $\xi$ is a variational parameter. The approximate expression becomes equal to the exact expression when $\xi = (2t-1)y$. From the expression of the lower bound (2.25), it can be seen that it is quadratic in $y$ once the logarithmic transformation is applied. Hence, variational inference becomes tractable as the expressions resembles the Gaussian form.

The update equations for the classification model are almost similar to the regression case except for minor changes in the update of $Q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma})$:

$$\mathbf{\Sigma} = \left(\mathbf{A} + 2\sum\nolimits_i f(\xi_i)\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} \quad \mathbf{m} = \frac{1}{2}\mathbf{\Sigma}\mathbf{X}^\top(2\mathbf{t} - \mathbf{1}) \tag{2.26}$$

where $\mathbf{A}$ is the same as in the regression.

Additionally, maximization of the lower bound (2.25) allows an updating procedure for the parameter $\xi_i$:

$$\xi_i^2 = \mathbf{x}_i^\top \langle \mathbf{w}\mathbf{w}^\top \rangle \mathbf{x}_i. \tag{2.27}$$

## 2.4 Related work

Despite their success in many applications, previous sparse learning methods are limited by several factors for the integration of correlation structure information into the learning framework. For example, group lasso [71] can be used to utilize memberships of features in groups via a $l_{1/2}$ norm to select relevant groups of features, but they ignore structural information about the group. Additionally, they select all the features of the selected groups leading to dense estimation. NaNOS, on the other hand, incorporates correlation structure information through the generalized spike-and-slab prior, and avoids dense estimation of the selected groups due to its two layered sparsity structure accomplished through the hybrid combination of its conditional and generative components. Adaptive group lasso [72] extends the group lasso by assigning different weights to different groups, but it still can not avoid the dense estimation problem. An excellent work by [56] overcomes the limitation of ignoring structural information by incorporating group structures in a Laplacian

matrix of a global graph to guide the selection of relevant features. In addition to graph Laplacians, binary Markov random field priors can be used to represent correlation structure information to influence feature selection [57, 73, 74, 75]. However, unlike NaNOS, these network-regularized approaches do not explicitly select relevant groups. However, not all groups are relevant and group selection can yield insight into underlying generative processes. For example, in genomic data applications, there is a need to determine relevant pathways for disease diagnosis and prognosis. A pioneering approach to joint group and feature selection by [76] uses binary Markov random field priors and couples feature and group selection by hard constraints – for example, if a feature is selected, all the groups it belongs to will be selected. However, this consistency constraint might be too rigid for certain applications: an active gene for cancer progression does not necessarily imply that *all* the pathways it belongs to are active. NaNOS overcomes this constraint by duplicating the weights of the features appearing in multiple groups, and treating each weight as a separate model parameter. Given the Markov random field priors and the nonlinear constraints, posterior distributions are inferred by a Markov Chain Monte Carlo method [76]. But the convergence of MCMC for high dimensional problems is known to take a long time. NaNOS, on the other hand, employs variational inference approach that converges much faster than MCMC methods.

## 2.5 Experiments

For the purposes of evaluating the performance of NaNOS both in terms of predictive power and selection accuracy, we conducted thorough synthetic and real data experiments. We specifically focused on genomic data applications. We simulated gene expression data sets consisting of pathways and genes, and examined the quality of results generated by NaNOS on these datasets. We also tested NaNOS on real gene expression data sets, and analyzed the results. In order to demonstrate the superior performance of NaNOS, we compared its results with other alternative methods such as lasso [77], elastic net [78], group lasso [71, 79], the method proposed by [56] and denoted as“LL”, and the classical spike-

and-slab prior [67]. We employed the Glmnet software package [1] to generate results for lasso and elastic net. In case of group lasso, all the results were generated using the SLEP software package [2]. Since group lasso does not handle over lapping groups, we applied the duplication operator [79] on the expression level of genes appearing in multiple pathways or groups before feeding the data into SLEP package. The classical spike-and-slab model was implemented in a way similar to NaNOS. We used the same variational inference strategy as in NaNOS to generate prediction and selection results. We did not change the default configurations of all the software packages, and used 10-fold cross validation to tune all free parameters. We give a brief summary of all the CV grids we used for various methods: (1) lasso: $\alpha = [0 : 0.01 : 1]$; (2)elastic net: $\alpha = [0 : 0.01 : 1]$ and $\beta = [0 : 0.01 : 1]$; (3) group lasso: $\alpha = [0 : 0.01 : 1]$;(4) LL: $\lambda_1 = [1 : 25 : 300]$ and $\lambda_2 = [1 : 25 : 300]$; (5) NaNOS: $s_1 = r_1 = [0.1, 1, 3]$ and $s_2 = r_2 = [10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$.

We also compared NaNOS with GSEA [80, 81]. GSEA is a popular method for extracting relevant gene sets. For applying GSEA on our synthetic data sets, we treated each pathway as a gene set, and selected all pathways with FDR $< 25\%$. This is the same criteria that GSEA uses for gene set selection. We also did not change the default settings in the GSEA package. Once the gene sets were extracted, we assumed all the genes in these sets to be relevant. Since GSEA can not perform prediction, and we do not know the true relevant pathways in real data sets, we did not use it for real data analysis.

### 2.5.1 Simulation studies

We conducted the following three simulation experiments in order to compare all the methods on synthetic data.

**Experiment 1.** We followed the approach proposed by [56] to conduct these experiments. First we construct 200 tree-structured regulatory networks consisting of a transcription factor (TF) and 10 other genes controlled and regulated by it. Out of these 200 pathways, only 4 – including *all* of their genes – are assumed to have an effect on the response $\mathbf{t}$.

[1] `www-stat.stanford.edu/~tibs/glmnet-matlab/`
[2] `www.public.asu.edu/~jye02/Software/SLEP/`

The expression levels of each TF ($x_{TF}$) and the corresponding regulated genes are sampled from $\mathcal{N}(0,1)$ and $\mathcal{N}(0.7x_{TF}, 0.51)$ respectively. This sampling procedure establishes a correlation of $0.7$ between the TF and its regulated genes.

For the first data generation model, in a regression setting, the weight vector for each pathway is given by $\boldsymbol{\rho} = [1, \frac{1}{\sqrt{10}}, \cdots, \frac{1}{\sqrt{10}}]$; the first component of $\boldsymbol{\rho}$ corresponds to the TF and remaining components are the weight values for the 10 regulated genes. The outcome $\mathbf{t}$ is then sampled as follows:

$$\begin{aligned} \mathbf{w} &= [5\boldsymbol{\rho}, -5\boldsymbol{\rho}, 3\boldsymbol{\rho}, -3\boldsymbol{\rho}, \mathbf{0}^\top]^\top \\ \mathbf{t} &= \mathbf{X}\mathbf{w} + \epsilon \end{aligned} \tag{2.28}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$.

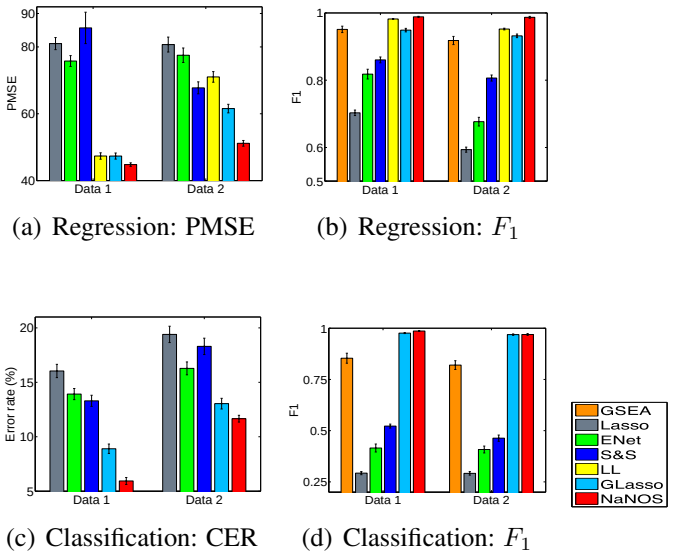
For the second data generation model, the only difference from the first model is that the regulated genes of the same TF can have both positive and negative influence on $\mathbf{t}$ [56]. Specifically, the weight vector for each pathway is now given by

$$\boldsymbol{\rho} = [1, \frac{-1}{\sqrt{10}}, \frac{-1}{\sqrt{10}}, \frac{-1}{\sqrt{10}}, \underbrace{\frac{1}{\sqrt{10}}, \cdots, \frac{1}{\sqrt{10}}}_{7}]. \tag{2.29}$$

For the classification case, the procedure for generating $\mathbf{X}$ and $\mathbf{w}$ remains the same. Once these are generated, the outcome $\mathbf{t}$ is sampled from (2.2).

For both data generating models, we conducted 50 simulation, and in each experiment we simulated 100 training and 100 test samples. To compare the prediction accuracy of competing methods, we computed the prediction mean-squared error (PMSE) [56] for regression, and the error rate for classification. We also computed sensitivity, specificity and $F_1$ score to examine the gene and pathway selection capability of all methods. $F_1$ score is defined as the harmonic average of the sensitivity and specificity, and is given by $F_1 = 2\ (\text{sensitivity} \times \text{specificity})/(\text{sensitivity} + \text{specificity})$. Therefore higher values of the $F_1$ score indicate more accurate selection results.

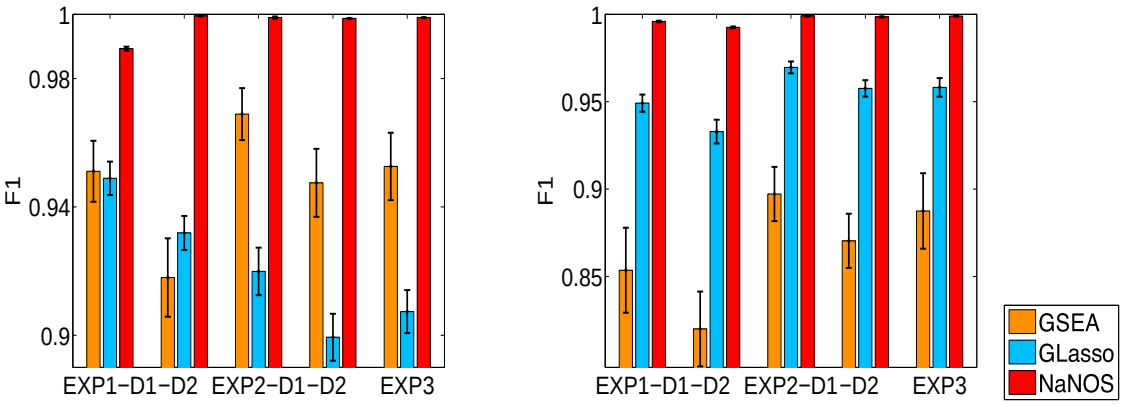
Figure 2.2 presents all the results, error bars indicate the standard errors. Apart from the classification case in the second data model where NaNOS and group lasso achieve comparable $F_1$ scores, NaNOS outperforms alternative methods, both in terms of predic-




(a) Regression: PMSE (b) Regression: $F_1$

(c) Classification: CER (d) Classification: $F_1$

Figure 2.2.: Prediction errors and $F_1$ scores for gene selection in Experiment 1. ENet, S&S, and GLasso stand for elastic net, the spike-and-slab model, and group lasso, respectively; and Data 1 and 2 indicate the first and second data generation models. CER stands for classification error rate.




(a) Regression: $F_1$ (b) Classification: $F_1$

Figure 2.3.: $F_1$ scores for pathway selection. "EXP" stands for "Experiment" and "D" stands for "Data model".

tion accuracy and selection results. We also performed a two-sample t-test, using 5 percent significance level, to determine whether the improvements achieved by NaNOS were significant or not. It was clear from the results that all improvements were significant. We also

compare the pathways selection accuracy of group lasso and GSEA with our model. Figure 2.3 shows the pathways selection accuracy plots. As can be seen from the plots, NaNOS significantly outperforms the other two methods. The improved performance of NaNOS can be attributed to its modelling power. By applying sparse prior over each pathway, NaNOS induces sparsity at the group level, and can explicitly select relevant pathways. Other methods, for example the LL approach, do not have this leverage. Despite the fact that LL uses the topological information, it does not generate sparsity at the group level. It treats the whole network as one big global structure, and extracts important sub networks from the global network. This approach is more suitable for discovering new pathway structure, but less helpful for determining the relevance of already existing pathways to the outcome.

**Experiment 2.** Under the settings of experiment 2, we do not assume all genes in relevant pathways to be influential in the outcome. Secondly, we simulate expression levels of 100 transcription factors (TFs), each TF now regulates 21 genes to form a tree like network. Expression levels are sampled in the same way as Experiment 1, except for some minor change in $\boldsymbol{\rho}$. The expression for $\boldsymbol{\rho}$ is given by

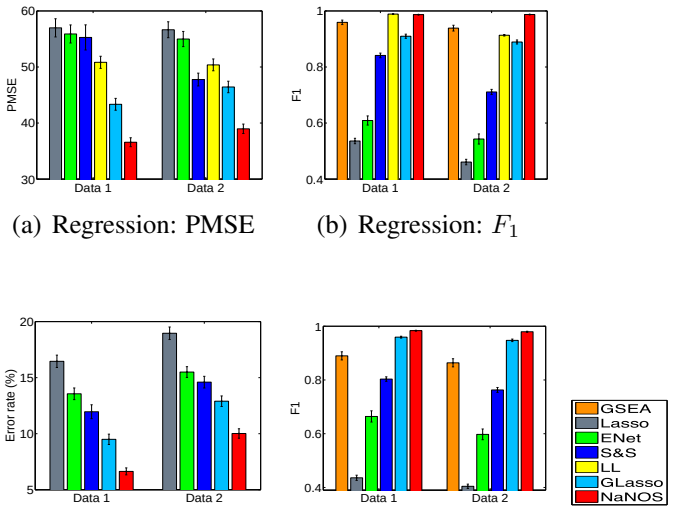
$$\boldsymbol{\rho} = [1, \underbrace{\frac{1}{\sqrt{21}}, \cdots, \frac{1}{\sqrt{21}}}_{10}, \underbrace{0, \ldots, 0}_{11}] \tag{2.30}$$

for the first data generation model and

$$\boldsymbol{\rho} = [1, \frac{-1}{\sqrt{21}}, \frac{-1}{\sqrt{21}}, \frac{-1}{\sqrt{21}}, \underbrace{\frac{1}{\sqrt{21}}, \cdots, \frac{1}{\sqrt{21}}}_{7}, \underbrace{0, \ldots, 0}_{11}] \tag{2.31}$$

for the second data generation model.

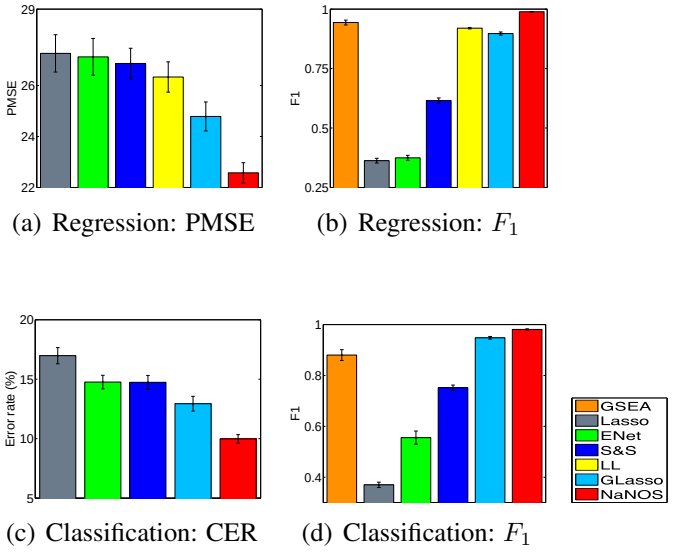
The results for both classification and regression are shown in Figures 2.4 and 2.3. Apart from the regression case in the first data generation model where NaNOS and LL show comparable performance in terms of $F_1$ score, in all other settings, NaNOS shows superior performance compared to other methods, improvement is again tested at the 5 percent significance level.




(a) Regression: PMSE (b) Regression: $F_1$

(c) Classification: CER (d) Classification: $F_1$

Figure 2.4.: Prediction errors and $F_1$ scores for gene selection in Experiment 2.




(a) Regression: PMSE (b) Regression: $F_1$

(c) Classification: CER (d) Classification: $F_1$

Figure 2.5.: Prediction errors and $F_1$ scores for gene selection in Experiment 3.

**Experiment 3.** In this experiment, the only change in the data generating process is in the expression of $\boldsymbol{\rho}$. The change reflects a weaker influence of TF on its regulated genes. As the Figures 2.3 and 2.5 demonstrate, NaNOS shows superior performance to alternative methods.

### 2.5.2 Application to gene expression data

In order to demonstrate the effectiveness of our model on real data, we analyze three cancer related gene expression data sets: diffuse large B cell lymphoma (DLBCL) [82], colorectal cancer (CRC) [83], and pancreatic ductal adenocarcinoma (PDAC) [84]. We employed the probeset-to-gene mapping provided in cancer studies of these data sets. In case we had expression levels of multiple probes corresponding to the same gene, we took the average expression levels of these probes. We followed this approach for the CRC and PDAC datasets in which multiple probes were mapped to the same genes. Information about the pathways was collected from the KEGG pathway database (`www.genome.jp/kegg/pathway.html`).

**Diffuse large B cell lymphoma (DLBCL)[82].** We collected gene expression profiles of 240 DLBCL patients from [82]. KEGG dataset provided 752 genes and 46 pathways for the gene expression data set. The data set also provides the survival time information about all the patients. We used the logarithm of survival times of patients as the target variable.

Out of the total of 240 samples, half of them were randomly chosen for training, and other half for testing. We performed this splitting 100 times and recorded the results of all methods in each case. 2.6.a shows the average test results over 100 runs for all method. Superior performance of NaNOS is quite evident from the figure. An obvious advantage of NaNOS when compared to LL approach is that while NaNOS explicitly selects relevant pathways, LL method extracts connected sub-networks. These sub-networks may or may not correspond to biological pathways. It is highly likely that they may consist of portions from multiple overlapping pathways. Based on the results generated by NaNOS, the top two pathways in terms of the frequency of selection across all the runs, and having selection posterior probabilities larger than 0.95 were (1) antigen processing and presentation pathway, and (2) cell adhesion molecules (CAMs). Existing literature supports the relevance of these pathways to the growth of Diffuse large B cell lymphoma [82, 85, 86, 87, 88, 89].

**Colorectal cancer (CRC).** The colorectal cancer dataset [83] contains gene expression profiles from 22 normal patients and 25 cancer patients. 2455 genes from 22,283 probes

were mapped into 67 KEGG pathways. The aim of this analysis was to predict the medical condition of the tissues: whether a tissue is cancerous or not, and select pathways and genes relevant to the cancerous phenotype.

We randomly selected 23 samples for training, and remaining 24 for testing. We performed this splitting 50 times and recorded the results of all methods in each case. 2.6.b shows the average test results over 50 runs for all method. Again, based on the 5 percent significance level, NaNOS performs superior to other methods. The top three pathways that were most consistently selected by NaNOS, with the selection posterior probabilities larger than 0.95, were (1)cell cycle pathway,(2) the intestinal immune network for IgA production, and (3) cytokine-cytokine receptor interaction pathway. All these pathways are well recognized for CRC. NaNOS also selected relevant genes within these pathways. For cell cycle pathway, the selected genes were: Bub1, Mad1, Mad2, BubR1, Bub3, CycD/CDK4, CDK1, CDK2, CycE, MCM2, MCM5, TP53, c-Myc; for the second pathway, the corresponding selected genes were: CXCR4. and CXCL12; and for the third one: CXCL13, CXCL10, and IL10. All these genes and pathways are supported by published literature [83, 90, 90, 91, 92, 93, 94, 95, 96, 97, 98].

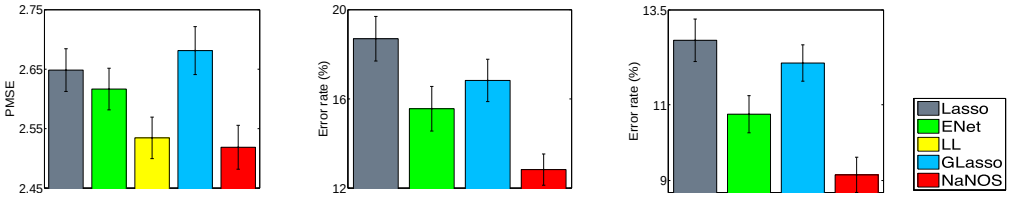
**Pancreatic ductal adenocarcinoma (PDAC).** The data set contains gene expression profiles from 39 normal patients and 39 cancer patients. 2781 genes from 54677 probes were mapped into 67 KEGG pathways. The aim of this analysis was to predict whether a tissue has the pancreatic cancer or not, and select pathways and genes relevant to the pancreatic cancer phenotype. We randomly divided the dataset into two equal parts, one for training and the other for testing. We performed this splitting 50 time, and recorded the output generated by all methods in each case. The average test results are shown in Figure 2.6.c. Again, based on a 5 percent significance level, NaNOS shows significant improvement over other competing methods.

The pathways and genes selected by NaNOS for the PDAC data set are mentioned below:

The first selected pathway was the TGF-$\beta$ signaling pathway. The associated related genes were: IFNG, TNF-$\alpha$, LTBP1, DCN, TGF-$\beta$, TGF-$\beta$ R1, Smad 4,EMT, BMP2. The

second identified pathway was extracellular matrix (ECM)-receptor interaction. In this pathway, NaNOS selected ITGB1, ITGA2, ITGA3, ITGA5, ITGA6 ,COL1A1, COL1A2, LAMC2 and LAMB3. The third chosen pathway was CAMs. In this pathway, the selected molecules include CDH2, CDH3, and neural-related molecules (MAG). The relevance of these pathways and molecules to PDAC can be confirmed from published literature [99, 100, 101, 102, 103, 104].

All the above discussed pathways and genes are shown in Figures 2.7 a b and c.




(a) DLBCL (b) CRC (c) PDAC

Figure 2.6.: Predictive performance on three gene expression studies of cancer.

To demonstrate the robustness of our model to structural noise in pathway database, we randomly removed 20%, 50%, 80% and 100% edges in each pathway and applied NaNOS in each case. The average test error is reported in Figure 2.8. Consistent with our intuition,

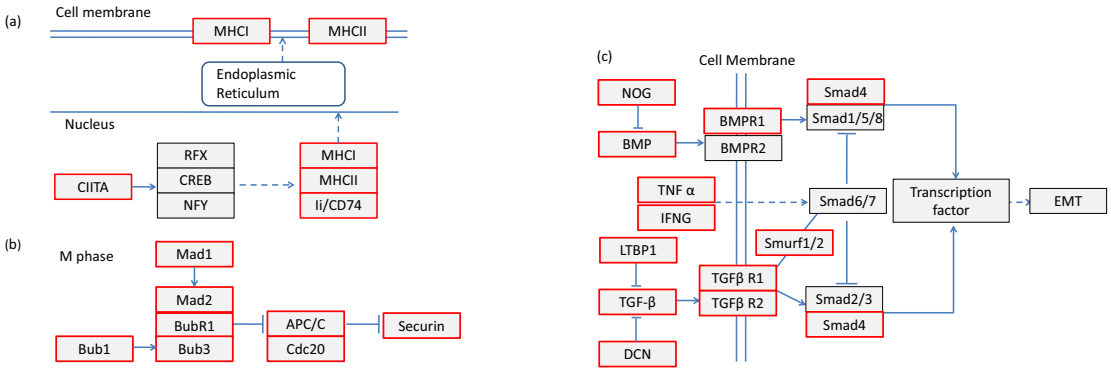



Figure 2.7.: Examples of part of identified pathways. (a): the antigen processing and presentation pathway for DLBCL; (b): the cell cycle pathway for CRC; (c): the TGF-$\beta$ signaling pathway for PDAC. Red and black boxes indicate selected and not selected genes, respectively.

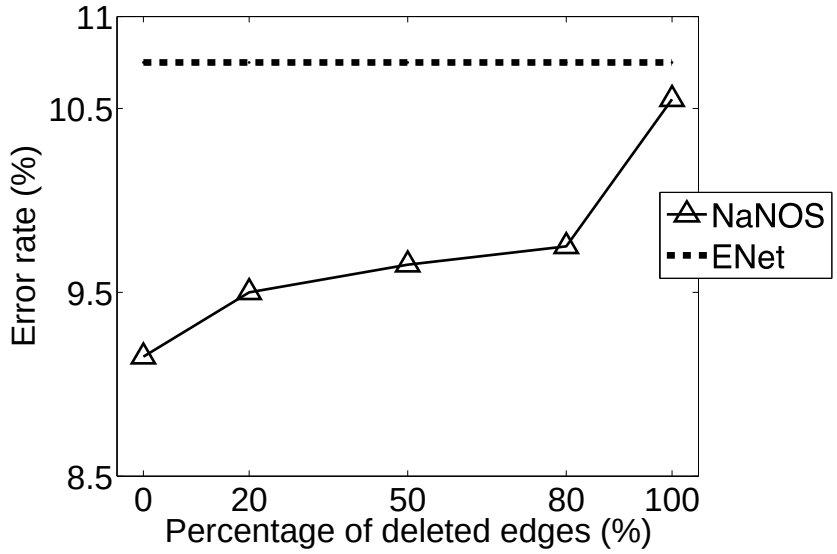



Figure 2.8.: The predictive performance of NaNOS when the pathway structures are inaccurate. When more edges are randomly selected and removed from each pathway, the performance of NaNOS degrades smoothly, but still better than the competing methods.

test error rate increases as more edges are removed from the network. The drop in performance is due to the loss of topological information contained in the network structure. However, despite this loss of information, NaNOS consistently outperforms all the other alternative methods including elastic net, the second best method on this dataset. These observations lead to the following conclusions: (1) NaNOS can enhance its modelling capability and predictive power by incorporating pathway topology information. (2) NaNOS is robust to small changes in network topology.

In order to demonstrate the robustness of NaNOS to the choice of prior distributions on pathway and gene selection probabilities $u_k$ and $v_{kj}$, we examined the performance of NaNOS over a wide range of prior choices. To cover the whole spectrum of prior choices we specifically tested NaNOS on a highly sparse prior: Beta(1,10) (mean $0.09$ and standard deviation $0.083$); highly dense prior: Beta(10,1) (mean $0.91$ and standard deviation $0.083$); and an uninformative or weak prior: (e.g., Beta(0.5,0.5)). The average test error based on the uninformative prior is $9.15 \pm 0.5$ (Figure 2.6.c). If we replace this prior with the sparsity favoring prior, the test error rate slightly jumps to $10.0 \pm 0.4$. Next, if we use a dense prior

that favors dense estimation, then the average test error increases to $11.2 \pm 0.5$. In the first case, the decline in performance is due to over sparsification of the results, and in the second case, dense estimator tends to select all pathways and genes which is obviously wrong. However, in both cases, NaNOS outperforms other methods as shown in 2.6.c The above two cases correspond to the extremes of the spectrum of prior choices. If we use an uninformative or weak sparse prior that lies somewhere in the middle of this spectrum, NaNOS generates prediction error rates very close to that in 2.6.c. The above examination leads to the conclusion that NaNOS is robust to the change in prior distributions.

# 3 FAST LAPLACE APPROXIMATION FOR SPARSE BAYESIAN SPIKE-AND-SLAB MODELS

## 3.1 Motivation

As an intersection of machine learning, statistics, and signal processing, sparse modeling has numerous applications. For developing various sparse models, $L_1$ regularization has played a central role. $L_1$-type methods not only enjoy provable properties relating to the estimation optimality and oracle properties [10, 105], but also have the convenience of using well-developed computational tools from convex optimization to obtain sparse solutions. As a result, they have been widely used in many applications including feature selection, compress sensing[106], multi task learning [107], and time-varying network reconstruction [108].

Despite the popularity of $L_1$ regularization, [109] examined the performance of $L_1$-type methods and compared them with Bayesian spike-and-slab methods [63], which are relatively under used in the machine learning community. [109] revealed the improved performance of the spike-and-slab methods over the $L_1$-type methods—in unsupervised settings. This improvement probably stems from a selective shrinkage property of the spike-and-slab prior [59]. Unlike the $L_1$ penalization, (i.e., equivalently, the Laplace prior) which shrinks all features—regardless of relevance or not—in the same way, the spike-and-slab prior is a mixture of two components: one component regularizes relevant variables mildly while the other one shrinks irrelevant variables aggressively (Section 2). Furthermore, the spike-and-slab method has the advantage of uncertainty quantification in feature selection which is not possible with $L_1$ methods

In this chapter, we examine the performance of the Bayesian spike-and-slab models for very high dimensional problems in the supervised learning setting. For very high dimensional problems, existing Monte Carlo methods [59] converge slowly with tens of thou-

sands of features in data; and the variational Bayes (VB) and expectation propagation (EP) approaches [60, 60, 61] either need a fully factorized approximation to obtain a linear cost but at the price of a reduced approximation quality, or have a quadratic cost, making them impractical for large data. By contrast, the frequentist $L_1$-type methods have fast solvers developed over years, making them a practical tool. To address the computational issue associated with the spike-and-slab model, we develop the Fast Laplace Approximation for Spike-and-slab model. Our approach not only maintains the benefits of the Bayesian treatment (e.g. uncertainty quantification) but also possesses the computational efficiency, and oracle properties of the frequentist methods.

Specifically, we apply the Laplace approximation to the marginal posterior distribution of each weight parameter. For the Laplace approximation we need to obtain the mode of the posterior distribution. To this end, we exploit two efficient optimization methods, the recently developed GIST method [64] and the popular limited-memory BFGS (L-BFGS) [65]. First, we present a MAP estimation procedure based on L-BFGS [65] method, denoted by FLAS. Second, we present two approaches for joint MAP estimation of model weights and selection probabilities based on GIST [64] method. The first joint optimization approach employs an alternating optimization strategy, denoted by FLAS*, with convergence guarantees for both regression and classification, and oracle properties for regression model. In each iteration of the alternating optimization procedure, the model weights are optimized through the GIST method. The second joint approach, denoted by FLAS**, performs a direct optimization on the joint space of model weights and selection probabilities, again using the GIST method. Then we propose an ensemble Nyström approach to calculate the diagonal of the inverse Hessian over the mode to obtain the approximate posterior marginals in $O(knp)$ time, where $n$ and $p$ are the numbers of samples and features respectively, and $k << p$. The theoretical analysis of the ensemble method is also provided. With the posterior marginals of model weights, we use quadrature integration to estimate the marginal posteriors of selection probabilities and indicator variables for all features, which quantify the selection uncertainty. While a factorized joint posterior assumption is

usually not true, VB and EP often adopt it for computational efficiency. By contrast, our method is free of this assumption, but still enjoys a linear cost in $p$.

On simulated data, our methods perform feature selection better than or comparable to the alternative approximate methods, with less running time, and provide higher prediction accuracy than various sparse methods including VB, EP, automatic relevance determination, lasso, elastic net and a capped-$L_1$ method (Section 3.5). On large real benchmark datasets, our methods often achieve improved prediction accuracy compared to alternative methods, but with a convergence time comparable to frequentist $l_1$ methods. Finally, we apply our approach to Region-Of-Interest (ROI) study on brain image data with tens of thousands of features. We find interesting brain regions for face and Chinese character recognition, many of which are supported by existing literature.

## 3.2 Spike-and-slab models

We first present sparse linear models with spike-and-slab priors. Suppose we have $n$ independent and identically distributed samples $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$, where $\mathbf{x}_i$ is the $p$ dimensional feature vector of the $i$-th sample, and $t_i$ is its response. We aim at predicting the response vector $\mathbf{t} = [t_1, \ldots, t_n]^\top$ based on the feature set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ and selecting a small number of features relevant to the prediction. For real-world applications, we often have $n \ll p$.

For regression, the Gaussian data likelihood is used:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \tau) = \prod_{i=1}^{n} \mathcal{N}(t_i|\mathbf{x}_i^\top \mathbf{w}, \tau^{-1}) \tag{3.1}$$

where $\mathbf{w}$ are regression weights, and $\tau$ is the precision parameter.

For classification, the logistic likelihood is used:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \sigma(\mathbf{x}_i^\top \mathbf{w})^{t_i}[1 - \sigma(\mathbf{x}_i^\top \mathbf{w})]^{1-t_i} \tag{3.2}$$

where $t_i \in \{0, 1\}$, $\mathbf{w}$ are classifier weights, and $\sigma(a) = 1/(1 + \exp(-a))$.

A set of latent binary variables $\{z_j\}$ are introduced to indicate the feature selection: $z_j = 1$ means the $j$-th feature is selected; otherwise, it is not. Then a spike-and-slab prior [59] over $\mathbf{w}$ is assigned:

$$p(\mathbf{w}|\mathbf{z}) = \prod_{j=1}^{p} \mathcal{N}(w_j|0, r_0)^{(1-z_j)} \mathcal{N}(w_j|0, r_1)^{z_j}, \tag{3.3}$$

$$p(z_j = 1|s_j) = s_j \quad (1 \leq j \leq p) \tag{3.4}$$

where $r_0$ and $r_1$ are the variances of the two Gaussian components and $s_j \in [0, 1]$ represents the selection probability for the $j$-feature. We set $r_1 \gg r_0$ so that if the $j$-th feature is selected, the prior over $w_j$ has a large variance $r_1$ (as a regular $L_2$ penalty in the frequentist framework) and, if not, the zero-mean prior has a very small variance $r_0$, leading to aggressive shrinkage of the irrelevant feature. We further assign a Beta prior over $s_j$: $p(s_j) = \text{Beta}(a_0, b_0)$. In the experiments, we set $a_0 = b_0 = 1$ such that this prior is an uninformative uniform prior.

## 3.3 Algorithm

Given high dimensional data, current inference methods such as Gibbs sampling or VB can suffer from high computational cost. To overcome the computational bottleneck, we use Laplace's method to approximate the posteriors of each $\{w_j\}$ and apply the quadrature integration [110] to estimate the selection probability $s_j$ and indicator variable $z_j$.

### 3.3.1 Laplace approximation

To obtain the Laplace approximation, we need to compute the mode and the second-order derivative of the log posterior distribution at the mode. We describe two approaches for computing MAP estimation: marginalized MAP estimation ,and joint MAP estimation. Details of the two approaches are described below.

L-BFGS optimization of the marginalized model

For the FLAS method, we marginalize out both $\mathbf{z}$ and $\mathbf{s}$. The negative log probability of the marginalized model is then given by

$$\mathcal{F}(\mathbf{w}) = L(\mathbf{w}) - \sum_{j=1}^{p} \log\Big(\frac{1}{2}\mathcal{N}(w_j|0, r_1) + \frac{1}{2}\mathcal{N}(w_j|0, r_0)\Big), \tag{3.5}$$

where $L(\mathbf{w})$ is the negative log likelihood for regression or classification. To minimize the negative log probability, we use the L-BFGS method [65] because of its low computational and memory cost. As a quasi-Newton method, the L-BFGS method uses last $M$ function/gradient pairs to approximate the inverse Hessian matrix of the parameters $\mathbf{w}$. Because $M$ is set to be much smaller than $p$, often as small as 3-10, the computational cost is linear in $p$.

To use L-BFGS, we need to compute the gradient over $\mathbf{w}$:

$$\Big[\frac{\mathrm{d}\mathcal{F}}{\mathrm{d}\mathbf{w}}\Big]_j = \Big[\frac{\mathrm{d}L(\mathbf{w})}{\mathrm{d}\mathbf{w}}\Big]_j + \frac{r_0 + r_1 g(w_j)}{r_0 r_1 (1 + g(w_j))} w_j \tag{3.6}$$

where $g(w_j) = \sqrt{\frac{r_1}{r_0}} \exp(\frac{1}{2}(\frac{1}{r_1} - \frac{1}{r_0})w_j^2)$, and $\frac{\mathrm{d}L(\mathbf{w})}{\mathrm{d}\mathbf{w}} = \tau \mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{t})$, for regression and $\frac{\mathrm{d}L(\mathbf{w})}{\mathrm{d}\mathbf{w}} = \sum_{n=1}^{N} \big(\frac{t_n}{1+\exp(\mathbf{x}_n^\top \mathbf{w})} - \frac{1-t_n}{1+\exp(-\mathbf{x}_n^\top \mathbf{w})}\big)\mathbf{x}_n$, for classification.

Using the gradient in the L-BFGS method, we can compute the mode of $w_j$ efficiently. Then we can approximate the posteriors of $s_j$ and $z_j$ as explained in Section 3.3.3.

Optimization of the joint model

For the joint MAP estimation approach we only marginalize out $\mathbf{z}$ and jointly optimize over the weights $\mathbf{w}$ and the selection probability $\mathbf{s}$. From a Bayesian perspective, we prefer the FLAS approach because by marginalizing out $\mathbf{s}$, it essentially takes all possible values of $\mathbf{s}$ into account. However, the joint estimation approach can provide a more pronounced selective shrinkage effect than the first approach. We first describe the alternating optimization (AO) procedure(FLAS*) for joint MAP estimation. We use the (AO) approach for both regression and classification, and employ GIST [64], which converges to a local

optimum with a cost per iteration linear in $n$ and $p$, for finding the minimizer of $\mathbf{w}$ during the AO iterations. This alternating scheme is guaranteed to converge to a joint local minimum. Additionally, the alternating optimization approach for regression leads to nice oracle properties for the $\mathbf{w}$ estimator: estimation, sign, and selection consistency.

In joint optimization, we minimize the negative log joint probability:

$$\min_{\mathbf{w},\mathbf{s}} \mathcal{F}(\mathbf{w},\mathbf{s}) = \min_{\mathbf{w}} L(\mathbf{w}) - \min_{\mathbf{s}} R(\mathbf{w},\mathbf{s}) \tag{3.7}$$

where

$$R(\mathbf{w},\mathbf{s}) = \sum\nolimits_{j=1}^{p} R_j(w_j, s_j) \tag{3.8}$$

$$R_j(w_j, s_j) = \log\left(s_j \mathcal{N}(w_j|0, r_1) + (1 - s_j)\mathcal{N}(w_j|0, r_0)\right) \tag{3.9}$$

We perform alternating optimization by keeping one variable fixed, and optimize over the other. We start the optimization procedure by randomly initializing $\mathbf{w}$. Given $\mathbf{w}$ as fixed, $\mathcal{F}(\mathbf{w},\mathbf{s})$ is a monotone function of each $s_j$, hence it attains minimum either at $s_j = 1$ or $s_j = 0$. The update of $s_j$ is given by:

$$s_j = \begin{cases} 1 & \text{if } |w_j| \geq a \\ 0 & \text{if } |w_j| < a \end{cases} \tag{3.10}$$

where $a = \sqrt{\left(\frac{2r_0 r_1}{r_1 - r_0}\right) \log \sqrt{\frac{r_1}{r_0}}}$. Given $\mathbf{s}$, the optimization of $\mathbf{w}$ has a closed form solution for regression that is a special case of generalized ridge regression [111]:

$$\mathbf{w}_{opt} = (\tau \mathbf{X}^\top \mathbf{X} + \mathrm{diag}(\mathrm{d}))^{-1} \tau \mathbf{X}^\top \mathbf{t} \tag{3.11}$$

where $\mathrm{d}$ is such that $d_j = (\frac{1}{r_1})^{s_j} (\frac{1}{r_0})^{1-s_j}$.

As can be seen from the above equations, the update of $\mathbf{w}$ requires the inversion of $p$ by $p$ matrix which has a complexity of $O(p^3)$. This is prohibitively expensive at higher dimensions. Therefore, instead of directly using the closed form solution, we employ GIST for minimizing $\mathbf{w}$. Since the function to be optimized is strictly convex, GIST is guaranteed to converge to the unique minimum (closed form solution), but with cost per iteration linear in $n$ and $p$ [64]. In case of classification, we do not have have a closed form update for $\mathbf{w}$, but

with the logistic loss function the optimization problem is still strictly convex, hence GIST again converges to the unique minimum. Below we briefly describe the GIST updates:

GIST iteratively minimizes $\mathcal{F}(\mathbf{w}, \mathbf{s}^l)$ with respect to $\mathbf{w}$, where $l$ is the index for the AO iterations, using the following step [64]:

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\, L(\mathbf{w}^{(k)}) + \langle \nabla L(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{\rho^k}{2}||\mathbf{w} - \mathbf{w}^{(k)}||^2 + R(\mathbf{w}, \mathbf{s}^l). \tag{3.12}$$

where $\mathbf{w}^{(k)}$ is the value of $\mathbf{w}$ at step $k$, and $\rho^{(k)}$ is the stepsize at step $k$. Due to the form of $R(\mathbf{w}, \mathbf{s}^l)$, the minimization problem can be further cast into $p$ independent univariate proximal operator problems [64]:

$$w_j^{(k+1)} = \underset{w_j}{\operatorname{argmin}} \frac{1}{2}(w_j - u_j^{(k)})^2 + \frac{1}{\rho^{(k)}} R(w_j, s_j^l) \tag{3.13}$$

where $j = 1, \ldots, p$, and $u_j^{(k)} = w_j^{(k)} - \nabla L(w_j^{(k)})/\rho^{(k)}$. To solve the univariate optimization problem, we calculate the value of $w_j$ for the following two cases. For the first case, $s_j = 1$, the function has its minimal at $w_j^{(k+1)} = b_1$, where $b_1 = \frac{u_j^{(k)}}{1+1/(r_1\rho^{(k)})}$; for the second case, $s_j = 0$, the function has its minimal at $w_j^{(k+1)} = b_0$, where $b_0 = \frac{u_j^{(k)}}{1+1/(r_0\rho^{(k)})}$.

Next we present the FLAS** approach by directly applying GIST on the joint space of $\mathbf{w}$ and $\mathbf{s}$. By directly exploring the joint space, this approach is expected to perform joint MAP estimate efficiently. For FLAS**, GIST iteratively minimizes (3.7) with respect to $\mathbf{w}$ and $\mathbf{s}$ using the following two steps [64]:

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\; L(\mathbf{w}^{(k)}) + \langle \nabla L(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{\rho^k}{2}||\mathbf{w} - \mathbf{w}^{(k)}||^2 + R(\mathbf{w}, \Phi(\mathbf{w})) \tag{3.14}$$

$$\Phi(\mathbf{w}) = \underset{\mathbf{s}}{\operatorname{argmin}}\, R(\mathbf{w}, \mathbf{s}) \tag{3.15}$$

$$\mathbf{s}^{(k+1)} = \underset{\mathbf{s}}{\operatorname{argmin}}\, R(\mathbf{w}^{(k+1)}, \mathbf{s}) \tag{3.16}$$

As before, the minimization problem can be cast into $p$ independent univariate proximal operator problems [64]:

$$w_j^{(k+1)} = \underset{w_j}{\operatorname{argmin}} \frac{1}{2}(w_j - u_j^{(k)})^2 + \frac{1}{\rho^{(k)}} R(w_j, \phi(w_j)) \tag{3.17}$$

$$\phi(w_j) = \underset{s_j}{\operatorname{argmin}} R(w_j, s_j) \tag{3.18}$$

$$s_j^{(k+1)} = \underset{s_j}{\operatorname{argmin}} R_j(w_j^{(k+1)}, s_j) \tag{3.19}$$

where $j = 1, \ldots, p$, and $u_j^{(k)} = w_j^{(k)} - \nabla L(w_j^{(k)})/\rho^{(k)}$.

We again calculate the value of $w_j$ for two cases: for the first case, $s_j = 1$ and the update of $w_j$ is given by:

$$w_j = \begin{cases} b_1 & \text{if } |b_1| > a \\ \text{sign}(b_1)a & \text{if } |b_1| \leq a \end{cases} \tag{3.20}$$

where $b_1 = \frac{u_j^{(k)}}{1+1/(r_1\rho^k)}$.

For the second case, $s_j = 0$ and the updates for $w_j$ are:

$$w_j = \begin{cases} b_0 & \text{if } |b_0| < a \\ \text{sign}(b_0)a & \text{if } |b_0| \geq a \end{cases} \tag{3.21}$$

where $b_0 = \frac{u_j^{(k)}}{1+1/(r_0\rho^k)}$. Then, comparing the minimum values for these two cases and taking the smaller one, we can easily obtain the new $w_j^{(k+1)}$ and $s_j^{(k+1)}$. Note that, when $w_j = a$, $s_j$ can be either 1 or 0, which gives the same function values.

**Estimation, Selection and Sign consistency for regression:** Similar to the estimation consistency proof given in [69], and using the approach presented in [30] for ridge regression, the above estimator can be shown to have an estimation consistency property: the estimated weight vector approaches the true vector in the $l_2$ norm sense as $n \to \infty$. Let us assume that $\mathbf{w}^*$ is the true coefficient vector of the regression model. Define $S^* = \{j : w_j^* \neq 0\}$, and $S_{opt} = \{j : w_{optj} \neq 0\}$. Let $\mathcal{S}$ denote the space in which $S^*$ lies. Selection consistency implies that $S^* = S_{opt}$, and sign consistency requires $sign(\mathbf{w}^*) = sign(\mathbf{w}_{opt})$, where $sign(a) = 1, 0, -1$ for $a > 0, a = 0, a < 0$ respectively,

sign operator is applied component wise. In addition to using the following assumption, we will use some of the assumptions employed in the previous chapter:

**Assumption 3 [69].** There exist finite constant $c_3 > 0$ such that $(w_j^*)^2 < c_3$ for all $j = 1, , \cdots p$.

**Assumption 4.** For parameters $r_0$, $r_1$, assume that there exist finite positive constants $r_{low}$ and $r_{up}$ such that $r_{low} \leq r_0, r_1 \leq r_{up}$.

**Assumption 5.** Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ be a sample from $p$ dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ with mean $\mathbf{0}$ and unit covariance matrix. Then, for sufficiently large $n$ with $p$ fixed, $\mathbf{X}^\top \mathbf{X} \to n\mathbf{I}_p$. Let $\zeta = \mathbf{X}^\top \epsilon$ then there exist a finite positive constant $\zeta_0$ such that $|\zeta_i| < \zeta_0$ for all $i = 1, ..., p$.

Assumption 3 is needed to make sure that the true weight vector does not grow without bound. This is required because in theorem 3.3.1 the true weight vector changes with sample size.

Similar to assumption 2 in the previous chapter, assumption 4 only requires a compact support for the parameters.

Assumption 5 is a strong assumption, but it can find its application, for example, in compressed sensing where the user has control over the design of the data matrix $\mathbf{X}$.

The form of the argument presented in the following theorem is exactly similar to the one given in [69], but it applies to our case as it is.

**Theorem 3.3.1** *Given that 1, 2, 3, and 4 are satisfied and $p \propto n^\alpha$ with $\alpha > 0$, then*

$$P(||\mathbf{w}_{opt} - \mathbf{w}^*||^2 > \xi_n) \leq c_0 \exp\{-\log(n^{1-\alpha}\xi_n)\} \tag{3.22}$$

*for some positive finite constant $c_0$ and $\xi_n$. Assume that $\xi_n \propto n^{-\alpha^*}$ for some $\alpha^* > 0$. Then if $0 < \alpha^* < \alpha < 1/2$, $P(||\mathbf{w}_{opt} - \mathbf{w}^*||^2 > \xi_n) \to 0$ as $n \to \infty$, and hence $\mathbf{w}_{opt}$ has estimation consistency..*

**Proof.**

Let $\mathrm{diag(d)} = \mathbf{D}$

$$\mathbf{w}_{opt} - \mathbf{w}^* = -(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})^{-1}\mathbf{D}\mathbf{w}^* + \tau(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})^{-1}\mathbf{X}^\top\epsilon$$

which implies

$$\begin{aligned}
E(||\mathbf{w}_{opt} - \mathbf{w}^*||_2^2) &\leq 2||(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})^{-1}\mathbf{D}\mathbf{w}^*||_2^2 + \\
&\quad 2E(||\tau(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})^{-1}\mathbf{X}^\top\epsilon||_2^2) \\
&\leq 2||(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})^{-1}||_2^2||\mathbf{D}\mathbf{w}^*||_2^2 + \\
&\quad 2\tau^2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})E(\epsilon^\top\mathbf{X}^\top\mathbf{X}\epsilon) \\
&\leq 2r_0^{-2}\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})||\mathbf{w}^*||_2^2 + \\
&\quad 2\tau^2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D})E(\epsilon^\top\mathbf{X}^\top\mathbf{X}\epsilon) \\
&\leq 2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D}) \\
&\quad (r_{low}^{-2}||\mathbf{w}^*||_2^2 + \tau Tr(\mathbf{X}^\top\mathbf{X})) \\
&\leq 2\lambda_{min}^{-2}(\tau\mathbf{X}^\top\mathbf{X} + \mathbf{D}) \\
&\quad (r_{low}^{-2}||\mathbf{w}^*||_2^2 + p\lambda_{max}(\mathbf{X}^\top\mathbf{X})\tau) \\
&\leq \frac{2r_{low}^{-2}||\mathbf{w}^*||_2^2 + 2\tau pnc_2}{(\tau nc_1 + r_1^{-1})^2} \\
&\leq \frac{2n^{-1}r_{low}^{-2}pc_3 + 2\tau pc_2}{\tau^2 nc_1^2} \\
&\leq \frac{2n^{-1}r_{low}^{-2}pc_3 + 2\tau_{max}pc_2}{\tau_{min}^2 nc_1^2}
\end{aligned}$$

Now by using Markov inequality

$$\begin{aligned}
P(||\mathbf{w}_{opt} - \mathbf{w}^*||_2^2 > \xi_n) &\leq \frac{2n^{-1}r_{low}^{-2}pc_3 + 2\tau_{max}pc_2}{\tau_{min}^2 nc_1^2\xi_n} \\
&\leq \frac{2r_{low}^{-2}pc_3 + 2\tau_{max}pc_2}{\tau_{min}^2 nc_1^2\xi_n}
\end{aligned}$$

for $n > 1$. Let $c_4 = \frac{2r_{low}^{-2}c_3+2\tau_{max}c_2}{\tau_{min}^2 c_1^2}$, then with the assumption that $p \propto n^{\alpha}$, the right hand side becomes $c_4 p(\xi_n n)^{-1} = c_4 c_5 n^{\alpha-1}\xi_n^{-1}$, where $c_5$ is a positive constant. Then,

$$P(||\mathbf{w}_{opt} - \mathbf{w}^*||^2 > \xi_n) \leq c_0 \exp\{-\log(n^{1-\alpha}\xi_n)\}$$

where $c_0 = c_4 c_5$. If $\xi_n \propto n^{-\alpha^*}$, then $n^{1-\alpha}\xi_n = n^{1-(\alpha+\alpha^*)}$. Then under condition $0 < \alpha^* < \alpha < 1/2$, the term $n^{1-(\alpha+\alpha^*)} \to \infty$ as $n \to \infty$. Therefore if $0 < \alpha^* < \alpha < 1/2$, the right hand side will approach $0$, which completes the proof.

**Theorem 3.3.2** *Under assumption 4 and 5,* $\mathbf{w}_{opt} \to \mathbf{w}^*$ *as* $n \to \infty$ *with* $p$ *fixed.*

**Proof.**

$$\mathbf{w}_{opt} = n\tau(n\tau\mathbf{I}_p + \mathbf{D})^{-1}\mathbf{w}^* + \tau(n\tau\mathbf{I}_p + \mathbf{D})^{-1}\mathbf{X}^\top\epsilon$$

As $n \to \infty$, $n\tau(n\tau\mathbf{I}_n + \mathbf{D})^{-1} \to \mathbf{I}_p$, and $\tau(n\tau\mathbf{I}_n + \mathbf{D})^{-1} \to \text{diag}(\mathbf{0}_p)$. Therefore, since $|\zeta_i| < \zeta_0$ for all $i = 1, ..., p$, $\mathbf{w}_{opt} \to \mathbf{w}_*$ as $n \to \infty$.

The results of theorem 3.3.2 imply strong consistency, but since the shrinkage of coefficients is not absolute, selection and sign consistency does not immediately follow from the results of the theorem. Our estimator is selection and sign consistent only in the limit. In order to enforce absolute shrinkage, we make use of another assumption:

**Assumption 6.** Assume that their exist a positive finite constant $M$ such that $|\mathbf{w}_i^*| \geq M$, $i \in S^*$. Also assume a small positive constant $\delta$ such that $0 < \delta < M$

Let $\mathbf{w}_{opt}^c = \mathbf{e} \circ \mathbf{w}_{opt}$, where $e_i = 1$ if $|\mathbf{w}_{opti}| \geq M - \delta$, and $0$ otherwise.

**Corollary 3.3.2.1** *Under assumptions 4,5, and 6,* $\mathbf{w}_{opt}^c$ *will be sign and selection consistent as* $n \to \infty$ *with* $p$ *fixed.*

**Proof.**

Based on theorem 3.3.2, there exist a finite positive integer $n_0$ such that for $n > n_0$, $|\mathbf{w}_{opti}| \geq M - \delta$ for $i \in S^*$, and $|\mathbf{w}_{opti}| < M - \delta$ for $i \notin S^*$. This completes the proof.

**Sparsity condition for the case p > n:** [112] describes the geometrical properties of the generalized ridge regression (GRR) estimator for the case $p > n$. In [112] it was

shown that the GRR estimator is constrained to lie in a subspace of dimensions at most $n$. Since the MAP estimator of our AO algorithm is a special case of GRR estimators, the above mentioned property has direct implications for our approach. In order to have accurate estimation, the true weight vector needs to be sparse, and it should not have more than $n$ non zero coefficients. For non sparse settings, there are no guarantees for accurate estimation.

**Convergence of Alternating Optimization:** Since both the subproblems in the AO iterations have a unique minimizer, the alternating optimization scheme for our joint model satisfies the Existence and uniqueness (EU) assumption. Below we will describe the EU assumption, and the theorem that states the convergence of AO algorithm to the joint local minimum.

**Existence and uniqueness (EU) assumption [113]:** Let $\psi_1, \psi_2 \subseteq \mathcal{R}^p$; and let $\Psi = \psi_1 \times \psi_2$. Assume $\mathbf{v} = (\mathbf{w}, \mathbf{s})^\top$, $\mathbf{w}, \mathbf{s} \in \mathcal{R}^p$. Let $g_{\mathbf{w}}(\mathbf{w}) = \mathcal{F}(\mathbf{w}, \mathbf{s}_0)$, and $g_{\mathbf{s}}(\mathbf{s}) = \mathcal{F}(\mathbf{w}_0, \mathbf{s})$, where $\mathbf{w}_0$ and $\mathbf{s}_0$ are some fixed values. If $\mathbf{v} \in \Psi$, then $g_{\mathbf{w}}(\mathbf{w})$ has a unique global minimizer for $\mathbf{w} \in \psi_1$, and $g_{\mathbf{s}}(\mathbf{s})$ has a unique global minimizer for $\mathbf{s} \in \psi_2$.

**Theorem 3.3.3** *[113]. Suppose the EU assumption is satisfied by $\mathcal{F}$. Let $\mathbf{v} = (\mathbf{w}, \mathbf{s})^\top$, and $\Psi = \psi_1 \times \psi_2$, where $\psi_1$ and $\psi_2$ are compact subsets of $\mathcal{R}^p$. Let $\{\mathbf{v}^{(r+1)} = T(\mathbf{v}^{(r)})\}$ denote the sequence of AO iterations begun at $\mathbf{v}^{(0)} \in \Psi$, and denote the fixed points of $T$ as $\Omega = \{\mathbf{v} \in \Psi : \mathbf{v} = T(\mathbf{v})\}$. Then:*

*(i) If $\mathbf{v}^* \in \Omega$, then $\mathbf{v}^* = (\mathbf{w}^*, \mathbf{s}^*)^\top$ satisfies,*

$$\mathbf{w}^* = \underset{\mathbf{w} \in \psi_1 \subset \mathcal{R}^p}{\operatorname{argmin}} \; g_{\mathbf{w}}(\mathbf{w})$$

$$\mathbf{s}^* = \underset{\mathbf{s} \in \psi_2 \subset \mathcal{R}^p}{\operatorname{argmin}} \; g_{\mathbf{s}}(\mathbf{s})$$

*(ii) $\mathcal{F}(\mathbf{v}^{(r+1)}) \leq \mathcal{F}(\mathbf{v}^{(r)})$, equality if and only if $\mathbf{v}^{(r)} \in \Omega$;*

*(iii) either: (a) $\exists \mathbf{v}^* \in \Omega$ and $r_0 \in \mathcal{R}$ so that $\mathbf{v}^{(r)} = \mathbf{v}^*$ for all $r \geq r_0$; or (b) the limit of every convergence subsequence of $\{\mathbf{v}^{(r)}\}$ is in $\Omega$.*

### 3.3.2 Marginal posterior of weights

Standard Laplace approximation requires to invert the Hessian matrix of the negative log probability at the mode, via which we can obtain a joint approximate posterior. For prediction and feature selection, however, we only need marginal posterior of each weight $w_j$, which only requires the diagonal entry of the inverse Hessian. Nevertheless, we still have to invert the Hessian matrix, which has time complexity of $O(p^3)$ and is unacceptable for large problems. To resolve this issue, we resort to Nyström method. Specifically, let us denote the mode of the model weights by $\tilde{\mathbf{w}}$ and consider the Hessian matrix in regression case first,

$$\mathbf{H} = \tau \mathbf{X}^\top \mathbf{X} + \mathrm{diag}(\mathbf{v})$$

where $v_j = -\frac{d^2 \log(p(w_j))}{dw_j^2}\Big|_{w_j=\tilde{w}_j}$, and $p(w_j)$ is the marginalized prior for $w_j$ (after marginalizing both $s_j$ and $z_j$). Then the Nyström approach is used to approximate $\mathbf{X}^\top\mathbf{X}$: A subset of columns of $\mathbf{X}$ are sampled to form a low-rank $n \times k$ matrix $\mathbf{X}_k = [\mathbf{f}_{i_1}, \ldots, \mathbf{f}_{i_k}]$, where $\mathbf{f}_{i_t}$ is the $i_t$-th column of $\mathbf{X}$; and $\mathbf{X}^\top\mathbf{X} \approx \mathbf{X}^\top\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^\dagger\mathbf{X}_k^\top\mathbf{X}$ where $(\cdot)^\dagger$ is the generalized inverse operation. The inverse of Hessian is then approximated by

$$\mathbf{H}^{-1} \approx \tilde{\mathbf{H}}^{-1}, \quad \tilde{\mathbf{H}} = \tau \mathbf{X}^\top\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^\dagger\mathbf{X}_k^\top\mathbf{X} + \mathrm{diag}(\mathbf{v}).$$

Applying Woodbury matrix identity [114], we can readily reduce the complexity to $O(nkp)$:

$$\begin{aligned}\tilde{\mathbf{H}}^{-1} = \ & \mathrm{diag}(\mathbf{v})^{-1} - \mathrm{diag}(\mathbf{v})^{-1}\mathbf{X}^\top\mathbf{X}_k(\tau^{-1}\mathbf{X}_k^\top\mathbf{X}_k \\ & + \mathbf{X}_k^\top\mathbf{X}\mathrm{diag}(\mathbf{v})^{-1}\mathbf{X}^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top\mathbf{X}\mathrm{diag}(\mathbf{v})^{-1}. \end{aligned} \tag{3.23}$$

Since we can choose $k \ll p$, the inversion cost will still be linear in $p$. We can then read off the diagonal of $\tilde{\mathbf{H}}^{-1}$ to calculate the marginal posterior approximation of each $w_j$: a Gaussian with mean $m_j$ being the posterior mode $\tilde{w}_j$ and variance $\sigma_j^2$ equal to the $j$-th entry of the diagonal of $\tilde{\mathbf{H}}^{-1}$.

For classification, the Hessian matrix has a slightly different form: $\mathbf{H} = \mathbf{X}^\top\mathrm{diag}(\mathbf{b})\mathbf{X} + \mathrm{diag}(\mathbf{v})$, where $b_i = \sigma(\mathbf{x}_i^\top\tilde{\mathbf{w}})(1 - \sigma(\mathbf{x}_i^\top\tilde{\mathbf{w}}))$. We can first multiply $\mathrm{diag}(\sqrt{\mathbf{b}})$ into $\mathbf{X}$, i.e., $\tilde{\mathbf{X}} = \mathbf{X}\mathrm{diag}(\sqrt{\mathbf{b}})$ and obtain $\mathbf{H} = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \mathrm{diag}(\mathbf{v})$. Then we follow the same procedure as in the regression case to calculate the Laplace approximation for each $w_j$.

Using Nyström approach to estimate the diagonal of inverse Hessian will inevitably bring some approximation error. To improve accuracy, a simple ensemble approach is proposed. Specifically, we first sample $d$ disjoint sets of columns of $\mathbf{X}$, each set is of the same size $k$. For each set $r$, we can calculate an approximate inverse Hessian $\tilde{\mathbf{H}}_r^{-1}$. The estimation of the $j$-th diagonal entry of inverse Hessian is then obtained by

$$\mathbf{H}^{-1}(j,j) \approx \frac{1}{d}\sum_{r=1}^{d} \tilde{\mathbf{H}}_r^{-1}(j,j). \tag{3.24}$$

Using Taylor expansion and error bounds of Nyström approximations[115], we can prove that the proposed ensemble approach can have a smaller estimation error bound. First we present the theoretical results for ensemble Nyström method presented in [115] which will be used in our theoretical analysis for our ensemble approach.

**Theorem 3.3.4** *[115, 116]. Let $Z_1, \cdots, Z_m$ be a sequence of random variables sampled uniformly without replacement from a fixed set of $m+u$ elements $Z$, and let $\phi : Z^m \to \mathrm{R}$ be a symmetric function such that for all $i \in [1,m]$ and for all $z_1, \cdots, z_m \in Z$ and $z_1^{'}, \cdots, z_m^{'} \in Z$, $|\phi(z_1, \cdots, z_m) - \phi(z_1, \cdots, z_{i-1}, z_i^{'}, z_{i+1} \cdots, z_m)| \leq c$. Then for all $\epsilon > 0$, the following inequality holds:*

$$Pr[\phi - E[\phi] \geq \epsilon] \leq \exp[\frac{-2\epsilon^2}{\alpha(m,u)c^2}] \tag{3.25}$$

*where* $\alpha(m,u) = \frac{mu}{m+u-1/2}\frac{1}{1-1/(2\max\{m,u\})}$

**Theorem 3.3.5** *[115]. Let $\tilde{\mathbf{H}}$ denote the rank-q Nystrom approximation of Hessian $\mathbf{H}$ based on $k$ columns sampled uniformly at random without replacement from $\mathbf{H}$ , and $\mathbf{H}_q$ the best rank-q approximation of $\mathbf{H}$. Then, with probability at least $1-\delta$ , the following inequality holds for any sample of size $k$:*

$$\begin{aligned} ||\mathbf{H} - \tilde{\mathbf{H}}||_F \leq ||\mathbf{H} - \mathbf{H}_k||_F + [\frac{64q}{k}]^{\frac{1}{4}} n\mathbf{H}_{max}\Big[1 + \sqrt{\frac{n-k}{n-1/2}\frac{1}{\beta(k,n)}log\frac{1}{\delta}} \\ d_{max}^{\mathbf{H}}/\mathbf{H}_{max}^{\frac{1}{2}}\Big]^{\frac{1}{2}} \\ = D_0 \end{aligned} \tag{3.26}$$

*where* $q \leq k$, $\beta(k,n) = 1 - \frac{1}{2max\{k,n-k\}}$, $\mathbf{H}_{max}$ *is the maximum diagonal entry of* $\mathbf{H}$, *and* $d^{\mathbf{H}}_{max} = \max_{ij}\sqrt{\mathbf{H}_{ii} + \mathbf{H}_{jj} - 2\mathbf{H}_{ij}}$.

**Theorem 3.3.6** *[115]. Let* $S$ *be a sample of* $dk$ *columns drawn uniformly at random without replacement from Hessian* $\mathbf{H}$, *decomposed into* $d$ *subsamples of size* $k$, $S_1, \cdots, S_d$. *For* $r \in [1,d]$, *let* $\tilde{\mathbf{H}}_r$ *denote the rank-q Nystrom approximation of Hessian* $\mathbf{H}$ *based on the sample* $S_r$, *and let* $\mathbf{H}_q$ *denote the best rank-q approximation of* $\mathbf{H}$. *Then, with probability at least* $1-\delta$ *, the following inequality holds for any sample* $S$ *of size* $dk$ *and for any* $\mu$ *in the simplex* $\Delta$ *and* $\tilde{\mathbf{H}}^{ens} = \sum_{r=1}^{d}\mu_r\tilde{\mathbf{H}}_r$*:*

$$\begin{aligned} ||\mathbf{H} - \tilde{\mathbf{H}}^{ens}||_F &\leq ||\mathbf{H} - \mathbf{H}_q||_F + [\frac{64q}{k}]^{\frac{1}{4}} n\mathbf{H}_{max}\Big[1 + \mu_{max}p^{\frac{1}{2}} \\ &\sqrt{\frac{n-dk}{n-1/2}\frac{1}{\beta(dk,n)}log\frac{1}{\delta}}d^{\mathbf{H}}_{max}/\mathbf{H}^{\frac{1}{2}}_{max}\Big]^{\frac{1}{2}} \\ &= D_1 \end{aligned} \tag{3.27}$$

*where* $\beta(k,n) = 1 - \frac{1}{2max\{dk,n-dk\}}$ *and* $\mu_{max} = max^{d}_{r=1}\mu_r$

Next we present results for our own ensemble approach.

**Theorem 3.3.7** *Define* $\Omega = \{\mathbf{A} \in \mathbb{R}^{p\times p} | \mathbf{A} \succ \mathbf{0}, \lambda_{min}(\mathbf{A}) \geq c, \lambda_{max}(\mathbf{A}) < \infty\}$. *Assume Hessian* $\mathbf{H}$ *and rank-q Nystrom approximation of* $\mathbf{H}$ *based on* $k$ *samples,* $\tilde{\mathbf{H}}$, *both belong to* $\mathbf{\Omega}$. *Consider a function* $f(\mathbf{A}) = \mathbf{e}_j^\top \mathbf{A}^{-1}\mathbf{e}_j, \mathbf{A} \in \mathbf{\Omega}$. *Then,* $\|\nabla f(\mathbf{A})\|_F \leq L$, $(1-\eta)\mathbf{H} + \eta\tilde{\mathbf{H}} \in \mathbf{\Omega}\ \forall\ \eta \in [0,1]$, *and with probability at least* $1-\delta$,

$$|\mathbf{H}^{-1}(j,j) - \tilde{\mathbf{H}}^{-1}(j,j)| \leq L \cdot D_0 \tag{3.28}$$

*where* $c$ *is a small positive constant, and* $L = p/c^2$. $\mathbf{e}_j$ *is a standard basis vector with* $1$ *in* $j$*-th coordinate and* $0$*'s elsewhere, and* $D_0$ *is the Nyström error bound based on Frobenius norm in theorem 3.3.5 [115].*

**Proof.**

The derivative of $f(\mathbf{A})$ can be calculated by

$$\nabla f(\mathbf{A}) = -\mathbf{A}^{-1}\mathbf{e}_j\mathbf{e}_j^\top\mathbf{A}^{-1}.$$

Now since $\mathbf{A} \succ \mathbf{0}$ and consequently $\mathbf{A}^{-1} \succ \mathbf{0}$, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{p} \lambda_i^2}$ and $\|\mathbf{A}^{-1}\|_F = \sqrt{\sum_{i=1}^{p} \frac{1}{\lambda_i^2}}$. Since $\lambda_{max}(\mathbf{A}) < \infty$ and $\lambda_{max}(\mathbf{A}^{-1}) < \infty$, $\|\mathbf{A}\|_F < \infty$ and $\|\mathbf{A}^{-1}\|_F < \infty$ $\implies$ $\|\nabla f(\mathbf{A})\|_F < \infty$. Now $\|\nabla f(\mathbf{A})\|_F \leq \|\mathbf{A}^{-1}\|_F^2 = \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \leq p(1/\lambda_{min}^2(\mathbf{A})) \leq p/c^2 = L$.

Since $f(\mathbf{A}) = \mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j$ for $\mathbf{A} \in \mathbf{\Omega}$, we can write $\mathbf{H}^{-1}(j,j) = f(\mathbf{H})$. Now, let us consider $|\mathbf{H}^{-1}(j,j) - \tilde{\mathbf{H}}^{-1}(j,j)| = |f(\tilde{\mathbf{H}}) - f(\mathbf{H})|$. We define $\mathbf{\Delta} = \tilde{\mathbf{H}} - \mathbf{H}$. Then $f(\tilde{\mathbf{H}}) = f(\mathbf{H} + \mathbf{\Delta})$. Now for any $0 \leq \eta \leq 1$, we have $(\mathbf{H} + \eta\mathbf{\Delta}) = ((1-\eta)\mathbf{H} + \eta\tilde{\mathbf{H}}) \succ \mathbf{0}$, and since, based on Weyl's inequality, $\lambda_{max}((1-\eta)\mathbf{H} + \eta\tilde{\mathbf{H}}) \leq (1-\eta)\lambda_{max}(\mathbf{H}) + \eta\lambda_{max}(\tilde{\mathbf{H}}) < \infty$, and $\lambda_{min}((1-\eta)\mathbf{H} + \eta\tilde{\mathbf{H}}) \geq (1-\eta)\lambda_{min}(\mathbf{H}) + \eta\lambda_{max}(\tilde{\mathbf{H}}) \geq (1-\eta)c + \eta c = c$, $\mathbf{H} + \eta\mathbf{\Delta} \in \mathbf{\Omega}$. This implies that $\|\nabla f(\mathbf{H} + \eta\mathbf{\Delta})\|_F \leq L$ for any $0 \leq \eta \leq 1$. Since $\frac{df(\mathbf{H}+\eta\mathbf{\Delta})}{d\eta} = \mathrm{tr}(\nabla f(\mathbf{H} + \eta\mathbf{\Delta})^\top \cdot \mathbf{\Delta})$, it is defined and bounded for all $\eta \in [0,1]$, hence it is continuous with respect to $\eta$. Therefore, by mean value theorem, there exist a number $t \in [0,1]$ such that:

$$f(\mathbf{H} + \mathbf{\Delta}) = f(\mathbf{H}) + \mathrm{tr}(\nabla f(\mathbf{H} + t\mathbf{\Delta})^\top \cdot \mathbf{\Delta}).$$

Thus by cauchy schwarz inequality,

$$|f(\mathbf{H} + \mathbf{\Delta}) - f(\mathbf{H})| \leq \|\nabla f(\mathbf{H} + t\mathbf{\Delta})\|_F \cdot \|\mathbf{\Delta}\|_F \leq L \cdot \|\mathbf{\Delta}\|_F.$$

Note that $\|\mathbf{\Delta}\|_F$ is the Nyström approximation error for $\mathbf{X}^\top\mathbf{X}$ and therefore we can readily apply the Nyström error bound $D_0$ [115].

**Theorem 3.3.8** *Define set $S$ to be a collection of $dk$ columns of Hessian $\mathbf{H}$ sampled uniformly at random without replacement, and partitioned into $d$ subsets of size $k$, $S_1, \cdots, S_d$. Assume Hessian $\mathbf{H}$ and $d$ rank-$q$ Nystrom approximations of $\mathbf{H}$, $\{\tilde{\mathbf{H}}_1, \ldots, \tilde{\mathbf{H}}_d\}$ where $\tilde{\mathbf{H}}_r$ denotes the rank-$q$ Nystrom approximation of Hessian $\mathbf{H}$ based on the subset $S_r$, all belong to $\mathbf{\Omega}$, then with probability at least $1 - \delta$,*

$$|\mathbf{H}^{-1}(j,j) - \frac{1}{d}\sum_{r=1}^{d} \tilde{\mathbf{H}}_r^{-1}(j,j)| \leq L \cdot D_1 \tag{3.29}$$

*where $D_1$ is the error bound for ensemble Nyström based on Frobenius norm in theorem 3.3.6 [115].*

If we use uniform weighting for ensemble Nystrom, $D_1 < D_0$ [115] and hence the ensemble approach for diagonal entry estimation of $\mathbf{H}^{-1}$ has a smaller error bound.

**Proof.**

First, we have

$$|\mathbf{H}^{-1}(j,j) - \frac{1}{d}\sum_r \tilde{\mathbf{H}}_r^{-1}(j,j)| = \frac{1}{d}|\sum_r f(\mathbf{H}) - f(\tilde{\mathbf{H}}_r)| \leq \frac{1}{d}\sum_r |f(\mathbf{H}) - f(\tilde{\mathbf{H}}_r)|.$$

Following (3.29), we have

$$|\mathbf{H}^{-1}(j,j) - \frac{1}{d}\sum_r \tilde{\mathbf{H}}_r^{-1}(j,j)| \leq L \cdot \frac{1}{d}\sum_{r=1}^{d} \|\mathbf{\Delta}_r\|_F$$

where $\mathbf{\Delta}_r = \tilde{\mathbf{H}}_r - \mathbf{H}$. From the proof of Theorem 3 in [115], we can see that the error bound for ensemble Nyström is obtained by calculating the bound for $\sqrt{\frac{1}{d}\sum_{r=1}^{d} \|\mathbf{\Delta}_r\|_F^2}$ (note that the error for our ensemble approach is upper bounded by $\frac{1}{d}\sum_{r=1}^{d} \|\mathbf{\Delta}_r\|_F$). Therefore, using Jensen's inequality, we can directly apply the resulting error bound $D_1$ to obtain

$$|\mathbf{H}^{-1}(j,j) - \frac{1}{d}\sum_r \tilde{\mathbf{H}}_r^{-1}(j,j)| \leq L \cdot D_1.$$

**Proposition 1** *Assume that $\lambda_{max}(\mathbf{X}^\top\mathbf{X}) < \infty$, and $\forall j\ c \leq v_j < \infty$. Then both Hessian $\mathbf{H}$ and any approximate Hessian $\tilde{\mathbf{H}}$ based on Nyström method belong to $\mathbf{\Omega}$, and hence satisfy theorems 3.3.7 and 3.3.8.*

**Proof.**

Since $\tau\mathbf{X}^\top\mathbf{X} \succeq \mathbf{0}$ and $\mathrm{diag}(\mathbf{v}) \succ \mathbf{0}$, $H = \tau\mathbf{X}^\top\mathbf{X} + \mathrm{diag}(\mathbf{v}) \succ \mathbf{0}$. Now by Weyl's inequality, $\lambda_{max}(\tau\mathbf{X}^\top\mathbf{X} + \mathrm{diag}(\mathbf{v})) \leq \lambda_{max}(\tau\mathbf{X}^\top\mathbf{X}) + \lambda_{max}(\mathrm{diag}(\mathbf{v})) < \infty$; $\lambda_{min}(\tau\mathbf{X}^\top\mathbf{X} + \mathrm{diag}(\mathbf{v})) \geq \lambda_{min}(\tau\mathbf{X}^\top\mathbf{X}) + \lambda_{min}(\mathrm{diag}(\mathbf{v})) \geq c$, $\lambda_{min}(\tau\mathbf{X}^\top\mathbf{X}) \geq 0$. Therefore, $\mathbf{H} \in \mathbf{\Omega}$.

Using theorem 3.5 in [117] we can conclude that $\tau\mathbf{X}^\top\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^\dagger\mathbf{X}_k^\top\mathbf{X} \succeq \mathbf{0}$, therefore $\tilde{\mathbf{H}} \succ \mathbf{0}$. Based on theorem 3.8 in [117] $\lambda_{max}(\tau\mathbf{X}^\top\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^\dagger\mathbf{X}_k^\top\mathbf{X}) \leq \lambda_{max}(\tau\mathbf{X}^\top\mathbf{X}) < \infty$, and $\lambda_{min}(\tau\mathbf{X}^\top\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^\dagger\mathbf{X}_k^\top\mathbf{X}) \geq 0$. Therefore, combined with Weyl's inequality, $\lambda_{max}(\tilde{\mathbf{H}}) < \infty$, and $\lambda_{min}(\tilde{\mathbf{H}}) \geq c$. Therefore, $\tilde{\mathbf{H}} \in \mathbf{\Omega}$.

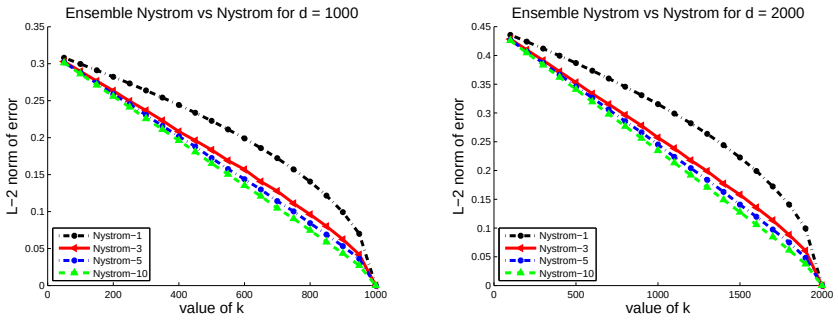
For the joint model, we approximate the marginalized distribution based on its mode, i.e., $p(\mathbf{w}, \mathbf{t}, \mathbf{X}) \approx p(\mathbf{w}, \hat{\mathbf{s}}, \mathbf{t}, \mathbf{X})$ where $\hat{\mathbf{s}}$ is the mode of $\mathbf{s}$, and $v_j = 1/r_1$ or $v_j = 1/r_0$.

If the value of $r_1$ is such that $c \leq 1/r_1$, Hessian $\mathbf{H}$ and any approximate Hessian $\tilde{\mathbf{H}}$ will satisfy theorems 3.3.7 and 3.3.8 for the joint model.

To empirically demonstrate the effectiveness of the ensemble method, we estimate the diagonal of the inverse of synthetically generated symmetric positive definite matrices defined as [118]:

$$a_{ij} = \begin{cases} \frac{1}{|i-j|^2} & \text{if } i \neq j \\ 1 + \sqrt{i} & \text{if } i = j \end{cases} \tag{3.30}$$

Figs 3.1 a and b show the RMSE plots for $p = 1000$, and $p = 2000$ respectively. The results clearly demonstrate that ,if $k$ remains fixed, the increase in the number of ensembles, $d$, decreases the error value.




(a) p = 1000 (b) p = 2000

Figure 3.1.: Root mean square error of the diagonal of the inverse matrix.

### 3.3.3 Posteriors moments of $s_j$ and $z_j$

Given the approximate marginal posterior of $w_j$, we can estimate marginal posterior moments of $s_j$—the probability of selecting the $j$-th feature. Specifically, we first invert the conditional relationship between $s_j$ and $w_j$ based on Bayes rule,

$$p(s_j|w_j) = \frac{s_j \mathcal{N}(w_j|0, r_1) + (1 - s_j)\mathcal{N}(w_j|0, r_0)}{\frac{1}{2}\mathcal{N}(w_j|0, r_1) + \frac{1}{2}\mathcal{N}(w_j|0, r_0)}. \tag{3.31}$$

Then the marginal posterior of $s_j$ can be computed by

$$p(s_j|\mathbf{t}, \mathbf{X}) = \int p(s_j|w_j)\mathcal{N}(w_j|m_j, \sigma_j^2)\mathrm{d}w_j \tag{3.32}$$

where $\mathcal{N}(w_j|m_j, \sigma_j^2)$ is the estimated posterior marginal of $w_j$. Then, the posterior mean and variance of $s_j$ are calculated by

$$\mathrm{E}[s_j] = \int \frac{2\mathcal{N}_1(w_j) + \mathcal{N}_0(w_j)}{3(\mathcal{N}_1(w_j) + \mathcal{N}_0(w_j))} q(w_j)\mathrm{d}w_j \tag{3.33}$$

$$\mathrm{Var}[s_j] = \int \frac{3\mathcal{N}_1(w_j) + \mathcal{N}_0(w_j)}{6(\mathcal{N}_1(w_j) + \mathcal{N}_0(w_j))} q(w_j)\mathrm{d}w_j - \mathrm{E}^2[s_j] \tag{3.34}$$

where $\mathcal{N}_g(w_j)$ (for $g = 0, 1$) and $q(w_j)$ are the shorthand for $\mathcal{N}(w_j|0, r_g)$ and $\mathcal{N}(w_j|m_j, \sigma_j^2)$ respectively.

A similar procedure can be used to calculate the posterior moments of $z_j$—the selection indicator of $j$-th feature; the poster mean and variance of $z_j$ are given by

$$\mathrm{E}[z_j] = \int \frac{\mathcal{N}_1(w_j)}{\mathcal{N}_1(w_j) + \mathcal{N}_0(w_j)} q(w_j)\mathrm{d}w_j \tag{3.35}$$

$$\mathrm{Var}[z_j] = \int \frac{\mathcal{N}_1(w_j)}{\mathcal{N}_1(w_j) + \mathcal{N}_0(w_j)} q(w_j)\mathrm{d}w_j - \mathrm{E}^2[z_j]. \tag{3.36}$$

The integrations to calculate the means and variances of $s_j$ and $z_j$ do not have a closed-form solution. So we apply Gauss-Hermite quadrature method [110] to obtain an estimation. Since the integration is one dimensional and smooth, the quadrature method is computationally efficient and accurate: With only 5 quadrature nodes (or function evaluations), we can estimate $\mathrm{E}(s_j)$, $\mathrm{E}(z_j)$, $\mathrm{Var}(s_j)$, and $\mathrm{Var}(z_j)$ with high accuracy (e.g. the numerical difference from the true integration is often on the order of $10^{-4}$).

The over all time complexity of our algorithms is $O(dknp)$, $d, k << p$, including estimating the posterior mean and variance of $\mathbf{w}$, $\mathbf{s}$, and $\mathbf{z}$. The linear cost makes our algorithm scalable for high dimensional data.

## 3.4 Related work

[69] proposed a MAP estimation of spike-and-slab models with delta spikes. They approximate the delta spike by a continuous bound via an elegant majorization and minimization (MM) algorithm. They also provide consistency results for their MAP estimate. We, on the other hand, assume continuous spikes to make use of efficient continuous optimization strategies. Secondly, while they only focus on the MAP estimate, we provide

a full Bayesian inference strategy, and also show oracle properties for our MAP estimate. Another closely related work is proposed by [62]. There are few differences between our approach and their method. First, while they employ a rescaled spike-and-slab model with a bimodal continuous prior on the variances of regression weights, we do not performing any rescaling, and use a two point discrete prior for the variances. Secondly, they present estimation and selection consistency results for the posterior mean of the regression weights in a Gibbs sampling framework, whereas we provide consistency results for the MAP estimate in a Laplace approximation settings. Gibbs sampling framework is not suitable for high dimensional settings, because the sampler will be very slow to converge. Our approach, on the other hand, utilizes highly efficient optimization strategies, and hence is scalable to high dimensions.

EP and VB approximations have been developed to conduct Bayesian inference on the spike-and-slab model. In [61], EP was applied to learn the spike-and-slab model for multi-task learning, where the weights $\mathbf{w}$ were factorized over multiple tasks. For one task, the computational complexity is $O(n^2p)$ when $n < p$ (or $O(np^2)$ when $n > p$). Further, [60] imposed a fully factorized approximate posterior of $\mathbf{w}$ in EP and achieved a cost of $O(np)$ with $n < p$ in the classification context. Similarly, a cost of $O(np^2)$ or $O(n^2p)$ was spent for the VB approximation with fully factorized posterior assumption [107, 119]. In addition, to estimate the hyperparameters, such as selection probabilities, [107] used variational EM to obtain the point estimate, while [119] used importance sampling.

Unlike previous methods, we neither assume the joint posterior $p(\mathbf{w}, \mathbf{z}, \mathbf{s}|\mathbf{X}, \mathbf{y})$ to have a factorized form such as $\prod_j q(w_j, z_j, s_j)$, nor try to find such an approximate posterior close to the true posterior (in terms of KL divergence). Instead, we start from Laplace approximation and calculate the approximate posterior marginal of each $w_j$ separately. Then we use these marginal posteriors to quantity the selection uncertainty, including selection probabilities and indicators. In this way, our method not only enjoys a linear cost in $p$, but also avoids the strong factorization assumptions which could hurt the inference quality [119].

However, similar to the existing approximate approaches including VB, EP and MM, our method is not able to recover the multimodal nature of the true posterior of spike-and-slab models. Our Laplace based method, may trap in a local mode and result in poor approximations like VB and MM. EP can alleviate this problem by summarizing the information from all the modes, but it still returns a unimodal approximation. Moreover, EP has issues in convergence guarantees; due to its fixed point iteration nature, the algorithm can diverge; some heuristic tricks such as damping can be used to avoid divergence [120], but without assurance. In summary, the posterior multimodality for spike-and-slab models remains an open problem for efficient approximate inference algorithm design.

## 3.5 Experiments

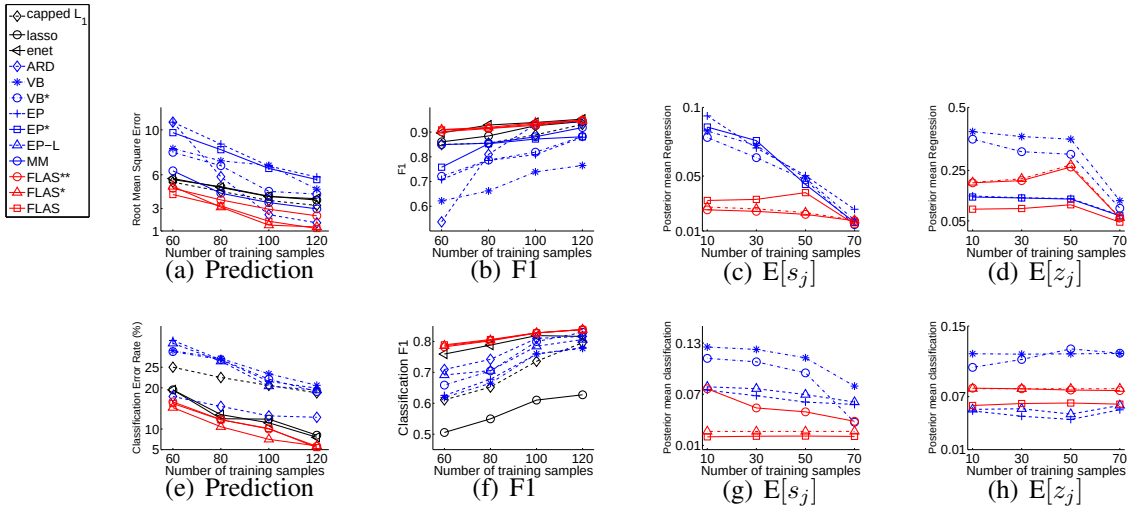



Figure 3.2.: Simulation results, including the prediction accuracy, the F1 score of feature selection, and the root mean squared error for the posterior mean estimation of $\{s_j\}$ and $\{z_j\}$. Results are averaged over 50 runs.

Table 3.1.: The training time (seconds) on simulated data ($p = 1000$). Results generated by our best method are highlighted in red. In case an alternative method generates best result, it is highlighted in blue.

(a) Regression

| method | 60 | 80 | 100 | 120 |
|---|---|---|---|---|
| capped $L_1$ | **0.0054**± **0.0049** | 0.0705 ±0.0015 | **0.0103**± **0.0002** | **0.0108**±**0.0003** |
| lasso | 0.0312± 0.0100 | 0.0313± 0.0028 | 0.0321 ± 0.0043 | 0.0329±0.0044 |
| elastic net | 0.0360 ± 0.0204 | 0.0346 ± 0.0045 | 0.0352 ± 0.0047 | 0.0349±0.0038 |
| ARD | 0.03 ±0.042 | 0.17 ±0.017 | 0.20±0.0093 | 0.67 ±0.0088 |
| VB | 2.4161± 0.0702 | 2.3999 ± 0.0795 | 2.4794 ± 0.0773 | 2.4404±0.0470 |
| VB* | 2.7544± 0.0413 | 3.0118 ±0.0241 | 2.6012 ±0.0324 | 2.7795±0.0564 |
| EP | 0.9345 ± 0.0341 | 1.0478 ±0.0195 | 1.1160 ± 0.0058 | 1.1468±0.0078 |
| EP* | 0.505± 0.0102 | 0.681 ±0.0119 | 1.047 ± 0.0086 | 1.936±0.0091 |
| MM | 2.5230 ±0.1036 | 1.1047±0.1209 | 0.4314 ±0.1551 | 0.5282±0.0864 |
| FLAS** | 0.0664 ±0.0055 | 0.0642 ±0.0041 | 0.0704 ±0.0045 | **0.0855**± **0.006** |
| FLAS* | 0.1419 ± 0.0107 | 0.1321 ±0.0091 | 0.1718 ±0.0139 | 0.1923±0.0084 |
| FLAS | **0.0140**± **0.0015** | **0.0154** ± **0.0003** | **0.0216** ±**0.0007** | 1.4526±0.0438 |

(b) Classification

| method | 60 | 80 | 100 | 120 |
|---|---|---|---|---|
| capped $L_1$ | 0.0180±0.017 | 0.0499±0.0001 | 0.0427±0.0004 | 0.0559±0.0005 |
| lasso | 0.1033 ± 0.0185 | 0.1289 ±0.0157 | 0.1555 ± 0.0316 | 0.1821 ±0.0277 |
| elastic net | 0.08690 ± 0.0268 | 0.1009 ± 0.0095 | 0.1163 ± 0.0182 | 0.1356 ±0.0195 |
| ARD | 0.06 ±0.011 | 0.07± 0.023 | 0.15± 0.032 | 0.45±0.0091 |
| VB | 10.3312± 0.1850 | 11.2570 ± 0.1144 | 12.3317± 0.1364 | 13.3366±0.1470 |
| VB* | 0.0812± 0.087 | 0.1570 ±0.017 | 2.8915 ±0.01102 | 3.0194±0.0221 |
| EP | 1.1165± 0.0303 | 1.1695 ± 0.0257 | 1.2400 ± 0.0132 | 1.3090±0.0076 |
| EP-L | 0.0132± 0.0085 | 0.0581± 0.0081 | 0.0598± 0.0092 | 0.1631±0.0045 |
| FLAS** | 0.0344± 0.02 | 0.0736 ± 0.02 | 0.0794± 0.03 | 0.1929 ±0.06 |
| FLAS* | 0.0696± 0.0026 | 0.0832± 0.0046 | 0.1047±0.0052 | 0.1594±0.0077 |
| FLAS | **0.0097** ± **0.0002** | **0.0111** ± **0.0003** | **0.0139** ± **0.0002** | **0.0152**±**0.0005** |

### 3.5.1 Simulation

First we examine our method in a simulation study. The study aims to evaluate our algorithm in three aspects: (i) the predictive performance when $p \gg n$, (ii) the capability to select relevant features and (iii) the accuracy of the estimated posteriors of $s_j$ and $z_j$.

**Data Generation.** The feature dimension $p$ is set to $1000$. We assume $20$ out of the $1000$ features are relevant to the response. The irrelevant features are generated independently from the standard Gaussian distribution. The relevant features are generated from a multivariate Gaussian distribution with a block diagonal covariance matrix. The covariance matrix consists of two 10 by 10 sub-covariance matrices on the main diagonal. In each sub-covariance matrix, the diagonal elements are set to 1 and the off-diagonal elements are set to 0.81. Therefore, the 20 features are generated from two different groups. The weights $\mathbf{w}$ are set as

$$\mathbf{w} = [\underbrace{0, \ldots, 0}_{980}, \mathbf{v}, \mathbf{v}/\sqrt{10}, -\mathbf{v}, -\mathbf{v}/\sqrt{10}] \tag{3.37}$$

where $\mathbf{v} = [5, 5, 5, 5, 5]$. Given the sampled $\mathbf{X}$, for regression the response vector $\mathbf{t}$ is generated by $\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where each $\epsilon_i$ is sampled independently from the standard Gaussian. For classification, we generate each response by $t_i = -1 \cdot \delta(\mathbf{x}_i^\top \mathbf{w} < 0) + 1 \cdot \delta(\mathbf{x}_i^\top \mathbf{w} > 0)$, where $\delta(x) = 1$ if $x = 1$ and $0$ otherwise. We fix the number of test samples to $200$ and vary the number of training samples $n$ from $\{60, 80, 100, 120\}$. For each $n$, we randomly generate 50 datasets and report the average results. To evaluate the accuracy of posterior inference, we run another simulation with similar sampling procedure but the feature dimension $p$ is set to $100$. The reason we choose a relatively small number of features is that we need to evaluate the accuracy of posterior inference results via comparing with Gibbs sampling, which converges slowly for high dimensional problems.

**Competing methods.** We compare our approach with alternative approximate inference algorithms for the spike-and-slab model, including VB, EP, and MM [69] that only provides MAP estimation. We implement two versions of EP algorithms, where for regression, one is based on continuous spikes proposed by [61](EP) and the other is based on delta spikes (EP$^*$); for classification, one is used by [121] and is similar to [61] (and

thus we also denote it by EP); the other has a better time complexity [60], and we denote it by EP-L. Both EP and EP* have the cost $O(np^2)$, while EP-L uses fully factorized posterior assumption for model weights to obtain a linear cost $O(np)$. For VB, we use two versions: [122](VB) having cubic cost $O(p^3)$ but without a factorized posterior assumption over model weights, and [107](VB*) using a fully factorized posterior assumption with reduced cost $O(np^2)$. For all these methods, including Gibbs sampling, we apply the same model in Section 2 where the selection probabilities $\{s_j\}$ are not integrated out. Because VB and EP only provide point estimates of the selection probabilities $\{s_j\}$, we modify them to obtain their posteriors using an approach similar to [123]. We also test other popular sparse learning methods, including ARD, lasso, elastic net, and capped $L_1$. We use the Glmnet[1] software package for lasso and elastic net (the package performs the tuning of hyper parameters through cross validation), and the Gist[2] software package for capped $L_1$. For these software packages, we use the default settings (e.g. initial value settings and maximum iteration number). For our methods we use the solution of $L_2$ regularization as initialization. The variances for spike-and-slab components, i.e., $r_0$ and $r_1$ are chosen from cross validation. The grids used are $r_0 = [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}]$ and $r_1 = [1 : 1 : 5]$. We use the same cross validation grid for competing methods. In the step of using Nyström approach to calculate Laplace approximation, we sample $5$ columns for each Nyström approximation and repeat $5$ times for ensemble estimation of the inverse Hessian diagonal.

**Results.** Figures 3.2 a and e show the predictive performance of all the methods for regression and classification. Our methods consistently outperform the alternative methods apart from ARD whose performance becomes better than FLAS** beyond $n = 100$ for regression case. Figures 3.2 b and f report the feature selection accuracy based on the F1 score, i.e., the harmonic average of the sensitivity and the specificity of the selected feature set. To compute the F1 score, we select features when the posterior mean of the selection indicators, $\mathrm{E}(z_j)$, is over $0.5$ for Bayesian spike-and-slab models, or when model weights $|w_j| > 0.001$ for other methods. As we can see, our methods achieve higher F1 scores for classification and comparable F1 score than the best alternatives in regression.

[1] `www-stat.stanford.edu/~tibs/glmnet-matlab`
[2] `www.public.asu.edu/~jye02/Software/GIST/`

To examine the quality of the estimated uncertainty for feature selection, we calculate the posterior mean of each selection probability $s_j$ as well as the selection indicator $z_j$, and compare them with the ground truth obtained by Gibbs sampling with $100,000$ samples. We calculate the root mean square error to evaluate the difference from the ground truth and report the results in Figure 3.2 c, d, g, and h. It is quite evident from the plots, that our methods consistently obtain better uncertainty estimation than competing methods, except in the classification case in which they are slightly worse than EP and EP-L in recovering the posteriors of the selection indicators. This confirms the inference quality of our algorithms.

Finally, the running time of all the algorithms is reported in Table 3.1 a and b. It turns out that for both regression and classification, our methods converge faster than EP and VB, and are comparable to $L_1$ type methods. Therefore, our methods not only achieve superior Bayesian inference quality, but are computationally as efficient as the frequentist approaches.

Table 3.2.: Regression training data sets sizes

| datasets | GSE5680 | 10k corpus | House-census | tied | Yearprediction | dlbcl |
|---|---|---|---|---|---|---|
| $n$ | 120 | 3308 | 22784 | 750 | 463715 | 240 |
| $p$ | 31041 | 150358 | 138 | 999 | 90 | 752 |

Table 3.3.: Classification training data sets sizes

| datasets | classic | hitech | k1b | reviews | sports | ng3sim | ohscal | la12 |
|---|---|---|---|---|---|---|---|---|
| $n$ | 709 | 230 | 234 | 406 | 858 | 299 | 1116 | 627 |
| $p$ | 41681 | 10080 | 21819 | 18483 | 14870 | 15810 | 11465 | 31472 |

Table 3.4.: Root mean square error on regression datasets (the first 6 rows) and classification error rates (%) on large binary classification datasets (the last 8 rows). The results are averaged over 10 runs. Results generated by our best method are highlighted in red. In case an alternative method generates best result, it is highlighted in blue.

| dataset | lasso | elast net | capped $L_1$ | ARD | EP-L | FLAS** | FLAS* | FLAS |
|---|---|---|---|---|---|---|---|---|
| gse5680 | 0.107 ± 0.003 | 0.107 ± 0.003 | 0.107± 0.003 | 0.136 ± 0.005 | 0.72 ± 0.001 | 0.122 ± 0.002 | 0.111 ± 0.002 | **0.089± 0.002** |
| 10k corpus | 0.382 ± 0.002 | 0.382 ± 0.002 | 0.382 ± 0.002 | 0.382 ± 0.002 | 0.385 ± 0.003 | 0.383 ± 0.003 | 0.383 ± 0.003 | **0.372 ± 0.003** |
| tied | 0.656 ± 0.013 | 0.627 ± 0.014 | 0.656 ± 0.013 | **0.532 ± 0.017** | 1.11 ± 0.2 | 0.719 ± 0.012 | **0.632 ± 0.017** | 0.656 ± 0.013 |
| House | 1.576 ± 0.011 | 1.578 ± 0.017 | 1.587 ± 0.012 | 0.435 ±0.0006 | 0.430 ± 0.0002 | 0.561 ± 0.015 | 0.441 ± 9.5e-4 | **0.425 ±0.002** |
| Year | 0.296 ± 0.009 | 0.293 ± 0.007 | 0.307 ± 0.004 | 0.306 ± 0.006 | 0.32 ± 0.002 | 0.248 ± 0.0005 | **0.232± 5.04e-4** | 0.234 ± 0.0001 |
| dlbcl | 1.76 ± 0.026 | 1.75 ± 0.027 | 1.75 ± 0.028 | 2.38 ± 0.063 | 1.61 ± 0.050 | 1.60 ± 0.047 | **1.56 ± 0.043** | 1.60± 0.047 |
| classic | 6.69 ± 0.002 | 5.94 ± 0.002 | **4.14 ± 0.002** | 18.2 ± 0.002 | 8.94 ± 0.002 | 5.76 ± 0.002 | 4.2 ± 0.002 | **4.20 ± 0.001** |
| hitech | 23.2 ± 0.005 | 21.4 ± 0.004 | 21.3 ± 0.003 | 28.5 ± 0.019 | 25.2 ± 0.001 | **19.4 ± 0.003** | 19.9 ± 0.002 | 19.9 ± 0.003 |
| k1b | 5.44 ± 0.005 | 4.91 ± 0.004 | **4.42 ± 0.004** | 23.0 ± 0.013 | 7.94 ± 0.004 | 5.03 ± 0.005 | **4.73 ± 0.005** | 4.74 ± 0.005 |
| reviews | 7.68 ± 0.003 | 6.47 ± 0.002 | 6.09 ± 0.001 | 35.4 ± 0.05 | 8.28 ± 0.002 | 5.93 ± 0.002 | 5.55 ± 0.001 | **5.54 ± 0.001** |
| sports | 3.72 ± 0.001 | 3.15 ± 0.0008 | 3.25 ± 0.0009 | 24.1 ± 0.032 | 10.9 ± 0.008 | 2.78 ±0.001 | **2.77 ±0.0006** | 2.77 ±0.007 |
| ng3sim | 19.3 ± 0.005 | 16.2 ± 0.003 | 15.4 ± 0.003 | 21.3 ± 0.006 | 14.5 ± 0.002 | 13.7 ± 0.003 | 13.7 ± 0.002 | **13.6 ± 0.002** |
| ohscal | 13.8 ± 0.001 | 13.7 ± 0.001 | 13.8 ± 0.001 | 37.3 ± 0.02 | 13.7 ±0.002 | **11.9 ± 0.001** | 13.05 ±0.001 | 13.1 ± 0.001 |
| la12 | 13.6 ± 0.002 | 12.5 ± 0.002 | 12.2 ± 0.002 | 30.1 ± 0.025 | 13.2 ± 0.002 | 11.1 ± 0.002 | **11.04 ± 0.001** | 11.1 ± 0.001 |

Table 3.5.: Root mean square error on regression datasets (the first 3 rows) and classification error rates (%) on binary classification datasets (the last 4 rows) after dimension reduction. The results are averaged over 10 runs. FLAS is applied to reduce the data dimensions before the test. Results generated by our best method are highlighted in red. In case an alternative method generates best result, it is highlighted in blue.

| dataset | EP | VB | FLAS** | FLAS* | FLAS |
|---|---|---|---|---|---|
| gse5680 | 0.191 ± 0.008 | 0.238 ± 0.009 | 0.195±0.008 | **0.101±0.002** | 0.197±0.008 |
| 10k corpus | 0.382 ± 0.002 | 0.382 ± 0.002 | **0.381 ± 0.002** | **0.381 ± 0.002** | **0.381± 0.002** |
| tied | **0.5787± 0.013** | 0.7983 ± 0.011 | **0.5868 ± 0.013** | 0.6571 ± 0.013 | 0.5877±0.013 |
| hitech | 24.31 ± 0.046 | 22.48 ± 0.006 | 19.64 ± 0.004 | 20.07 ± 0.004 | **19.38 ± 0.004** |
| k1b | 9.37 ± 0.003 | 9.34 ± 0.002 | 5.4 ± 0.006 | **4.82 ± 0.005** | 5.91 ± 0.006 |
| reviews | 10.2 ± 0.0004 | 10.1 ± 0.004 | 6.39 ± 0.002 | **5.55 ± 0.002** | 6.16 ± 0.002 |
| ng3sim | 21.3 ± 0.006 | 24.37 ± 0.007 | 14.2 ± 0.004 | **14.07 ± 0.002** | 14.65 ± 0.005 |

### 3.5.2 Large real benchmark data

We then examine all the algorithms on 14 published large real datasets, including 8 classification datasets[3] and 6 regression datasets: Diffuse large B cell lymphoma (DLBCL) [82], GSE5680 [124], Yearprediction[4](Year), House-census[5](House), 10K corpus [125] and TIED[6]. Among the 14 datasets, the feature numbers are often at tens of thousands, while the sample sizes are often at hundreds or thousands. Detailed information about these datasets is provided in tables 3.2 and 3.3.

We compare our algorithms with lasso, elastic net, capped $L_1$, ARD and EP-L. Note that we implement lasso and elastic net based on GIST, because the Glmnet software used in simulation is no longer feasible. For the intractable EP and VB methods—the ones with computational cost $O(p^3)$ or $O(np^2)$, we first reduce the dimensionality of the datasets, by running FLAS and pruning all features with posterior mean selection probability less than $0.5$, and then perform the comparison. We randomly split each dataset into two parts—$10\%$ samples for training and the rest for test—for 10 times and run all the methods on each partition. In each run, we use 10-fold cross validation on the training data to tune the free parameters. Table 3.4 lists the average prediction accuracy and standard errors on the original datasets. As we can see, in all datasets, except for *Tied* in regression, and *classic* and *k1b* in classification, our algorithms obtain smaller root mean square errors or classification error rates. Table 3.5 shows the prediction accuracy on the datasets with reduced dimensionality; that is for the comparison with intractable EP and VB algorithms. It turns out that our methods perform better than or comparable to the intractable EP and VB methods; however, our methods have the scalability advantage in high dimensional problems. We also examine the average training time of all the methods and it turns out that our approach spends comparable time to the best $l_1$ type method, and less time than EP and ARD approaches. Table 3.6 shows all the training convergence times. All algorithms were initialized using $l_2$ regularization based solution.

[3] www.shi-zhong.com/software/docdata.zip
[4] archive.ics.uci.edu/ml/datasets.html
[5] www.cs.toronto.edu/~delve/data/census-house/desc.html
[6] www.causality.inf.ethz.ch/repository.php

Table 3.6.: The average convergence time on real training data sets. Results generated by our best method are highlighted in red. In case an alternative method generates best result, it is highlighted in blue.

| dataset | lasso | elast net | capped $L_1$ | ARD | FLAS** | FLAS* | FLAS | EP-L |
|---|---|---|---|---|---|---|---|---|
| gse5680 | 2.03± 4.3e-3 | 2.26± 9.5e-3 | 1.53±2e-3 | 3.52±1e-3 | 2.1±1.1e-1 | **0.15±7.8e-3** | 0.3±1.3e-2 | 6.52±3e-3 |
| 10k corpus | 0.71±1e-1 | **0.49± 1.9e-2** | 1.69 ±3.2e-1 | 6.1±2.1e-2 | 3.39±3.3e-1 | 2.1±1.2e-1 | **1.1±2.5e-2** | 2.18±2.2e-2 |
| tied | 0.30±1e-2 | 0.32±2.2e-2 | **0.01±1.6e-3** | 5.9±1.5e-2 | 0.06±2.7e-3 | **0.02±7.8e-4** | 0.03±1.9e-3 | 1.69±1.6e-3 |
| dlbcl | 0.44±9.9e-2 | 0.39 ±4e-2 | **0.02±6.5e-3** | 0.8±3.1e-3 | **0.08±1.9e-2** | 0.08±4e-2 | 0.09±3.2e-3 | 0.82±2e-3 |
| classic | 2.1±2.5e-1 | 0.63±2.9e-2 | **0.12±6.2e-3** | 19.4±2.3e-2 | 11.9±1.5 | 1.31±1.9e-1 | **0.78±3.6e-2** | 2.57±2.2e-2 |
| hitech | 1.33±5.4e-2 | 0.44±1.8e-2 | **0.07±5.7e-3** | 12.1 ±1.2e-3 | 2.15 ±3.3e-1 | 1.2±1.3e-1 | **0.28±5.9e-2** | 2.46±1.8e-2 |
| k1b | 1.6 ± 1.5e-1 | 0.49±3.8e-2 | **0.06±1e-2** | 16.4±1.7e-2 | 1.5±8.1e-2 | 1.1±1.4e-1 | **0.2±1.8e-2** | 30.5±1.3e-3 |
| reviews | 0.32 ± 2.2e-1 | 0.29 ± 6.8e-2 | 2.3± 1.4e-2 | 26.7 ± 1.5e-3 | 0.15±4.4e-1 | 1.41±9.4e-2 | **0.10±1e-2** | 1.02±1.1e-2 |
| sports | 2.61 ± 5.4e-1 | 1.15 ± 7.3e-2 | **0.22±1.1e-2** | 0.45±1.9e-2 | 5.26±1.29 | 4.7 ±3.9e-1 | **0.4±9.6e-3** | 2.23±2.1e-2 |
| ng3sim | 3.43± 6.1e-2 | 1.17 ± 4.9e-2 | 0.22± 1.6e-2 | 1.4 ±5.1e-2 | 2.23±3.2e-1 | 2.14±5.5e-1 | **0.21±1.2e-2** | 0.89±1.4e-2 |
| ohscal | 2.71±3.5e-2 | 1.76 ± 8.8e-2 | 0.64 ±3e-2 | 3.1± 2e-2 | 8.53±6.6e-1 | 3.33±2.6e-1 | **0.59±4.8e-2** | 0.72±2.7e-2 |
| la12 | 5.06 ± 1.4e-1 | 2.26± 1e-1 | **0.46 ±2.1e-2** | 33.6 ± 1.3e-3 | 5.63±3.1e-1 | 4.50±2.7e-1 | **0.88±8.4e-2** | 4.23±1.2e-2 |

### 3.5.3 Application on region-of-interest analysis for brain image data

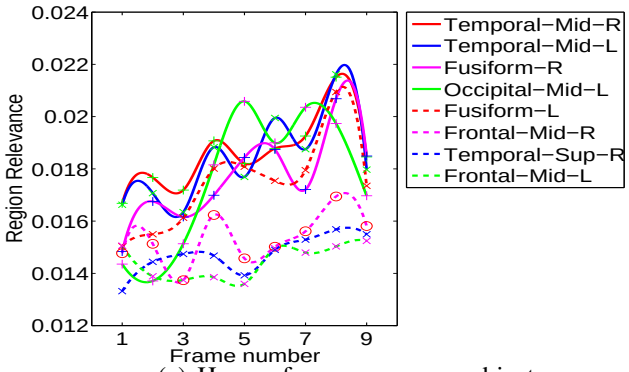
Finally, we apply our algorithm to carry out Region-of-Interest (ROI) analysis on a brain image data. The data is collected through an fMRI scan of $28$ subjects that were exposed to three types of stimuli: human face, Chinese character, and common object. Every subject was stimulated by $24$ objects, $8$ for each stimuli type. To each stimulus, a subject was exposed $9$ times. The whole brain area was divided into $31285$ voxels and the activities of $31285$ voxels at each time were recorded. Therefore the datasets, for each stimulus, are of size $28$ by $31285 \times 9$, which are very high dimensional. The voxels can be further divided into disjoint groups, named ROIs. An ROI defines a specific region of the brain. In this data, we have $116$ ROIs defined in a template, which describes the coordinates of voxels and their mappings to each ROI.

We use spike-and-slab models to determine ROIs that are relevant to face recognition and Chinese character recognition. Specifically, we use voxels as features to predict whether the subject is doing face recognition or looking at common objects. This is a binary classification problem, so we can use spike-and-slab models and apply our inference algorithm to select the related voxels. Then based on the selected voxels, we can determine the related ROIs: we calculate the $L_2$ norm of the weight vector of voxels in a ROI, which is named by relevance weight, to evaluate the relatedness of the ROI to the task; ROIs with biggest relevance weights are considered to be most relevant. Similarly, we construct another binary classification problem to determine related ROIs for Chinese character recognition.

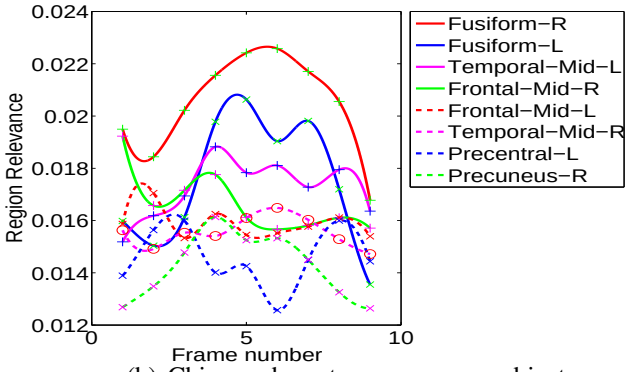
Fig 3.3 shows the top $8$ ROIs selected by our algorithm FLAS as the most discriminant regions in the brain to differentiate between human face stimuli *vs.* base (common objects) or Chinese character stimuli *vs.* base (common objects). Several research studies have shown highly similar activated regions between human face stimuli and Chinese character stimuli due to the similar properties of these two tasks, such as omni-presentation, expertise from childhood and upright orientation [126, 127]. Consistently, some common activated regions are selected in both stimuli by our model. Specifically, we select middle

temporal gyrus, fusiform and frontal region as the most relevant areas in both human face and Chinese character stimuli. These results are supported by several references. For middle temporal gyrus, it shows a connection with both face recognition and word meaning accessing [128]. The lesion in the middle temporal gyrus might cause alexia and agraphia for Kanji characters [129]. For fusiform, Liu *et al.*[126, 127] conducted a similar fMRI research stimulated by both human face and Chinese character. They showed that both face and Chinese character stimuli activated bilateral fusiform with great similarity, especially in the right hemisphere, which was consistent with our selection of both fusiform-R and fusiform-L as the most relevant, and fusiform-R showing a stronger relevance than fusiform-L, on average, in both cases. For the frontal region, Liu *et al.* demonstrated that it was highly activated in both human face and Chinese character stimuli [127]. Also, Tan *et al.* [130] studied the activated brain regions by the precise and vague meaning of Chinese characters. Their fMRI results showed that the left frontal regions were much more strongly activated than the right frontal regions [130]. Both Liu and Tan's results suggested the frontal regions to be related to human face recognition and Chinese character accessing. In addition to the common regions, our results also identify specialized regions for either human face recognition or Chinese characters stimuli. For instance, the occipital region is selected in human face stimuli. An important region called occipital face area (OFA) is located in the occipital region. OFA is mainly in charge of representing face parts and coordinates with fusiform face area (FFA) to perceive human faces [131]. In contrast, precuneus-R and precentral-L are selected as discriminant regions only in Chinese character stimuli, which is consistent with the findings of [130].

In addition, we compare the prediction accuracy of FLAS with capped $L_1$ in the two classification problems.The average error rates and the standard error for a $5$-fold cross validation are $\{$FLAS: **0.1900**$\pm$**0.0153**, capped $L_1$ :$0.1959 \pm 0.0158\}$, and $\{$FLAS: **0.2113**$\pm$ **0.0204**, capped $L_1$ :$0.3832 \pm 0.011\}$ respectively. Improved performance is evident from the results.




(a) Human face *vs.* common object




(b) Chinese character *vs.* common object

Figure 3.3.: Relevance weights of the top eight ROIs in nine time frames.

# 4 SUMMARY

Spike-and-slab priors have been very useful in sparse Bayesian learning due to their selective shrinkage effect. However, classical formulation of spike-and-slab priors does not explicitly take into account the correlation structure information between variables provided by various domains. Additionally, Bayesian inference of classical spike-and-slab models is computationally challenging due to intractable posterior distribution. Consequently, approximate inference techniques have to be employed at the cost of reduced quality of the posterior. In this dissertation we have proposed a general formulation of the spike-and-slab priors that incorporates domain based correlation structure information, and presented a principled framework for efficiently performing joint group and feature selection from a set of highly correlated variables. The dissertation also presents a Bayesian inference strategy for classical spike-and-slab models that assumes minimal structural constraints on the joint posterior, but still enjoys time complexity linear in the number of variables. The conclusion and future work are summarized as follows:

- In chapter 2 we proposed a new sparse Bayesian approach, called NaNOS, for joint network and node selection. NaNOS is a sparse hybrid Bayesian model that integrates conditional and generative components in a principled Bayesian framework [66]. The conditional component includes the generalized spike-and slab prior that induces network level sparsity via the selective shrinkage effect, and imposes structural constraints on each network through the use of graph Laplacian matrices. For the generative component, we use the classical spike and slab prior to choose relevant nodes in selected networks. This hybrid approach allows our model to combine the selective shrinkage of the classical spike-and-slab prior with the network constraint regularization effect, and hence gives our model the capability to not only select rel-

evant networks, but also induces structured sparsity, guided by domain knowledge, within selected networks.

  - In the current model, an edge between two nodes in a network represents some association or correlation between variables. We can extend this idea to directed networks where an edge not only represents a relation between the two nodes, but also carries information about the direction of influence. For example, in genomic data applications, transcription factors are the regulatory genes that have a direct influence on some other set of genes, but not the other way round. Capturing the relation between transcription factor and its set of regulated genes through an undirected graph will not be accurate. In order to accomplish this task, we can construct a Laplacian matrix for directed graphs by using the approach given in [132], and incorporate it into our modelling framework.

- In chapter 3 we proposed a new scalable sparse Bayesian inference procedure for the spike-and-slab model. Our approach achieves linear time complexity without imposing factorization constraints on the joint posterior. From a frequentist perspective, our approach has nice asymptotic consistency properties for linear regression. Our alternating optimization strategy for the joint MAP estimation also possesses convergence guarantees. Additionally, it provides uncertainty quantification as a Bayesian method. To some extent, we can view it as a hybrid of frequentist and Bayesian treatment, enjoying the benefits of both worlds. Our empirical results suggest that the nonconvex spike and slab model can yield improved selection and predictive accuracy over the classical convex $l_1$-type methods, and show better or comparable performance with respect to alternative approximate Bayesian inference methods.

  - One possible future direction is to extend the model to perform selection at the group level.

  - Secondly, we can further enhance the computational complexity of our joint optimization approaches by parallelizing the closed form updates of the GIST algorithm.

- FLAS method addresses the computational bottleneck, but since it employs the classical spike-and-slab prior, it does not incorporate domain based information about the correlation structure between variables into the learning process. Another possible direction for future work is to integrate NaNOS approach with the FLAS framework. A combination of these two approaches will lead to a very efficient technique for spare Bayesian learning.

# BIBLIOGRAPHY

# BIBLIOGRAPHY

[1] Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *arXiv preprint arXiv:1411.3230*, 2014.

[2] Dorothy Wrinch and Harold Jeffreys. On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 42(249):369–390, 1921.

[3] C. L. Mallow. Choosing variables in a linear regression: A graphical aid. In *Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas*, 1964.

[4] Colin L. Mallows. Choosing a subset regression. In *Technometrics*, volume 9, page 190. American Statistical Association, 1967.

[5] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*. Academiai Kiado, 1973.

[6] Ronald R. Hocking. A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[7] Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[8] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[9] Gideon Schwarz *et al*. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[10] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[11] Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[12] Yagyensh Chandra Pati, Ramin Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *The 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE, 1993.

[13] David L. Donoho and Jain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[14] Shane F. Cotter, J. Adler, R. Durga Rao, and Kenneth Kreutz-Delgado. Forward sequential algorithms for best basis selection. In *IEE Proceedings on Vision, Image and Signal Processing*, volume 146, pages 235–244. IEE, 1999.

[15] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[16] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[17] Emmanuel. J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[18] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[19] Jon F. Claerbout and Francis Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

[20] Howard L. Taylor, Stephen C. Banks, and John F. McCoy. Deconvolution with the $l_1$ norm. *Geophysics*, 44(1):39–52, 1979.

[21] Bruno A. Olshausen *et al*. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[22] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[23] Michael S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[24] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[25] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *IEEE 12th International Conference on Computer Vision*, pages 2272–2279. IEEE, 2009.

[26] Laurent El Ghaoui, Guan-Cheng Li, Viet-An Duong, Vu Pham, Ashok N. Srivastava, and Kanishka Bhaduri. Sparse machine learning methods for understanding large text corpora. In *Conference on Intelligent Data Understanding*, pages 159–173, 2011.

[27] Yash Deshpande and Andrea Montanari. Sparse pca via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.

[28] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[29] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[30] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4):1733, 2009.

[31] Jim E. Griffin and Philip J. Brown. Hierarchical sparsity priors for regression models. *arXiv preprint arXiv:1307.5231*, 2013.

[32] Jim E. Griffin and Philip J. Brown *et al*. Some priors for sparse regression modelling. *Bayesian Analysis*, 8(3):691–702, 2013.

[33] Jason Palmer, Kenneth Kreutz-Delgado, Bhaskar D. Rao, and David P. Wipf. Variational EM algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*, pages 1059–1066, 2005.

[34] Xianghang Liu, Xinhua Zhang, and Tibério Caetano. Bayesian models for structured sparse estimation via set cover prior. In *Machine Learning and Knowledge Discovery in Databases*, pages 273–289. Springer, 2014.

[35] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[36] A. C. Faul and M. E. Tipping. Analysis of sparse Bayesian learning. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2001.

[37] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In *International Conference on Artificial Intelligence and Statistics*, pages 73–80, 2009.

[38] Jim E. Griffin and Philip J. Brown *et al*. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[39] Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[40] Berwin A. Turlach, William. N. Venables, and Stephen. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

[41] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[42] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[43] Junzhou Huang and Tong Zhang *et al*. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[44] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[45] S. Canu and Y. Grandvalet. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. *Advances in Neural Information Processing Systems*, page 445, 1999.

[46] Sergey Bakin *et al*. Adaptive regression and model selection in data mining problems. 1999.

[47] Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Pierre Dupont. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research*, 14(1):1891–1945, 2013.

[48] Shihao Ji, David Dunson, and Lawrence Carin. Multitask compressive sensing. *IEEE Transactions on Signal Processing*, 57(1):92–106, 2009.

[49] Sudhir Raman, Thomas J. Fuchs, Peter J. Wild, Edgar Dahl, and Volker Roth. The Bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 881–888. ACM, 2009.

[50] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

[51] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.

[52] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011.

[53] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *The Journal of Machine Learning Research*, 12:3371–3412, 2011.

[54] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[55] Jerome M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

[56] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomics data. *Bioinformatics*, 24(9):1175–1182, December 2008.

[57] Fan Li and Nancy R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491), 2010.

[58] F. B. Lempers. *Posterior probabilities of alternative linear models: Some theoretical considerations and empirical experiments*. PhD thesis, Rotterdam University, 1971.

[59] Toby J. Mitchell and John J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[60] Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Alberto Suárez. Expectation propagation for microarray data classification. *Pattern Recognition Letters*, 31(12):1618–1626, 2010.

[61] José Miguel Hernández-Lobato. *Balancing flexibility and robustness in machine learning semi-parametric methods and sparse linear models*. PhD thesis, Autonomous University of Madrid, 2010.

[62] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

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

[64] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *The 30th International Conference on Machine Learning*, pages 37–45, 2013.

[65] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[66] J. A. Lasserre *et al*. Principled hybrids of generative and discriminative models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 87–94, June 2006.

[67] Edward I. George and Robert E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

[68] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[69] Tso-Jung Yen. A majorization–minimization approach to variable selection using spike and slab priors. *The Annals of Statistics*, 39(3):1748–1775, 2011.

[70] Tommi S. Jaakkola and Michael I. Jordan. Bayesian parameter estimation through varational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[71] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2007.

[72] Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52(12):5277–5286, 2008.

[73] Zhi Wei and Hongzhe Li. A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, May 2007.

[74] Zhi Wei and Hongzhe Li. A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Annals of Applied Statistics*, 2(1):408–429, 2008.

[75] Francesco C. Stingo and Marina Vannucci. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, 27(4):495–501, February 2010.

[76] Francesco C. Stingo and Yian A. Chen *et al*. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics*, 5(3):1978–2002, November 2011.

[77] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(B):267–288, 1996.

[78] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, March 2005.

[79] Laurent Jacob *et al*. Group lasso with overlap and graph lasso. In *Proceedings of the 26th International Conference on Machine Learning*, pages 433–440, New York, 2009.

[80] Vamsi K. Mootha *et al*. PGC-1$\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, June 2003.

[81] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, and Eric S. Lander *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[82] Andreas Rosenwald *et al*. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma (DLBCL). *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[83] N. Ancona *et al*. On the statistical assessment of classifiers using DNA microarray data. *BioMedCentral Bioinformatics*, 7(387), May 2006.

[84] L. Badea *et al*. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology*, 55(88):2016–2027, November 2008.

[85] K. A. Cycon *et al*. Alterations in CIITA constitute a common mechanism accounting for downregulation of MHC class II expression in diffuse large B-cell lymphoma (DLBCL). *Experimental Hematology*, 37(2):184–194, February 2009.

[86] L. Amiot *et al*. Loss of HLA molecules in B lymphomas is associated with an aggressive clinical course. *British Journal of Haematology*, 100(4):655–663, March 1998.

[87] S. Dupire and B. Coiffier. Targeted treatment and new agents in diffuse large B cell lymphoma (DLBCL). *International Journal of Hematology*, 92(1):12–24, July 2010.

[88] S. P. Lee *et al*. Clinicopathologic characteristics of CD99-positive diffuse large B-cell lymphoma. *Acta Haematologica.*, 125(3):167–174, December 2011.

[89] M. J. Terol *et al*. Expression of beta-integrin adhesion molecules in non-Hodgkin's lymphoma: Correlation with clinical and evolutive features. *Journal of Clinical Oncology*, 17(6):1869–1875, June 1999.

[90] Antje Menssen *et al*. c-MYC delays prometaphase by direct transactivation of MAD2 and BubR1: Identification of mechanisms underlying c-MYC-induced DNA damage and chromosomal instability. *Cell Cycle*, 6(3):339–352, February 2007.

[91] Q. S. Wang *et al*. Altered expression of cyclin D1 and cyclin-dependent kinase 4 in azoxymethane-induced mouse colon tumorigenesis. *Carcinogenesis*, 19(11):2001–2006, November 1998.

[92] K. Vermeulen *et al*. The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer. *Cell Proliferation*, 36(3):131–149, June 2003.

[93] C. Giaginis *et al*. Clinical significance of MCM-2 and MCM-5 expression in colon cancer: Association with clinicopathological parameters and tumor proliferative capacity. *Digestive Diseases and Sciences*, 54(2):282–291, Feburary 2009.

[94] A. Rizzo *et al*. Intestinal inflammation and colorectal cancer: A double-edged sword? *World Journal of Gastroenterology*, 17(26):3092–3100, July 2011.

[95] A. Chalkias *et al*. Patients with colorectal cancer are characterized by increased concentration of fecal hb-hp complex, myeloperoxidase, and secretory IgA. *American Journal of Clinical Oncology*, 34(6):561–566, December 2011.

[96] N. Sakai *et al*. CXCR4/CXCL12 expression profile is associated with tumor microenvironment and clinical outcome of liver metastases of colorectal cancer. *Clinical and Experimental Metastasis*, 29(2):101–110, Feburary 2012.

[97] Y. Toiyama *et al*. Evaluation of CXCL10 as a novel serum marker for predicting liver metastasis and prognosis in colorectal cancer. *International Journal of Oncology*, 40(2):560–566, February 2012.

[98] Y. Toiyama *et al*. Loss of tissue expression of interleukin-10 promotes the disease progression of colorectal carcinoma. *Surgery Today*, 40(1):46–53, December 2010.

[99] S. B. Krantz *et al*. Contribution of epithelial-to-mesenchymal transition and cancer stem cells to pancreatic cancer progression. *Journal of Surgical Research*, 173(1):105–112, March 2012.

[100] K. J. Gordon *et al*. Bone morphogenetic proteins induce pancreatic cancer cell invasiveness through a smad1-dependent mechanism that involves matrix metalloproteinase-2. *Carcinogenesis*, 30(2):238–248, February 2009.

[101] M. A. Shields *et al*. Biochemical role of the collagen-rich tumour microenvironment in pancreatic cancer progression. *Biochemical Journal*, 441(2):541–552, January 2012.

[102] R. J. Weinel *et al*. Expression and function of VLA-$\alpha_2$, -$\alpha_3$, -$\alpha_5$ and -$alpha_6$-integrin receptors in pancreatic carcinoma. *International Journal of Cancer*, 52(5):827–833, November 1992.

[103] S. Keleg *et al*. Invasion and metastasis in pancreatic cancer. *Molecular Cancer*, 2(14), January 2003.

[104] K. Kameda *et al*. Expression of highly polysialylated neural cell adhesion molecule in pancreatic cancer neural invasive lesion. *Cancer Letters*, 137(2):201–207, April 1999.

[105] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[106] Emmanuel J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians: Invited Lectures*, pages 1433–1452, 2006.

[107] Michalis K. Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 2339–2347, 2011.

[108] Amr Ahmed and Eric P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

[109] Shakir Mohamed, Katherine Heller, and Zoubin Ghahramani. Evaluating Bayesian and $L_1$ approaches to sparse unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2011.

[110] Thomas P. Minka. Deriving quadrature rules from Gaussian processes. Technical report, Statistics Department, Carnegie Mellon University, 2000.

[111] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[112] Hemant Ishwaran and J. Sunil Rao. Geometry and properties of generalized ridge regression in high dimensions. *Contemporary Mathematics*, 622:81–93, 2014.

[113] James C. Bezdek and Richard J. Hathaway. Convergence of alternating optimization. *Neural, Parallel and Scientific Computations*, 11(4):351–368, 2003.

[114] Max A. Woodbury. Inverting modified matrices. *Memorandum Report*, 42:106, 1950.

[115] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble nystrom method. In *Advances in Neural Information Processing Systems*, pages 1060–1068, 2009.

[116] Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability of transductive regression algorithms. In *Proceedings of the 25th International Conference on Machine Learning*, pages 176–183. ACM, 2008.

[117] Nicholas Francis Arcolano. *Approximation of positive semidefinite matrices using the Nystrom method*. PhD thesis, Harvard University, 2011.

[118] Constantine Bekas, Alessandro Curioni, and I. Fedulova. Low cost high performance uncertainty quantification. In *Proceedings of the 2nd Workshop on High Performance Computational Finance*, page 8. ACM, 2009.

[119] Peter Carbonetto and Matthew Stephens *et al*. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108, 2012.

[120] Tom Minka *et al*. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

[121] Daniel Hernández-Lobato, José Miguel Hernández-Lobato, Thibault Helleputte, and Pierre Dupont. Expectation propagation for Bayesian multi-task feature selection. In *Machine Learning and Knowledge Discovery in Databases*, pages 522–537. Springer, 2010.

[122] Shandian Zhe, Syed A. Z. Naqvi, Yifan Yang, and Yuan Qi. Joint network and node selection for pathway-based genomic data analysis. *Bioinformatics*, 2013.

[123] Y. Qi, T. S. Jaakkola, and D. K. Gifford. Approximate expectation propagation for Bayesian inference on large-scale problems. Technical report, MIT Computer Science and Artificial Intelligence Laboratory, October 2005.

[124] Todd E. Scheetz *et al*. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

[125] Shimon Kogan *et al*. Predicting risk from financial reports with regression. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 272–280, 2009.

[126] Jiangang Liu, Jie Tian, Kang Lee, and Jun Li. A study on neural mechanism of face processing based on functional MRI. *Progress in Natural Science*, 18(2):201–207, 2008.

[127] Jiangang Liu, Jie Tian, Jun Li, Qiyong Gong, and Kang Lee. Similarities in neural activations of face and chinese character discrimination. *Neuroreport*, 20(3):273–277, 2009.

[128] Daniel J. Acheson and Peter Hagoort. Stimulating the brain's language network: Syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. *Journal of Cognitive Neuroscience*, 25(10):1664–1677, 2013.

[129] Yasuhisa Sakurai, Imari Mimura, and Toru Mannen. Agraphia for kanji resulting from a left posterior middle temporal gyrus lesion. *Behavioural Neurology*, 19(3):93–106, 2008.

[130] Li Hai Tan, John A. Spinks, Jia-Hong Gao, Ho-Ling Liu, Charles A. Perfetti, Jinhu Xiong, Kathryn A. Stofer, Yonglin Pu, Yijun Liu, and Peter T. Fox. Brain activation in the processing of chinese characters and words: A functional MRI study. *Human Brain Mapping*, 10(1):16–27, 2000.

[131] David Pitcher, Vincent Walsh, and Bradley Duchaine. The role of the occipital face area in the cortical face perception network. *Experimental Brain Research*, 209(4):481–493, 2011.

[132] Fan Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

VITA

# VITA

Syed Abbas Zilqurnain Naqvi was born in Pakistan. He attended the University of Engineering and Technology Lahore from 2003 to 2007, where he was awarded a bachelor's degree in electrical engineering. He worked as a lecturer at the University of Engineering and Technology Lahore from 2007 to 2009. He started his PhD studies in the Department of Computer Science at Purdue University in August 2009. His PhD research topic was Bayesian machine learning with emphasis on sparse learning using spike-and-slab priors.