# Fast Laplace Approximation for Sparse Bayesian Spike and Slab Models

**Syed Abbas Z. Naqvi,**[1] **Shandian Zhe,**[1] **Yuan Qi,**[1] **Yifan Yang,**[2] and **Jieping Ye**[3]
[1]Department of Computer Science, Purdue University
[2]Department of Biology, Purdue University
[3]Department of EECS, University of Michigan, Ann Arbor

## Abstract

We consider the application of Bayesian spike-and-slab models in high-dimensional feature selection problems. To do so, we propose a simple yet effective fast approximate Bayesian inference algorithm based on Laplace's method. We exploit two efficient optimization methods, GIST [Gong *et al.*, 2013] and L-BFGS [Nocedal, 1980], to obtain the mode of the posterior distribution. Then we propose an ensemble Nyström based approach to calculate the diagonal of the inverse Hessian over the mode to obtain the approximate posterior marginals in $O(knp)$ time, $k \ll p$. Furthermore, we provide the theoretical analysis about the estimation consistency and approximation error bounds. With the posterior marginals of the 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. Our method 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. Simulation shows that our method estimates better or comparable selection probabilities and indicator variables than alternative approximate inference methods such as VB and EP, but with less running time. Extensive experiments on large real datasets demonstrate that our method often improves prediction accuracy over Bayesian automatic relevance determination, EP, and frequentist $L_1$ type methods.

## 1 Introduction

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 [Zou and Hastie, 2005; Tibshirani, 1996], 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[Candès, 2006], multi task learning [Titsias and Lázaro-Gredilla, 2011], and time-varying network reconstruction [Ahmed and Xing, 2009].

Recently there has been a shift from convex to nonconvex regularization approaches in the machine learning community. Specifically, within the Bayesian context, the spike-and-slab prior has been the focus of attention due to its selective shrinkage property. In this paper, 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 [Mitchell and Beauchamp, 1988] converge slowly with tens of thousands of features in data; and the variational Bayes (VB) and expectation propagation (EP) approaches [Hernández-Lobato *et al.*, 2010a; Hernández-Lobato, 2010; Hernández-Lobato *et al.*, 2010b] 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 (FLAS) algorithm. 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, in Section 3, 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 popular limited-memory BFGS (L-BFGS) [Nocedal, 1980] and the recently developed GIST method [Gong *et al.*, 2013]. Specifically, we use L-BFGS to obtain the MAP estimation for the marginalized model, and we use an alternating optimization strategy based on GIST [Gong *et al.*, 2013] for the joint model, with convergence guarantees for both regression and classification, and possessing oracle properties for the regression case. Then we propose an ensemble Nyström based 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 re-

spectively, and $k \ll p$. 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 cost linear in $p$. Detailed discussion on the related work is given in Section 4.

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 (Section 5). On large real benchmark datasets, our methods often achieve improved prediction accuracy with a comparable speed.

## 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: $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \tau) = \prod_{i=1}^n \mathcal{N}(t_i|\mathbf{x}_i^\top \mathbf{w}, \tau^{-1})$ where $\mathbf{w}$ are regression weights, and $\tau$ is the precision parameter; for classification, the logistic likelihood is used: $p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n \sigma(\mathbf{x}_i^\top \mathbf{w})^{t_i}[1 - \sigma(\mathbf{x}_i^\top \mathbf{w})]^{1-t_i}$ 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 [Ishwaran and Rao, 2005] over $\mathbf{w}$ is assigned:

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

$$p(z_j = 1|s_j) = s_j \quad (1 \le j \le p) \quad (2)$$

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 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 [Minka, 2000] to estimate the selection probability $s_j$ and indicator variable $z_j$.

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

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

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 [Nocedal, 1980] because of its low computational and memory cost, and due to the nonconvexity of the spike-and-slab model. 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 per iteration is linear in $p$.

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

$$\Big[\frac{\mathrm{d}\mathcal{F}}{\mathrm{d}\mathbf{w}}\Big]_j = \Big[\frac{\mathrm{d}L(\mathbf{w})}{\mathrm{d}\mathbf{w}}\Big]_j + \frac{r_0 + r_1 g(w_j)}{r_0 r_1 (1 + g(w_j))} w_j \quad (3)$$

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.

**Optimization of the joint model**
For the second approach, denoted by FLAS*, 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 first approach because by marginalizing out $\mathbf{s}$, it essentially takes all possible values of $\mathbf{s}$ into account. But the second approach can provide a more pronounced selective shrinkage effect than the first approach. We use an alternating optimization (AO) approach for both regression and classification, and employ GIST [Gong *et al.*, 2013] for finding the minimizer of $\mathbf{w}$ during the AO iterations.

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

$$\min_{\mathbf{w},\mathbf{s}} \mathcal{F}(\mathbf{w}, \mathbf{s}) = \min_{\mathbf{w}} L(\mathbf{w}) - \min_{\mathbf{s}} R(\mathbf{w}, \mathbf{s}) \quad (4)$$

where $R(\mathbf{w}, \mathbf{s}) = \sum_{j=1}^p R_j(w_j, s_j)$ and $R_j(w_j, s_j) = \log\big(s_j \mathcal{N}(w_j|0, r_1) + (1 - s_j)\mathcal{N}(w_j|0, r_0)\big)$.

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 a minimum at $s_j = 1$ if $|w_j| \geq a$, and $s_j = 0$ otherwise, $a = \sqrt{\left(\frac{2r_0 r_1}{r_1 - r_0}\right) \log \sqrt{\frac{r_1}{r_0}}}$. Given $\mathbf{s}$, the optimization of $\mathbf{w}$ has a closed form solution for regression that is a special case of generalized ridge regression [Hoerl and Kennard, 1970]:

$$\mathbf{w}_{opt} = (\tau \mathbf{X}^\top \mathbf{X} + \text{diag}(\text{d}))^{-1} \tau \mathbf{X}^\top \mathbf{t} \tag{5}$$

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

The update of $\mathbf{w}$ has a time complexity of $O(p^3)$. This is prohibitively expensive at higher dimensions. Therefore, we employ GIST to minimize $\mathbf{w}$. Since the problem is strictly convex, GIST is guaranteed to converge to the unique minimum (closed form solution), but with cost per iteration $O(np)$[Gong *et al.*, 2013]. 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. Our AO scheme also satisfies the Existence and Uniqueness (EU) assumption, and hence converges to a joint local minimum [Bezdek and Hathaway, 2003].

**Estimation, Selection and Sign consistency for regression**
Using the approach given in [Yen, 2011] and [Zou and Zhang, 2009], we will prove asymptotic consistency properties for the AO estimator. 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. We will use the following assumptions for our analysis:

**Assumption 1 [Yen, 2011].** 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:

$$0 \leq c_1 < \lambda_{min}(C_{SS}) \leq \lambda_{max}(C_{SS}) \leq c_2 < \infty \tag{6}$$

**Assumption 2.** For parameters $r_1$, $r_0$ and $\tau$, assume that they are fixed, and $0 < r_1, r_0, \tau < \infty$;

**Assumption 3 [Yen, 2011].** Assume a finite constant $c_3 > 0$ such that $(w_j^*)^2 < c_3$ for all $j = 1, , \cdots p$.

**Assumption 4.** 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 5.** 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 $0 < \delta < M$

Assumption 1 enforces positive definiteness of the sample covariance matrix. This assumption is reasonable for large sample sizes as the sample covariance matrix is of full rank. Assumption 2 simply indicates that we know the true value of the hyper parameters. However, our results are also valid for bounded support. Therefore, assuming the parameters as fixed and known is not necessary. Assumption 3 is needed to make sure that the true weight vector does not grow without bound. This is required because in theorem 1 the true weight vector changes with sample size. Assumption 4 can find its applications in situations where the user has control over the design of matrix $\mathbf{X}$, for example compressed sensing. Assumption 5 enforces absolute shrinkage, and hence selection and sign consistency.

**Theorem 1.** Given that 1, 2, and 3 are satisfied and $p \propto n^\alpha$ with $\alpha > 0$, then $P(||\mathbf{w}_{opt} - \mathbf{w}^*||^2 > \xi_n) \leq c_0 \exp\{-\log(n^{1-\alpha}\xi_n)\}$ for some positive finite constants $c_0$ 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.

**Theorem 2.** Under assumptions 2, 4, and 5 $\mathbf{w}_{opt} \to \mathbf{w}_*$ as $n \to \infty$ with $p$ fixed. 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. Then, $\mathbf{w}_{opt}^c$ will be sign and selection consistent as $n \to \infty$ with $p$ fixed.
(All proofs are omitted due to space limitations)

## 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,

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

where $v_j = -\frac{d^2 \log(p(w_j))}{dw_j^2}\Big|_{w_j = \tilde{w}_j}$. 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

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

Applying Woodbury matrix identity [Woodbury, 1950], we can readily reduce the complexity to $O(nkp)$:

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

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

Using Nyström approach to estimate the diagonal of inverse Hession 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 Hession is then obtained by

$$\mathbf{H}^{-1}(j,j) \approx \frac{1}{d}\sum_{r=1}^{d} \tilde{\mathbf{H}}_r^{-1}(j,j). \tag{7}$$

Using Taylor expansion and error bounds of Nyström approximations[Kumar *et al.*, 2009], we can prove that the proposed ensemble approach can have a smaller estimation error. This is expressed in the following theorems.

**Theorem 3.** 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 high probability,

$$|\mathbf{H}^{-1}(j,j) - \tilde{\mathbf{H}}^{-1}(j,j)| \leq L \cdot D_0 \tag{8}$$

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 [Kumar *et al.*, 2009].

**Theorem 4.** 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 high probability,

$$|\mathbf{H}^{-1}(j,j) - \frac{1}{d}\sum_{r=1}^{d} \tilde{\mathbf{H}}_r^{-1}(j,j)| \leq L \cdot D_1 \tag{9}$$

where $D_1$ is the error bound for ensemble Nyström based on Frobenius norm [Kumar *et al.*, 2009]. Because $D_1 < D_0$ (see [Kumar *et al.*, 2009]), the ensemble approach for the diagonal entry estimation of $\mathbf{H}^{-1}$ has a smaller error bound.

**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 and 4. (All proofs are omitted due to space limitations)

### 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,

$$p(s_j|w_j) = \frac{s_j\mathcal{N}(w_j|0,r_1) + (1-s_j)\mathcal{N}(w_j|0,r_0)}{\frac{1}{2}\mathcal{N}(w_j|0,r_1) + \frac{1}{2}\mathcal{N}(w_j|0,r_0)}. \tag{10}$$

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

$$p(s_j|\mathbf{t},\mathbf{X}) = \int p(s_j|w_j)\mathcal{N}(w_j|m_j,\sigma_j^2)\mathrm{d}w_j \tag{11}$$

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

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

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

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

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

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

We apply Gauss-Hermite quadrature method [Minka, 2000] to calculate the above one dimensional integrals with high accuracy. (*e.g.,* the numerical difference from the true integration is often on the order of $10^{-4}$). The over all cost for computing the posterior mean and variance of $\mathbf{w}$, $\mathbf{s}$, and $\mathbf{z}$ is $O(dknp)$, $d, k \ll p$. The linear cost makes our algorithm scalable for high dimensional data.

## 4 Related Work

A very closely related approach to our method proposes a MAP estimation of spike and slab models with delta spikes [Yen, 2011]. The method approximates the delta spike by a continuous bound via an elegant majorization and minimization (MM) algorithm. Consistency results for the MAP estimate are also provided. We, on the other hand, assume continuous spikes to make use of efficient continuous optimization strategies. Secondly, while the MM algorithm only focuses on the MAP estimate, we provide a full Bayesian inference strategy, and also show oracle properties for our MAP estimate. Another related method is the integrated nested Laplace approximation (INLA) [Rue *et al.*, 2009]. INLA is designed for the latent Gaussian models and is shown to be very efficient and accurate. However, Spike-and-slab priors are mixture priors and do not belong to the latent Gaussian family. Simply applying INLA to the spike-and-slab models will be computationally expensive ($O(p^3)$) due to the dense precision matrix and high dimensional feature space.

EP and VB approximations have also been developed to conduct Bayesian inference on the spike-and-slab model. In the context of multi-task learning, EP achieved a per task complexity of $O(n^2p)$ for $n < p$ (or $O(np^2)$ when $n > p$) [Hernández-Lobato, 2010]. Further, a fully factorized approximate posterior of $\mathbf{w}$ was imposed to achieve a cost of $O(np)$ with $n < p$ in the classification context

for EP [Hernández-Lobato *et al.*, 2010b]. Similarly, a cost of $O(np^2)$ was spent for the VB approximation with fully factorized posterior assumption [Titsias and Lázaro-Gredilla, 2011; Carbonetto *et al.*, 2012].

Our work differs from the above methods in that we do not impose any factorization assumption on the joint posterior. Instead, with minimal structural constraints, our method not only enjoys a linear cost in $p$, but also avoids the strong mean-field like assumption, which could hurt the inference quality [Carbonetto *et al.*, 2012].

# 5 Experiments

## 5.1 Simulation

First we examine our method in a simulation study.

**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 multi-variate 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

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

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. We also evaluate the accuracy of posterior inference by using Gibbs sampling results as a reference, and following the same simulation procedure as before, but with feature dimension $p$ set to 100.

**Competing methods.** We compare our approach with alternative approximate inference algorithms for the spike-and-slab model, including VB, EP, and MM [Yen, 2011] that only provides MAP estimation. We implement two versions of EP algorithms, where for regression, one is based on continuous spikes [Hernández-Lobato, 2010](EP) and the other is based on delta spikes (EP$^*$); for classification, we use a method similar to [Hernández-Lobato, 2010], and thus we also denote it by EP [Hernández-Lobato *et al.*, 2010a]; the other has a linear time complexity [Hernández-Lobato *et al.*, 2010b], 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: first is denoted by (VB) [Zhe *et al.*, 2013], the complexity is $O(p^3)$ but without a factorized posterior assumption over model weights; and the other is denoted by (VB$^*$) [Titsias and Lázaro-Gredilla, 2011], it uses a fully factorized posterior assumption with reduced cost $O(np^2)$. For all these methods, including Gibbs sampling, we apply the same model as in section 2 where the selection probabilities $\{s_j\}$ are not integrated out. Because the above VB and EP methods only provide point estimates of the selection probabilities $\{s_j\}$, we modify their Bayesian model by applying a prior on $\{s_j\}$, and infer their posteriors [Qi *et al.*, 2005]. 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 the initialization point. 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 1 a and e show the predictive performance of all the methods for regression and classification. FLAS* and FLAS show better performance than alternative methods, or at least comparable to them. Figures 1 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, FLAS* and FLAS achieve higher F1 scores for classification and comparable F1 score than the best alternatives in regression.

For selection uncertainty, we compare the posterior mean of $s_j$ and $z_j$ with the results of Gibbs sampling based on $100,000$ samples. We calculate the root mean square error to evaluate the difference from the ground truth and report the results in Figure 1 c, d, g, and h. It is clear that FLAS* and FLAS consistently obtain better or comparable uncertainty estimation to competing methods. This confirms the inference quality of our algorithm.

## 5.2 Large Real Benchmark Data sets

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) [Rosenwald *et al.*, 2002], GSE5680 [Scheetz *et al.*, 2006], Yearprediction[4](Year), House-census[5](House), 10K corpus [Kogan *et al.*, 2009] and TIED[6]. Among the 14 datasets, the feature numbers are often at tens of thousands, while the sample sizes

[1] www-stat.stanford.edu/~tibs/glmnet-matlab

[2] www.public.asu.edu/~jye02/Software/GIST/

[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

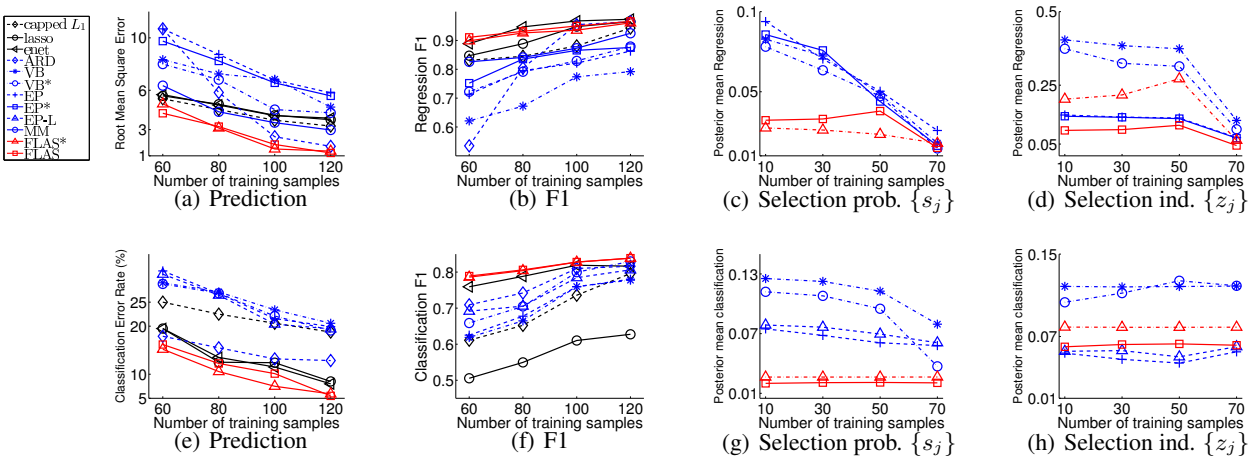


Figure 1: Simulation results, including the prediction accuracy, the F1 score of feature selection, and the root mean squared error for the posterior mean estimation of $\{s_j\}$ and $\{z_j\}$. Results are averaged over 50 runs.

Table 1: 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.

| dataset | lasso | elast net | capped $L_1$ | ARD | EP-L | FLAS* | FLAS |
|---|---|---|---|---|---|---|---|
| gse5680 | 0.107 ± 0.003 | 0.107 ± 0.003 | 0.107± 0.003 | 0.136 ± 0.005 | 0.72 ± 0.001 | 0.111 ± 0.002 | **0.089± 0.002** |
| 10k corpus | 0.382 ± 0.002 | 0.382 ± 0.002 | 0.382 ± 0.002 | 0.382 ± 0.384 | 0.385 ± 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 | 0.632 ± 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 | 0.441 ± 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 | **0.232± 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 | **1.56 ± 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 | 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 | **19.9 ± 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 | 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 | 5.55 ± 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 | **2.77 ±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 ± 0.002 | **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 | **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 | **11.04 ± 0.001** | 11.1 ± 0.001 |

are often at hundreds or thousands.

We compare our algorithms, FLAS* and FLAS, 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. 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 1 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, FLAS* or FLAS, obtain smaller root mean square errors or classification error rates. We also examine the average training time of all the methods and it turns out that our approach spends less or comparable time than the others. For example, the running time in seconds on *gse5680* and *reviews* are {lasso:2.03, elastic net:2.26, capped $L_1$:15.3, ARD:3.52, EP-L: 6.52, FLAS*:**0.15**, FLAS:0.3}, and {lasso:0.32, elastic net:0.29, capped $L_1$:2.3, ARD:26.7, EP-L:1.02, FLAS*:0.25, FLAS:**0.10**} respectively.

## 6 Conclusion

We have presented a new scalable sparse Bayesian inference method for the spike-and-slab model. From a frequentist perspective, our approach is computationally efficient, and possesses oracle properties, and from a Bayesian point of view, it quantifies selection uncertainty. Our empirical results suggest that the spike and slab model can yield improved selection and predictive accuracy over the classical $l_1$-type methods.

## Acknowledgments

This work was supported by NSF CAREER award IIS-1054903, and the Center for Science of Information, an NSF Science and Technology Center, under grant agreement CCF-0939370.

## References


[Ahmed and Xing, 2009] 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.

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

[Candès, 2006] Emmanuel J Candès. Compressive sampling. In *Proceedings oh the International Congress of Mathematicians: invited lectures*, pages 1433–1452, 2006.

[Carbonetto *et al.*, 2012] Peter Carbonetto, 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.

[Gong *et al.*, 2013] 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.

[Hernández-Lobato *et al.*, 2010a] 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.

[Hernández-Lobato *et al.*, 2010b] Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Alberto Suárez. Expectation propagation for microarray data classification. *Pattern recognition letters*, 31(12):1618–1626, 2010.

[Hernández-Lobato, 2010] José Miguel Hernández-Lobato. *Balancing flexibility and robustness in machine learning semi-parametric methods and sparse linear models*. PhD thesis, Ph. D. Thesis, Universidad Autó De Madrid, 2010.

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

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

[Kogan *et al.*, 2009] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics, 2009.

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

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

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

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

[Qi *et al.*, 2005] 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.

[Rosenwald *et al.*, 2002] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[Rue *et al.*, 2009] H**r**avard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[Scheetz *et al.*, 2006] Todd E Scheetz, Kwang-Youn A Kim, Ruth E Swiderski, Alisdair R Philp, Terry A Braun, Kevin L Knudtson, Anne M Dorrance, Gerald F DiBona, Jian Huang, Thomas L Casavant, 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.

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

[Titsias and Lázaro-Gredilla, 2011] Michalis K Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *NIPS*, volume 24, pages 2339–2347, 2011.

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

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

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

[Zou and Hastie, 2005] 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.

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