Variational Discriminant Analysis with Variable Selection

Weichang Yu\textsuperscript{1}, John T. Ormerod \textsuperscript{*1,2}, and Michael Stewart\textsuperscript{1}

\textsuperscript{1}School of Mathematics and Statistics, University of Sydney
\textsuperscript{2}ARC Centre of Excellence for Mathematical and Statistical Frontiers

Key words: Discriminant analysis, Variational Bayes approximation, Variable selection, Cake priors, Multiple hypothesis tests, Classification

Abstract

A Bayesian method that seamlessly fuses classification via discriminant analysis and hypothesis testing is developed. Building upon the original discriminant analysis classifier, modelling components are added to identify discriminative variables. A combination of cake priors and a novel form of variational Bayes we call reverse collapsed variational Bayes gives rise to variable selection that can be directly posed as a multiple hypothesis testing approach using likelihood ratio statistics. Some theoretical arguments are presented showing that Chernoff-consistency (asymptotically zero type I and type II error) is maintained across all hypotheses. We apply our method on some publicly available genomic datasets and show that our method performs well in practice. An \texttt{R} package \texttt{VaDA} has also been made available on Github.

1 Introduction and literature review

Classification is a fundamental component of machine learning that is applicable in many disciplines. A popular classification method, initially known as discriminant analysis, was first introduced by Fisher (1936) and has more recently been adapted by Dudoit et al. (2002) and Fernández-Delgado et al. (2014) to achieve consistently good performance for some high dimensional datasets. This class of methods involves a comparison of group proportions and group-conditional distributions of variables, also known as features in machine learning literature, to arrive at a classification decision rule. However, for high dimensional problems this decision rule cannot be computed when applying discriminant analysis (DA) to high dimensional data, i.e., when the number of observations, \( n \), less than number of variables, \( p \). Furthermore, the standard DA model of Fisher (1936) is not designed to identify discriminative variables. Without modification of the base DA model, its usage in high dimensional problems where identification of important variables is important is limited, e.g., Bioinformatics.

In discriminant analysis, the group-conditional distribution of variables are commonly assumed to be Gaussian. This simplifies the classification rule to a difference in Mahalanobis distances of a new observation from the group-conditional distributions. Since the MLE of the covariance matrix required to compute the Mahalanobis distance is singular in high dimension data, i.e., when \( n < p \), the classification decision rule cannot be computed. A straightforward solution to this problem is to use alternative estimators such as the Moore-Penrose inverse

\textsuperscript{*}This research was partially supported by an Australian Postgraduate Award (Yu) and the Australian Research Council Discovery Project grant DP170100654 (Ormerod).
Courrieu (2005), Chen and Feng (2014), Cai and Liu (2011) or a penalised covariance matrix. Another common solution is to utilise dimension reduction techniques such as principal components or t-distributed stochastic neighbour embeddings (t-SNEs) (van der Maaten and Hinton 2017) to project the original variables into a lower dimensional space. The high dimensionality issue has also been tackled by making the naïve Bayes assumption, i.e., the covariance matrix for the variables is assumed to be diagonal (Dudoit et al., 2002). Numerous examples of discriminant analysis models, sometimes called naïve Bayes classifiers, that have made this assumption can be found in Tibshirani et al. (2003), Pedro (2011), Witten (2011), Witten and Tibshirani (2011). While dimension reduction solutions are simple to implement, they do not necessarily address the need to identify discriminative variables commonly required in high dimensional data analysis.

An approach that addresses high dimensionality and identifies discriminative variables is to implement as a two-stage algorithm. In the first stage, a hypothesis test is performed on each of the variables to identify discriminative variables. In the second stage, variables that are identified as discriminative are retained and used to fit a DA model (e.g., Fan and Fan, 2008). Care must be taken when choosing an appropriate variable selection method as such methods can lead to inflated family-wise Type I error, false discovery rates or other multiple testing issues (Shaffer, 1995). This can be easily resolved with one of numerous remedial measures (see for example, Bonferroni, 1936; Benjamini and Hochberg, 1995; Benjamini and Daniel, 2001; Storey, 2003). A notable criterion, known as higher criticism thresholding, exhibits asymptotic optimality and good performance in several multiple testing metrics such as false discovery rate and missed detection rate under some sparsity assumptions (Donoho and Jin, 2004, 2008). This criterion has been incorporated as an option in two-stage DA algorithms such as shrinkage DA (R package: SDA) and factor-adjusted discriminant analysis (R package: FADA) (Ahdesmäki and Strimmer, 2010; Perthame et al., 2016). A modified version of the criterion, known as expanded higher criticism (EHC), has been incorporated into diagonal linear discriminant analysis (DLDA) and factor-based linear discriminant analysis (R package: HiDimDA). Although many of these algorithms utilise selection criteria that have good theoretical properties, information from the variable selection stage is lost when the variable selection and classification are done in two separate stages. For example, two variables yielding adjusted p-values of 0.001 and 0.04 may be selected, but their difference in signal strengths remain unaccounted for. This, in turn, may lead to an unnecessary loss of classification accuracy.

The loss of information can be circumvented by fusing the two stages. This fusion can be realised in penalised discriminant analysis models. In such methods a penalty function induces sparsity in the estimated discriminant vector (product of precision matrix and mean difference), that is used for both variable selection and classification. Witten and Tibshirani (2011) introduced two penalty options to the Fisher’s discriminant problem (R package: penalizedLDA) which will be further elaborated in the Section 6. Other examples may be found in Cai and Liu (2011), Shao et al. (2011) and Safo and Ahn (2016). While penalised DA models have demonstrated desirable theoretical properties and good numerical results in these papers, the results are often very sensitive to the setting of the tuning parameter of the penalty function. Usually costly cross-validation is often necessary to determine an appropriate value of this tuning parameter.

In this paper, we propose a Bayesian DA model that integrates both variable selection and classification. The model overcomes high dimensionality by adopting the naïve Bayes assumption leading to an invertible estimated covariance matrix. However, while Perthame et al. (2016) demonstrated the loss in variable selection stability and classification accuracy under this assumption, Bickel and Levina (2004) has shown the convergence of the worst-case misclassification error to the Bayes error (the classification error when the true population

2
distributions are used to derive a classification rule). By introducing variable selection parameters into our Bayesian model, we allow posterior inferences to be made on the “discriminativeness” of each variable.

Unlike two-stage methods, the Bayesian hierarchical setup of our proposed model fuses both variable selection and classification in an omnibus fashion. By doing so we avoided any loss of information between the variable selection and classification stages. The variable selection component fits naturally into a multiple hypothesis testing (MHT) paradigm. By choosing the cake priors (Ormerod et al., 2017) and approximating the computationally intensive posterior densities with a new variant of variational Bayes inference known as reverse collapsed variational Bayes (RCVB), the resultant decision rule overcomes the problems associated with MHTs such as inflated Type I errors. Since the variable selection rule depends on a set of approximate posterior probabilities, the choice of the selection threshold can also be intuitively determined. The resultant classification rule takes the form of a weighted naïve Bayes linear discriminant analysis (when variances are assumed equal). The computational cost of the algorithm is also reduced by utilising Taylor’s approximation to reduce the number of updates in the RCVB cycles. An implementation of our approach uses C++ for high performance computing.

In Section 2 we specify the model and discuss our choice of priors. Section 3 introduces our RCVB approximation. We will discuss its application to our model in Section 4. In Section 5, we state some asymptotic properties of our variable selection criterion induced by our proposed model and hence show that our variable selection rule circumvents issues with MHTs. In Section 6 we compare the performance of our proposed model with existing solutions by an application to simulated and publicly available datasets. Section 7 concludes.

2 The variational discriminant analysis model

Consider the training dataset \( \{x_i, y_i\}_{i=1}^n \) where for each \( i \) we have \( x_i = (x_{i1}, \ldots, x_{ip})^T \) as a vector of predictor-variables and \( y_i \in \{0, 1\} \) as an observed group label. In this paper, we present the model in the context of a binary classification problem but its extension to multiple groups should be possible. In addition, we have used bold-faced symbols to denote the vector of parameters across subscripts. We assume that \( y_i \) is observed for all \( i = 1, \ldots, n \). In line with machine learning terminology, we shall refer to this dataset with observed group labels as the training data.

We consider the following hierarchical model for our data set. For all \( i = 1, \ldots, n \), each \( (x_i, y_i) \) is distributed as follows. Let \( \gamma_j \in \{0, 1\}, 1 \leq j \leq p \) be a binary variables indicating whether variable \( j \) is discriminative. If \( \gamma_j = 1 \), then

\[
\begin{align*}
  x_{ij} \mid y_i, \mu_{j1}, \sigma_{j1}^2 & \sim \mathcal{N}(\mu_{j1}, \sigma_{j1}^2) \quad \text{if } y_i = 1; \\
  x_{ij} \mid y_i, \mu_{j0}, \sigma_{j1}^2 & \sim \mathcal{N}(\mu_{j0}, \sigma_{j1}^2) \quad \text{if } y_i = 0,
\end{align*}
\]

(1) and if \( \gamma_j = 0 \), then

\[
x_{ij} \mid \mu_j, \sigma_j^2 \sim \mathcal{N}(\mu_j, \sigma_j^2),
\]

(2)

where \( \mathcal{N}(m, s^2) \) denotes the Gaussian distribution with mean \( m \) and variance \( s^2 \). When \( \gamma_j = 1 \), “discriminativeness” is induced by imposing dependence between the Gaussian parameters and group label \( y_i \). We model the conditional distribution of the group labels as

\[
y_i \mid \rho_y \sim \text{Bernoulli}(\rho_y),
\]

where \( \rho_y \) denotes the probability of observing a group 1 sample from the population.
To avoid the problem of dealing with singular estimated covariance matrices, we make the naïve Bayes assumption by imposing conditional independence between the variables given the Gaussian parameters and group labels, i.e. $x_{ij} \perp x_{ik}$ for any $j \neq k$.

### 2.1 Homogeneity of group-specific variances

We have assumed that the group conditional variances are equal, i.e., $\text{Var}(x_{ij}|y_i = 1, \sigma^2_j) = \text{Var}(x_{ij}|y_i = 0, \sigma^2_j)$. Gaussian discriminant analysis models that adopt this assumption are known as linear discriminant analysis (LDA) since the resultant decision rule is linear in the variables $x$ (see Bickel and Levina, 2004; Ahdesmäki and Strimmer, 2010; Pedro, 2011; Witten and Tibshirani, 2011; Clemmensen et al., 2011; Perthame et al., 2016). Those that allow for different group conditional variances are known as quadratic discriminant analysis (QDA). When the true group conditional variances differ, it is well known that LDA performs poorly. However, LDA yields more precise estimates of variances in high dimensional setting and has less parameter estimates to compute. While a compromising solution between QDA and LDA has been proposed by Friedman (1989), we propose two variants of our model - a variational linear discriminant analysis (VQDA) and a variational quadratic discriminant analysis (VLDA), that correspond to LDA and QDA methods respectively. This allows us to study the sensitivity of the two types of models under both homogeneous and heterogeneous variance conditions.

### 2.2 Choice of priors

The choice of priors is of paramount importance in Bayesian modelling as it affects the computational course of the posterior inference and allows prior knowledge to influence results of the analysis. A natural choice of priors for $\gamma$ and $\rho_y$ uses

$$
\gamma_j \mid \rho_{\gamma} \overset{\text{iid}}{\sim} \text{Bernoulli}(\rho_{\gamma}), \quad \text{with} \quad \rho_{\gamma} \sim \text{Beta}(a_{\gamma}, b_{\gamma}),
$$

and

$$
\rho_y \sim \text{Beta}(a_y, b_y),
$$

where the random parameter $\rho_{\gamma}$ may be interpreted as the probability of including a discriminative variable in the dataset.

For the rest of this paper, we have chosen a flat prior for $\rho_y$ by assigning the hyperparameters $a_y = b_y = 1$. When considering settings for $a_{\gamma}$ and $b_{\gamma}$, a reasonable assumption is that the true model consists of a constant $p_1 < \infty$ discriminative variables, and that the probability of capturing all discriminative variables converges to 1 as $p \to \infty$. Here, we propose the setting $a_{\gamma} = 1$ and

$$
b_{\gamma} = \frac{\rho_{\gamma}^2}{\sqrt{n + 1}} \exp \left[ \frac{\kappa (n + 1)}{\log(n + 1)^r} \right] \tag{3}
$$

for some $r < 1$ and $\kappa > 0$, to induce desirable asymptotic properties in our resultant variable selection rule (see Section 5).

The choice of priors for the Gaussian parameters is a more complex issue. Most research is conducted in the absence of reliable prior knowledge and therefore diffuse priors are a popular choice. However, a diffuse prior may not retain the “diffuse” property after a model reparametrisation. Besides this property, we also require our priors to lead to model selection consistency in our variable selection rule and strong concordance with frequentist approaches (information consistency) in both variable selection and classification rules. The latter
criterion would help avoid unnecessary dilemma when users compare the inference results of our model with most other discriminant analysis models that are frequentist in approach. Recently, Ormerod et al. (2017) has proposed the cake prior which can be made diffuse under some setting of hyperparameters. For a two-sample test of Gaussian data, they have also obtained posterior Bayes factors which have good asymptotic properties. Furthermore, the posterior Bayes factor takes the form of penalized likelihood ratio statistics. The cake priors for our Gaussian parameters are as follows. For VLDA the model under $H_0$ for variable $j$ is (2) and the model under $H_1$ for variable $j$ is (1). The cake priors for these hypotheses are given by

$$\mu_j | \sigma_j^2 \sim N(0, h^{1/2} \sigma_j^2), \quad \mu_{jk} | \sigma_{jk}^2 \sim N(0, (n/n_k)h^{1/3} \sigma_{jk}^2) \quad \text{for } k = 0, 1,$$

$$\sigma_j^2 \overset{\text{iid}}{\sim} LN(0, 2h^{1/2}) \quad \text{and} \quad \sigma_{jk}^2 \overset{\text{iid}}{\sim} LN(0, 2h^{1/3}).$$

where $n_1 = \sum_{i=1}^{n} y_i$, $n_0 = n - n_1$, LN refers to the log-normal distribution and $h > 0$ is a common hyperparameter shared between $\{H_0\}_{p=1}^j$ and $\{H_1\}_{p=1}^j$.

Analogously, the VQDA model under $H_{1j}$ is

$$\begin{cases} x_{ij} | y_i, \mu_{j1}, \sigma_{j1}^2 \overset{\text{iid}}{\sim} N(\mu_{j1}, \sigma_{j1}^2) & \text{if } y_i = 1; \\ x_{ij} | y_i, \mu_{j0}, \sigma_{j0}^2 \overset{\text{iid}}{\sim} N(\mu_{j0}, \sigma_{j0}^2) & \text{if } y_i = 0, \end{cases}$$

and is equivalent to (2) under $H_{0j}$, so that the group conditional variances differ. The cake priors in this case are given by

$$\mu_j | \sigma_j^2 \sim N(0, h^{1/2} \sigma_j^2), \quad \mu_{jk} | \sigma_{jk}^2 \sim N(0, (n/n_k)h^{1/3} \sigma_{jk}^2) \quad \text{for } k = 0, 1,$$

$$\sigma_j^2 \overset{\text{iid}}{\sim} LN(0, 2h^{1/2}) \quad \text{and} \quad \sigma_{jk}^2 \overset{\text{iid}}{\sim} LN(0, 2(n/n_k)h^{1/4}) \quad \text{for } k = 0, 1.$$

The interested reader may refer to Ormerod et al. (2017) for details about the cake prior construction. Although cake priors do not lead to closed form expressions for the posterior of $\gamma$ and the group labels of new observations, we will observe in Section 4 that they yield approximate posteriors that satisfy both model selection and information consistency when implemented with a posterior inference method described in the next section.

### 3 Approximating the posterior

For most models posterior distributions are known only up to a normalising constant. While Markov chain Monte Carlo (MCMC) methods are the most widely used methods for making inferences from such posteriors, a new class of fast, deterministic algorithms known as variational Bayes approximation is gaining popularity in the computer science literature and has demonstrated comparable results at a fraction of MCMC’s computational cost in complex problems presented by Blei and Jordan (2006), Hall et al. (2011) and Luts and Ormerod (2014).

We now provide a review of variational Bayes (VB) approximation and then describe a modification of this method called reverse collapsed variational Bayes. For a more comprehensive introduction to VB the reader may refer to Ormerod and Wand (2010) and Blei et al. (2017).

#### 3.1 Variational Bayes

Given a model’s parameter $\theta$ and data $D$, the posterior density of $\theta$ may be expressed as

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)}.$$
The denominator is known as the marginal likelihood of the data and involves the evaluation of an integral or a sum that may be be computationally infeasible. The endpoint of the variational Bayes (VB) algorithm is to choose an approximation $q(\theta)$ to the posterior density $p(\theta|D)$ from a set of functions $\mathcal{F}$ that are more computationally feasible by minimising the Kullbeck-Leibler (KL) divergence

$$D_{KL}(q||p) = \mathbb{E}_q\left[\log\left\{\frac{q(\theta)}{p(\theta|D)}\right\}\right], \quad (5)$$

with respect to $q(\theta)$, where $\mathbb{E}_q$ refers to the expectation with respect to $q(\theta)$. Since

$$\mathbb{E}_q\left[\log\left\{\frac{q(\theta)}{p(\theta|D)}\right\}\right] = \mathbb{E}_{\theta}\left[\log\left\{\frac{q(\theta)p(\theta)}{p(\theta|D)}\right\}\right],$$

$$= \log p(D) - \mathbb{E}_q\left[\log\left\{\frac{p(\theta, D)}{q(\theta)}\right\}\right],$$

minimising (5) is also equivalent to maximising the Expected Lower Bound Order (ELBO) given by

$$\text{ELBO}_{\text{VB}} = \mathbb{E}_q\left[\log\left\{\frac{p(\theta, D)}{q(\theta)}\right\}\right]. \quad (6)$$

One common choice for $\mathcal{F}$ is the set of mean field functions $\mathcal{F} = \{q(\theta) \mid q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)\}$, where $\{\theta_j\}_{j=1}^{J}$ is a partition of $\theta$. This choice of $\mathcal{F}$ leads to the optimal approximating densities

$$q_j(\theta_j) \propto \exp\left[\mathbb{E}_{-q_j}\left\{\log p(\theta, D)\right\}\right], \quad \forall \ j = 1, \ldots, J, \quad (7)$$

as shown, for example, in Ormerod and Wand (2010). The notation $\mathbb{E}_{-q_j}$ refers to expectation with respect to $\prod_{l \neq j} q_l(\theta_l)$. The parameters of each $q_j(\theta_j)$ is updated iteratively with the batch coordinate-ascent variational inference algorithm as described in Zhang and Zhou (2017).

### 3.2 Reverse collapsed variational Bayes

A variant of VB called collapsed variational Bayes (CVB) was first coined in the context of latent Dirichlet allocation (Teh et al. 2007). The key idea behind these methods is to collapse, or marginalise over a subset of parameters before applying VB methodology. The CVB approach of Teh et al. (2007) results in a better approximation in comparison to VB, but is no different conceptually since it can be though of simply as applying VB to marginalized likelihood. We now introduce a reverse collapsed variational Bayes (RCVB) which can result in a different approximation to CVB. In this form the lower bound is calculated by using VB for one set of parameters, and the remaining set of parameters are collapsed over by marginalization.

To fix ideas, let $\theta_1$ and $\theta_2$ be a partition of the parameter vector $\theta$. Suppose we have a density $q_2(\theta_2)$ such that the quantity

$$\log p(D, \theta_1) \equiv \mathbb{E}_{q_2(\theta_2)}\log\left\{\frac{p(D, \theta_1, \theta_2)}{q_2(\theta_2)}\right\},$$

can be evaluated analytically for all $\theta_1$. Using Jensen’s inequality it is easy to show that

$$\log p(D, \theta_1) \geq \log p(D, \theta_1)$$

for all $\theta_1$. If we use $p(D, \theta_1)$ as an approximation of $\log p(D, \theta_1)$ we can then marginalise over $\theta_1$ to obtain the following lower bound on the log marginal likelihood:

$$\text{ELBO}_{\text{RCVB}} \equiv \log \int p(D, \theta_1) d\theta_1,$$

$$= \log \int \exp \left[\mathbb{E}_{q_2}\log\left\{\frac{p(D, \theta_1, \theta_2)}{q_2(\theta_2)}\right\}\right] d\theta_1,$$
where the integral can be interchanged with a sum when appropriate. Since \( \log p(D, \theta_1) \geq \log p(D, \theta_1) \) we have \( \text{ELBO}_{\text{RCVB}} \geq \text{ELBO}_{\text{VB}} \) and hence RCVB yields an approximation of the log marginal likelihood that is more accurate than VB.

Now suppose that we have partitioned \( \theta \) into three sets of parameters \( \theta_1, \theta_2 \) and \( \theta_3 \), and we want to apply VB-type approximations to \( \theta_2 \) and \( \theta_3 \) while integrating out \( \theta_1 \) analytically. The iterations for RCVB algorithms would be to repeat the following two steps for an arbitrarily large number of iterations until convergence:

1. \( q_2(\theta_2) \propto \int \exp \left[ \mathbb{E}_{q_3} \log \left( \frac{p(D, \theta_1, \theta_2, \theta_3)}{q_3(\theta_3)} \right) \right] d\theta_1. \)

2. \( q_3(\theta_3) \propto \int \exp \left[ \mathbb{E}_{q_2} \log \left( \frac{p(D, \theta_1, \theta_2, \theta_3)}{q_2(\theta_2)} \right) \right] d\theta_1. \)

Similarly to VB, the parameters of each \( q_j(\theta_j) \) is updated iteratively with the batch coordinate-ascent variational inference algorithm.

The above ideas are easily generalizable to an arbitrary partition size. In the next section we demonstrate the use of RCVB to perform posterior inference for our proposed model.

4 Posterior inference in variational discriminant analysis

The endpoint of variational discriminant analysis with variable selection (VaDA) is to identify the discriminative variables and the true group label of new observations. The calculations provided in the rest of this section will pertain only to \( \text{VLDA} \). Posterior inference calculations for \( \text{VQDA} \) are provided in the appendix. Given \( m \) new observations \( \{(X_{n+1}, Y_{n+1})\}_{i=1}^m \) where \( \{Y_{n+1}\}_{i=1}^m \) are latent, we may regard \( \gamma \) and \( \{Y_{n+1}\}_{i=1}^m \) as the parameters of interest. Let \( \theta_1 = (\mu_1, \mu_0, \sigma_1^2, \sigma^2, \rho_y, \rho) \) be the parameters not required for inference, which will be collapsed over. We provide details about the inference for \( m = 1 \) and will describe the generalisation to any \( m > 1 \) towards the end. Let \( X = [X_1, \ldots, X_n]^T = [\tilde{X}_1, \ldots, \tilde{X}_p] \in \mathbb{R}^{n \times p} \), where \( \tilde{X}_j \) is the column vector of variable \( j \) of the data and \( y = (y_1, \ldots, y_n)^T \) is the observed column vector of binary responses. Let \( D = (X, y) \) denote the training data. The posterior distribution of \( \gamma_j \) given \( D \) and \( X_{n+1} \) may be expressed as

\[
p(\gamma_j \mid D, X_{n+1}) = \frac{p(\gamma_j, D, X_{n+1})}{p(D, X_{n+1})}.
\]

Following arguments for a diffused prior in Ormerod et al. (2017), we let \( h \to \infty \). The marginal likelihood of the observed data \( (D \text{ and } X_{n+1}) \) in the denominator can be written as

\[
p(D, X_{n+1}) = \sum_{y_{n+1} \in \{0, 1\}} \sum_{\gamma \in \{0, 1\}^p} \int p(D, X_{n+1}, y_{n+1}, \gamma, \theta_1) \, d\theta_1
\]

\[
= \left[ \prod_{j=1}^p p(\tilde{X}_j, x_{n+1}, j \mid \gamma_j = 0) \right] \times \sum_{y_{n+1} \in \{0, 1\}} \left[ B(a_y + n_1 + y_{n+1}, b_y + n_0 + 1 - y_{n+1}) \times \sum_{\gamma \in \{0, 1\}^p} \exp \left\{ \log B(a_\gamma + 1^T \gamma, b_\gamma + p - 1^T \gamma) \right. 
\right.
\]

\[
+ \frac{1}{2} \gamma^T \lambda_{\text{Bayes}}(\tilde{X}_j, x_{n+1}, y, y_{n+1}) \right],
\]

where \( B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b) \) is the beta function, \( n_1 = 1^T y \), \( n_0 = n - n_1 \) and \( \lambda_{\text{Bayes}} \) is a column vector of size \( p \) which is defined as follows. The likelihood ratio statistic corresponding to the test \( \{H_{0j} : \gamma_j = 0\} \) using
model (2) against \{H_{1j} : \gamma_j = 1\} which uses (1) for variable \(j\) is

\[
\lambda_{\text{LRT}}(\tilde{x}_j, x_{n+1,j}, y, y_{n+1}) = (n + 1) \log \left( \frac{\sigma^2_{j1}}{\sigma^2_{j}} \right),
\]

where the maximum likelihood estimates (MLEs) are

\[
\hat{\sigma}^2_{j1} = \frac{1}{n + 1} \left[ ||y^T(\tilde{x}_j - \tilde{\mu}_{j1})||^2 + ||(1 - y^T)(\tilde{x}_j - \tilde{\mu}_{j0})||^2 \right] + y_{n+1}(x_{n+1,j} - \tilde{\mu}_{j1})^2 + (1 - y_{n+1})(x_{n+1,j} - \tilde{\mu}_{j0})^2,
\]

\[
\hat{\sigma}^2_j = \frac{1}{n+T} \left\{ ||\tilde{x}_j - \tilde{\mu}_{j0}||^2 + (x_{n+1,j} - \tilde{\mu}_{j})^2 \right\},
\]

\[
\tilde{\mu}_{j} = \frac{1}{n+1+y_{n+1}} \left\{ y^T\tilde{x}_j + y_{n+1}x_{n+1,j} \right\}, \quad \tilde{\mu}_{j0} = \frac{1}{n+1-y_{n+1}} \left\{ (1 - y)^T\tilde{x}_j + (1 - y_{n+1})x_{n+1,j} \right\},
\]

and \(\lambda_{\text{Bayes}}\) of \(\gamma_j\) is (as \(h \to \infty\)) satisfies

\[
\lambda_{\text{Bayes}}(\tilde{x}_j, x_{n+1,j}, y, y_{n+1}) \to \lambda_{\text{LRT}}(\tilde{x}_j, x_{n+1,j}, y, y_{n+1}) - \log(n + 1).
\]

The marginal likelihood in equation (9) involves a combinatorial sum over \(2^{p+1}\) binary combinations. Hence, exact Bayesian inference is computationally infeasible for large \(p\) and approximation is required. We will use RCVB to approximate the posterior \(p(\gamma, y_{n+1}|D, x_{n+1})\) using the partition

\[
q^{\text{RCVB}}(y_{n+1}, \gamma) = q^{\text{RCVB}}(y_{n+1}) \prod_{j=1}^{p} q^{\text{RCVB}}_j(\gamma_j).
\]

We shall henceforth drop the superscript RCVB. Note that since \(\gamma_j \in \{0, 1\}\) and \(y_{n+1} \in \{0, 1\}\) we have

\[
q_j(\gamma_j) = w_j^{\gamma_j}(1 - w_j)^{1-\gamma_j},
\]

\[
q(y_{n+1}) = \tilde{\gamma}^{y_{n+1}}(1 - \tilde{\gamma})^{1-y_{n+1}},
\]

where \(w\) and \(\tilde{\gamma}\) are variational parameters to be optimized over, i.e., \(q_j(\gamma_j)\) and \(q(y_{n+1})\) are densities corresponding to a Bernoulli(\(w_j\)) and Bernoulli(\(\tilde{\gamma}\)) distribution. The variational parameter \(w_j\) may be interpreted as the approximate posterior probability that the hypothesis \(H_{1j}\) is true when tested against \(H_{0j}\). The interpretation of \(\tilde{\gamma}\) is analogous.

### 4.1 Variable selection

With reference to the steps in (8), we will apply variational Bayes approximation over \(\gamma_{-j}\) and \(y_{n+1}\), and integrate analytically over \(\theta_1\), i.e., for \(1 \leq j \leq p\) we have

\[
q_j(\gamma_j) \propto \int \exp \left[ \mathbb{E}_{-q_j} \{ \log p(D, x_{n+1}, y_{n+1}, \gamma, \theta_1) \} \right] d\theta_1,
\]

\[
\propto \exp \left[ \mathbb{E}_{-q_j} \left\{ \log B(\alpha_{\gamma} + 1^T \gamma, b_\gamma + p - 1^T \gamma) \right\} \right.
\]

\[
+ \frac{\gamma}{2} \mathbb{E}_{-q_j} \left\{ \lambda_{\text{Bayes}}(\tilde{x}_j, x_{n+1,j}, y, y_{n+1}) \right\}.
\]
For a sufficiently large $n$, we can avoid the need to evaluate the expectation $E_{-q_j} \left\{ \lambda_{\text{Bayes}}(\tilde{x}_j, x_{n+1, j}, y, y_{n+1}) \right\}$ by applying Taylor’s expansion to approximate the MLEs with

$$
E_{-q_j} \log(a_\gamma + 1^T \gamma_{-j}) \approx \log(a_\gamma + 1^T w_{-j}),
$$
$$
E_{-q_j} \log(b_\gamma + p - 1^T \gamma_{-j} - 1)
\approx \log(b_\gamma + p - 1^T w_{-j} - 1),
$$
$$
\hat{\sigma}_j^2 \approx \frac{1}{n} \left( ||y^T (\tilde{x}_j - \hat{\mu}_j 1)||^2 + ||(1 - y^T) (\tilde{x}_j - \hat{\mu}_j 0)||^2 \right),
$$
$$
\hat{\sigma}_j^2 \approx \frac{1}{n} ||\tilde{x}_j - \hat{\mu}_j 1||^2,
$$

and hence $\lambda_{\text{Bayes}}$ does not depend on the new observation $(x_{n+1}, y_{n+1})$. By using the approximation in (10), we have

$$
w_j = \frac{q_j(\gamma_j = 1)}{q_j(\gamma_j = 1) + q_j(\gamma_j = 0)},
$$
$$
\approx \expit \left[ \log(a_\gamma + 1^T w_{-j}) - \log(b_\gamma + p - 1^T w_{-j} - 1) + \frac{1}{2} \lambda_{\text{Bayes}}(\tilde{x}_j, y) \right],
$$
$$
= \expit \left[ \log(a_\gamma + 1^T w_{-j}) - \log(b_\gamma + p - 1^T w_{-j} - 1) - \frac{1}{2} \log(n + 1) + \frac{1}{2} \lambda_{LRT}(\tilde{x}_j, y) \right],
$$
$$
= \expit \left[ \text{penalty}_j + \frac{1}{2} \lambda_{LRT}(\tilde{x}_j, y) \right],
$$

Each $w_j$ may be viewed, from a frequentist’s perspective, as a “test statistic” for $H_{1j}$ against $H_{0j}$. Hence, a natural decision rule is to identify variable $j$ as discriminative if $w_j > c_w$ for some constant $c_w \in (0, 1)$.

The penalty term in the expression for $w_j$ can be interpreted as a data dependent penalty term which trades off type I errors against power. This presents the task of identifying discriminative variables in our model as a multiple hypothesis testing problem using penalized likelihood ratio statistics. The constant $b_\gamma$ is particularly important. While we have chosen $a_\gamma = 1$ and $b_\gamma$ given by (3) for some $r < 1$ and $\kappa > 0$, one may specify other values of $b_\gamma$. If $b_\gamma$ is too small, then false positives will occur when $p$ is allowed to diverge with $n$. If $b_\gamma$ is too large, then there will be potentially too many false negatives. Ideally we want $b_\gamma$ to be as small as possible whilst having asymptotically zero false positives. Since the updates of each $w_j$ no longer depends on $\tilde{y}$ after applying the Taylor’s expansion results, each RCVB cycle only involves the update of $w_j$ over $1 \leq j \leq p$ until convergence.
4.2 Classification

We will apply variational Bayes approximation over \( \gamma \) and integrate analytically over \( \mathbf{\theta}_1 \) to obtain the approximate density for \( y_{n+1} \), i.e.,

\[
q(y_{n+1}) \propto \int \exp \left[ \mathbb{E}_{-y} \{ \log p(D, x_{n+1}, y_{n+1}, \gamma, \mathbf{\theta}_1) \} \right] d\mathbf{\theta}_1,
\]

\[
\propto \exp \left[ \log B(a_y + n_1 + y_{n+1}, b_y + n_0 + 1 - y_{n+1})
- \frac{1}{2} (n + 1) \mathbf{w}^T \log \{(n + 1) \hat{\mathbf{\sigma}}_1^2 \} \right],
\]

where \( \log \{(n + 1) \hat{\mathbf{\sigma}}_1^2 \} \) is the element-wise log of the column vector \( (n + 1) \hat{\mathbf{\sigma}}_1^2 \). Note that each element in \( \hat{\mathbf{\sigma}}_1^2 \) is function of \( y_{n+1} \).

Thus, the corresponding variational parameter \( \tilde{y} \) is

\[
\tilde{y} = \frac{q(y_{n+1} = 1)}{q(y_{n+1} = 1) + q(y_{n+1} = 0)},
\]

\[
\approx \expit \left[ \log \left( \frac{n_1 + a_y}{n_0 + b_y} \right) + \left( 1 + \frac{1}{n} \right) \text{LDA}(x_{n+1}) \right],
\]

where

\[
\text{LDA}(x_{n+1}) = \hat{\mathbf{\mu}}^T \mathbf{W}^{-1} \{ \hat{\mathbf{x}}_{n+1} - \frac{1}{2} (\hat{\mathbf{\mu}}_0 + \hat{\mathbf{\mu}}_1) \}
\]

is the naïve Bayes LDA classification rule assuming a balanced training dataset that is downweighted by \( \mathbf{W} = \text{diag}(w_1, \ldots, w_p) \), and \( \mathbf{\Sigma} \) is a diagonal matrix with entries \( \{\hat{\mathbf{\sigma}}_{1j}^2\}_{j=1}^p \). Note that when \( \mathbf{W} = \mathbf{1}_p \) and \( a_y = b_y = 0 \) the classification of \( \tilde{y} \) is the same as for a DLDA classifier.

In the general case whereby there are \( m \) new observations to be classified, the variational parameter for \( y_{n+i} \) is

\[
\tilde{y}_i = \expit \left[ \log \left( \frac{n_1 + \mathbb{E}_{-q_i}(y_{n+i+m}) + a_y}{n_0 + m - \mathbb{E}_{-q_i}(y_{n+i+m}) + b_y} \right) + \left( 1 + \frac{1}{n} \right) \text{LDA}(x_{n+i}) \right],
\]

where \( \mathbb{E}_{-q_i} \) denotes the expectation with respect to approximate densities of \( y_{(n+i)} \) and \( \gamma \). For computational efficiency, we may replace the moment estimates in the expression for \( \text{LDA}(x_{n+1}) \) with their respective Taylor’s expansion approximation in (10) and omit the term \( \mathbb{E}_{-q}(y_{n+i+m}) \). Hence, for \( 1 \leq i \leq m \), the updates in (11) may be approximated as

\[
\tilde{y}_i \approx \expit \left[ \log \left( \frac{n_1 + a_y}{n_0 + b_y} \right) + \left( 1 + \frac{1}{n} \right) \text{LDA}(x_{n+i}) \right].
\]

This approximation allows us to update \( w_j \) in each RCVB cycle and use only their converged values to update \( \tilde{y}_i \) for \( 1 \leq i \leq m \). In addition, the updates of \( \tilde{y}_i \) do not depend on each other. Hence, the length of each update cycle is kept at \( p \). We may use \( \tilde{y}_i \) to construct a classification rule by classifying observation \( n+i \) to group 1 if \( \tilde{y}_{n+i} > c_y \) for some \( c_y \in (0, 1) \).

The RCVB algorithm may be found Table[1]. Notice that at iteration \( t \), we use values of \( \mathbf{w} \) at iteration \( t-1 \) instead of the values at the current time step. This is in line with the batch update strategy described in Zhang and Zhou (2017).

Before ending this section, it is worth noting that the appearance of the terms \( \lambda_{\text{LRT}}(\hat{\mathbf{x}}_j) \) and \( \text{LDA}(x_{n+i}) \) in the updates for \( w_j \) and \( \tilde{y}_i \) provides reassurance that there is some concordance between the frequentist methods.
Table 1  Iterative scheme for obtaining the parameters in the optimal densities $q(\gamma; y_{n+1}, \ldots, y_{n+m})$ in VLDA

| Require: For each $j$, initialise $w_j^{(0)}$ with a number in $[0, 1]$. |
|--------------------------------------------------------------------------|
| while $||w^{(t)} - w^{(t-1)}||^2 > \epsilon$ do |
| At iteration $t$, |
| 1: $\eta_j^{(t)} \leftarrow \log \left\{ a_\gamma + 1^T w_j^{(t-1)} \right\} - \log \left\{ b_\gamma (r, \kappa) + p - 1^T w_j^{(t-1)} - 1 \right\} - \frac{1}{2} \log(n + 1) + \frac{1}{2}(n + 1) \log(\hat{\sigma}_j^2)$ |
| 2: $w_j^{(t)} \leftarrow \expit(\eta_j^{(t)})$ |
| Upon convergence of $w$, compute for $i = 1, \ldots, m$ |
| 3: $\hat{y}_i \leftarrow \expit \left[ \log \left( \frac{n_1 + a}{n_0 + b} \right) + (1 + \frac{1}{n})(\hat{\mu}_0 - \hat{\mu}_1)^T W \Sigma^{-1} \left\{ x_{n+i} - \frac{1}{2}(\hat{\mu}_0 + \hat{\mu}_1) \right\} \right]$ |

and our proposed variable selection and classification rules, and hence fulfills the consistency criterion which we considered in Section 2.2 when choosing our priors.

5 Variable selection asymptotics

Here, we establish a desirable asymptotic property, known as Chernoff-consistency, of the variable selection rule that identifies variable $j$ as discriminative if $w_j > c_w$ for any $c_w \in (0, 1)$. Although Wang and Blei (2018) has demonstrated the consistency of variational Bayes estimates in their recent work, their result cannot be directly applied to our high-dimensional setting as they have assumed a fixed model dimension.

The variable selection rule is Chernoff-consistent if the sum of Type I and II errors, $E^{(t)}$, given by

$$E^{(t)} = \sum_{j=1}^{p} |w_j^{(t)} - \gamma_j^*|$$

where $\gamma_j^* \in \{0, 1\}$ is the true value of $\gamma$, converges in probability to 0 as $n$ diverges. This sum can be broken down into components as

$$E^{(t)} = e_0^{(t)} + e_1^{(t)},$$

where the sum of Type I errors (non-discriminative variables identified as discriminative) after cycle $t$ is

$$e_0^{(t)} = \sum_{\ell \in J_0} w_\ell^{(t)},$$

the sum of Type II errors (discriminative variables identified as non-discriminative) after cycle $t$ is

$$e_1^{(t)} = \sum_{\ell \in J_1} \{1 - w_\ell^{(t)}\},$$

and $J_k = \{1 \leq j \leq p_n : \gamma_j^* = k\}$. Note that we have allowed the number of variables $p_n$ to increase with $n$ such that $|J_0| \uparrow \infty$ and $|J_1| \uparrow C$ as $n$ diverges for some constant $C$, but we have left this detail from the notations for
brevity. We have also used notations superscripted with * to denote the true value of the Gaussian parameters in the frequentist sense. For example, $\mu^*_{jk}$ is a fixed constant denoting the actual population mean of variable $j$ for observations from group $k$. Since $E^{(t)}$ and $w^{(t)}$ are iteratively related, we need only to prove that under the proposed initialisation $w_j^{(0)} \in [0,1]$ for each $j$, the total error after one cycle of the VB iterations $E^{(1)}$ converges in probability to 0.

We shall now state Theorem 1 in the main paper. Its proof has been provided in Appendix B.

**Theorem 1.** Consider $p_n$ sets of iid random variables $\tilde{X}_j = (X_{1j}, \ldots, X_{nj})^T$ that are independently drawn from multivariate normal populations, where $(X_{ij}, X_{ik})$ are independent for $j \neq k$, and $p_n$ corresponding pairs of distributional hypotheses

$$H_{0j}: X_{ij} \sim N(\mu_j,\sigma^2_{j}),$$

versus

$$H_{1j}: X_{ij} \sim \begin{cases} N(\mu_{j1},\sigma^2_{j1}), & \text{if } y_i = 1; \\
N(\mu_{j0},\sigma^2_{j1}), & \text{if } y_i = 0,
\end{cases}$$

such that exactly one hypothesis in each pair is true.
Let $\gamma^* \in \{0,1\}^{pn}$ be a binary indicator vector of length $p_n$ whose $j$th value is

$$\gamma^*_j = \begin{cases} 1 & \text{if } H_{1j} \text{ is true;} \\
0 & \text{if } H_{0j} \text{ is true,}
\end{cases}$$

Under the initialization $w_j^{(0)} \in [0,1]$ for each $j$ and the following set of conditions:

1. the number of truly discriminative variables $p_{1n} = |J_1|$ is an increasing sequence that is bounded by a constant,

2. the true parameters of the normal populations $\mu^*_{j1}, \mu^*_{j0}$ and $\sigma^2_{j1}$ are finite for all $j \geq 1$ and,

3. the total number of variables is bounded as $p_n \leq \tilde{p}_n$, where $\tilde{p}_n = \exp\{(n+1)/\log(n+1)\},$

the variable selection rule that identifies variable $j$ as discriminative if $w_j > c_w$ for any $c_w \in (0,1)$ is Chernoff-consistent, i.e.,

$$E^{(1)} = o_p(1).$$

The above theorem demonstrates the ability of the variable selection rule to avoid Type I error inflation due to the increase of $p_n$, and a tendency to select all truly discriminative variables. However, these properties are only guaranteed when we choose the Cake priors for the Gaussian parameters and the prior for $\rho_1$ in accordance with Section 2.2. We also make a disclaimer that our algorithm may not be optimal with respect to Type I and II errors in any sense.

Implicitly speaking, our resultant variable selection rule is justified by the asymptotic error rates that they induce. This differs from the existing variable selection rules such as higher criticism [Donoho and Jin 2008] and false-discovery rate controlling procedures [Benjamini and Hochberg 1995], whereby they are justified on the basis of a minimised Type II error rate for a user-specified family-wise Type I error rate and finite sample size $n$. 

12
6 Numerical results

In this section, we assess the performance of VaDA classifiers with eight simulation settings and two publicly available gene expression datasets. Both VLDA and VQDA variants will be compared with other discriminant analysis models such as two from the R package HiDimDA [Pedro, 2015]: the diagonal linear discriminant analysis with expanded higher criticism (EHC) - Dlda and the factor model linear discriminant analysis with EHC - RFlda. We also compare their performance with the factor-adjusted shrinkage discriminant analysis from the package FADA [Perthame et al., 2018]. Lastly, we included three penalised discriminant analysis models in our comparison: the nearest shrunken centroid (NSC) classifier from pamr [Hastie et al., 2014], lasso (penLDA-L1) and fused-lasso (penLDA-FL) penalised Fisher’s linear discriminant analysis from the R package penalizedLDA [Witten, 2015].

The codes for both variants of VaDA have been made publicly available at:

http://www.maths.usyd.edu.au/u/jormerod/

We have also found that a setting of $r = 0.98$ and $\kappa = 10^{-3}$ works well in both the simulated and gene expression datasets. The variable selection and classification thresholds are $c_w = 0.5$ and $c_y = 0.5$.

6.1 Competing classifiers

Dlda, RFlda and FADA are two-stage classification algorithm that perform variable selection and classification in un-integrated stages. In the first stage of FADA, the matrix of variables $X$ is de-correlated with a method described in Friguet et al. (2009), and is followed by an application of the higher criticism threshold (see Donoho and Jin, 2008) to select the discriminative variables. A modified version of the higher criticism, adapted for large or moderate signals, is used in Dlda and RFlda (see Pedro, 2011).

In the second stage, Dlda fits the naïve Bayes LDA model with the subset of selected variables whereas RFlda fits a factor-based linear discriminant analysis with user-specified $q$ as the number of factors. In our analysis, we have chosen the value of $q$ from $\{1, 2, 3\}$ that minimises CV error. In FADA, the selected variables from stage one are used to fit a shrinkage discriminant analysis model (Ahdesmäki and Strimmer, 2010).

NSC classifies observations according to the nearest (group) shrunken centroid. A tuning parameter $\Delta$ controls the amount of shrinkage, and consequently the sparsity of the estimated difference between the two group centroids. This filters individually weak but collectively strong false signals. In our analysis, the adaptive choice of $\Delta$ is adopted following procedures in Tibshirani et al. (2003). NSC is widely used in biomarker discovery (see Liu et al., 2005; Craig-Shapiro et al., 2011) and is therefore a good competing classifier.

Lastly, there has been a huge amount of attention given to sparse discriminant analysis models in the literature (see Clemmensen et al., 2011). Hence, we have also chosen to compare our classifiers with the two versions of penLDA that extended the solution to Fisher’s discriminant problem (Fisher, 1936) to high dimension settings. In the two-groups version of Fisher’s discriminant problem, one obtains the estimated discriminant vector $\hat{\beta}$ such that

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} (\beta^T \hat{\Sigma}_b \beta),$$

subject to $\beta^T \hat{\Sigma}_w \beta = 1$, where we have assumed that the estimated within-class covariance matrix $\hat{\Sigma}_w$ is full rank and $\hat{\Sigma}_b$ is the estimated between-class covariance matrix. A new observation, $x_{n+1}$, is mapped under

\[13\]
the transformation $x_{n+1}^T \hat{\beta}$ and classified in accordance to its nearest transformed centroid. In the penLDA extension, $\hat{\Sigma}_w$ is replaced with the diagonal estimate. Sparsity is induced on $\hat{\beta}$ by imposing either the lasso penalty (penLDA-L1) or the fused-lasso penalty (penLDA-FL) on the objective function in \[13\]. The fused lasso penalty assumes linear ordering in variable indices and is therefore not suitable for the publicly available datsets where such ordering cannot be ascertained. Tuning parameters are selected from $10^{-4}$, $10^{-3}$, $10^{-2}$, 0.1, 1 or 10 to minimise CV error.

6.2 Simulation setting

We assess the performance of both VLDA and VQDA with simulated data. The objective of this simulation study is to identify situations in which VaDA’s performance differ from the competing classifier. Performance of all classifiers will be assessed with eight simulation settings. The first four settings are motivated from those examined in Witten and Tibshirani (2011). The response values are generated as follow. We set $y_i$, for each $i$, to take values 1 or 0 with probability 0.5. This simulates a balanced study design and also gives adequate chance of having at least two observations per group in each simulation repetition. The variables are generated as follows.

$y_i \sim \begin{cases} \mathcal{N}(\mu^*_1, \sigma^*_2), & \text{if } y_i = 1; \\ \mathcal{N}(\mu^*_0, \sigma^*_0^2), & \text{if } y_i = 0. \end{cases}$

We use the term signal strength, denoted by $d$, to describe the absolute standardised difference in the group-conditional means, i.e., $d_j = 2|\mu^*_j - \mu^*_0|/(\sigma^*_j^2 + \sigma^*_0^2)$. Each simulation yields a dataset of $p = 500$ variables.

Details of their settings are as follow.

Simulation 1:

Here, we simulate the situation when there is a moderate proportion (10%) with non-zero true signals and that the signal strengths ($d_j = 0.7$) are moderately strong. We set $\mu^*_j = 0.7$ if $k = 1$ and $1 \leq j \leq 50$ and $\mu^*_j = 0$ otherwise. There is also some linear ordering in this setting. This is a homogenous variance setting such that $\sigma^*_j^2 = \sigma^*_0^2 = 1$ for all $1 \leq j \leq p$.

Simulation 2:

This setting assesses the impact on performance of both VLDA and VQDA when the independence assumption is violated and signal strength ($d_j = 0.6$) is moderate. We set $\mu^*_j = 0.6$ for $1 \leq j \leq 200$ and $\mu^*_j = 0$ otherwise. We grouped the variables into five networks - \{X$_{1+t}$, $X_{10+t}$, $X_{20+t}$, $X_{30+t}$, $X_{40+t}$\} for $t = 0, 100, 200, 300, 400$. This setting is representative of pathway models in genomic studies. An AR(1) covariance matrix is assigned as the dependency structure within each network, where the covariance between any two adjacent variables in the same network is $\rho = 0.6$ in either groups. Variables from different networks are uncorrelated. Here, there is some linear ordering in this simulation setting. This is a homogenous variance setting such that $\sigma^*_j^2 = \sigma^*_0^2 = 1$ for all $1 \leq j \leq p$.

Simulation 3:

We test the proposed classifiers in the situation when there is a sizeable proportion (20%) of variables with non-zero true signals and that the signal strengths ($d_j = 0.3$) are weak. We set $\mu^*_j = 0.3$ if $k = 1$ and $1 \leq j \leq 100$. There is also some linear ordering in this setting. This is a homogenous variance setting such that $\sigma^*_j^2 = \sigma^*_0^2 = 1$ for all $1 \leq j \leq p$. 

\[14\]
Simulation 4:
Here, we draw the group means of truly discriminative variables from a normal distribution. This setting examines performances of the classifiers when there is varying signal strengths in the data, while ensuring a large chance (≈90.2%) that there is at least one moderately strong ($|d_j| \approx 0.6$) signal. This time, we set $\mu_{jk}^* \sim \mathcal{N}(0, 0.3^2)$ if $k = 1$ and $1 \leq j \leq 50$, and $\mu_{jk}^* = 0$ otherwise. Linear ordering can no longer be guaranteed. This is a homogeneous variance setting such that $\sigma_{j1}^2 = \sigma_{j0}^2 = 1$ for all $1 \leq j \leq p$.

Simulation 5:
We simulate the situation when we have a very small proportion (1%) of truly discriminative variables and that the non-zero signal strengths each have high chance (≈78.9%) of being moderately strong ($|d_j| \approx 0.6$). We let $\mu_{j1}^* \sim \mathcal{N}(0, 0.3^2)$ and $\mu_{j0}^* = 0$ for $1 \leq j \leq 5$. Otherwise, $\mu_{jk}^* = 0$. This is a homogeneous variance setting such that $\sigma_{j1}^2 = \sigma_{j0}^2 = 0.6$ for all $1 \leq j \leq p$.

Simulation 6:
A heterogeneous version of Simulation 1. We set $\sigma_{j1}^2 = 0.7$ and $\sigma_{j0}^2 = 1.3$ for $1 \leq j \leq 25$, $\sigma_{j1}^2 = 1.3$ and $\sigma_{j0}^2 = 0.7$ for $26 \leq j \leq 50$ and $\sigma_{jk}^2 = 1$ otherwise.

Simulation 7:
A heterogeneous version of Simulation 3. We set $\sigma_{j1}^2 = 0.7$ and $\sigma_{j0}^2 = 1.3$ for $1 \leq j \leq 100$ and $\sigma_{jk}^2 = 1$ otherwise.

Simulation 8:
A heterogeneous version of Simulation 5. We set $\sigma_{j1}^2 = 0.4$ and $\sigma_{j0}^2 = 0.8$ for $1 \leq j \leq 5$ and $\sigma_{jk}^2 = 0.6$ otherwise.

6.3 Performance metrics for simulated datasets

The distribution of classification errors for each simulation setting is summarised over 25 simulation repetitions with $n = 100$. In each repetition, we generated the 1200 observations. The first 100 observations $\{(x_i, y_i)\}_{i=1}^{100}$ are assigned as the training set. The next 100 observations $\{(x_i, y_i)\}_{i=101}^{200}$ are designated as the validation set for choosing optimal tuning parameters in \texttt{penLDA-L1} and \texttt{penLDA-FL}. The remaining 1000 observations $\{(x_i, y_i)\}_{i=201}^{1200}$ will make up the testing dataset. At each iteration, the classification error is computed as

$$\text{Classification Error} = \frac{\sum_{i=201}^{1200} I\{y_i \neq \hat{y}_i\}}{m},$$

where $\hat{y}_i$ is the predicted classification of testing data $i$ and $m = 1000$.

Variable selection performance are compared using Matthew’s correlation coefficient (Matthews, 1975), computed as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where TP, TN, FP and FN is the number of true positives, true negatives, false positives and false negatives respectively in a particular repetition. A higher value of MCC indicates better variable selection performance. The MCC presents as a suitable variable selection metric for our simulation settings as it accounts for the imbalance in the total number of truly discriminative (TP + FP) and non-discriminative (TN + FN) variables (Chicco, 2017). Since computational cost has also been raised in Section II as a problem with existing models, we shall compare the computation time required by each classifier.
Figure 1: Classification errors for simulated datasets ($n = 100$)

Figure 2: Matthews correlation for simulated datasets ($n = 100$)
6.4 Simulation results

Boxplots of classification errors and MCCs may be found in Figures ?? and ??.

In simulation settings 1 to 5, all classifiers except penLDA-FL and VQDA performed well in comparison to one another. The superior performance by penLDA-FL may be attributed to its use of information in linear ordering, whereas the poorer performance by VQDA is a result of a loss of statistical power as expense for estimating a separate set of group-conditional variances.

First, we do a detailed comparison of the classification errors. Evidently, VLDA exhibits competitive or superior classification errors for settings 1, 4 and 5. Settings 4 and 5 are similar in that the true signal strength is drawn randomly from a normal distribution, thus there is high chance that there is at least one strong signal. The presence of at least one strong signal may explain the edge exhibited by VLDA. It is surprising that even when the naïve Bayes assumption is violated in underlying data (see simulation 2 results), the classification errors of the naïve Bayes classifiers (except VQDA) remain within an acceptable range. In fact, they are lower relative to RFlda(EHC) and FADA that allow for correlated variables. We expected the two penLDA to outperform other classifiers in Simulation 3 as regularisation has been shown to be effective in picking up weak signals while filtering most of the noise.

By comparing classification errors in settings 1, 3 and 5 with their heterogenous versions in settings 6 to 8, we found that all classifiers that assumed homogeneous variance exhibited a significant increase in classification errors. This may be contrasted with VQDA’s significant decrease in classification errors and is convincing evidence that one should test the homogeneity of variance assumption before choosing an appropriate classifier.

Next, we compare the variable selection performance. The fused-lasso penalty appears to do extremely well in all applicable circumstances. VLDA exhibited comparable performance with the L1-penalty in settings 1, 4 and 5. As expected, VQDA performed much better in the heterogeneous variance settings. Among the methods that assume homogeneous variance, the magnitude of change when comparing simulation 1, 3 and 5 against their heterogeneous versions is noticeable but not large. This provides some evidence of their robustness when the homogeneous variance assumption is violated.

Since the average computational time required for a single repetition remain below 5s for each classifier, they are considered acceptable and will therefore not be elaborated. The maximum computational time required for each of the competing classifier is approximately 2 to 400 times that required for VLDA and VQDA.

6.5 Gene expression datasets

The classifiers are also compared using two gene expression datasets. The description of the datasets may be found in the next two paragraphs. Since the subset of truly discriminative variables are unknown for each dataset, we shall omit the comparison of the variable selection performance. A filtering step is applied to each dataset to remove genes with mostly 0 readings. We then standardised each dataset to obtain $z_{ij} = (x_{ij} - \bar{\mu}_j)/s_j$ where $x_{ij}$ is the gene $j$ reading for observation $i$, $\bar{\mu}_j$ is the sample mean of gene $j$ and $s_j$ is the sample standard deviation. This standardisation procedure is similar to the one in Dudoit et al. (2002). A 5-fold cross validation over 50 repetitions is performed.

The total number of misclassifications at each iteration is summed across the 5 CV sub-iterations to compare performance between the classifiers. The classification errors and computational time are presented in Figure 3 and Table 2 respectively.
**Colorectal cancer dataset:** The colon cancer dataset is sourced from [Bioconductor](https://bioconductor.org) and has been analysed by [Jorissen et al.](https://www.ncbi.nlm.nih.gov/pubmed/18486855) (2008). The dataset consist of \( n = 155 \) observations of Affymetrix oligonucleotide arrays. The response variable is whether the tumour exhibited microsatellite instability, among which we have 78 microsatellite instable (MSI) tumours and 77 microsatellite stable (MSS) tumours. We implemented a filtering step that excludes genes with within-class outliers that are either 3 IQR above or below the median. This leaves us with \( p = 8212 \) genes.

**TCGA-LIHC dataset:** The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) dataset is a collection of clinical, genetic and pathological data residing in the Genomic Data Commons (GDC) Data Portal and is made publicly available ([Erickson et al.](https://www.ncbi.nlm.nih.gov/pubmed/27205649) 2016). The data underwent pre-processing to remove genes with IQR \( \leq 0.3 \) and the \( \log_2(1 + \text{FPKM}) \) transformation is taken. Patients whose survival time are lesser the the 20th percentile are considered as poor prognosis \( (n_{\text{poor}} = 47) \) while patients with survival time greater than 80th percentile are labelled as good prognosis \( (n_{\text{good}} = 62) \). Patients who belong to neither survival category were removed from the analysis. This leaves us with \( n = 109 \) observations of 15681 RNA-seq readings.

### 6.6 Gene expression dataset results

Classification errors for the colorectal cancer dataset are comparable with one another and are within the expected range when benchmarked with those presented in [Jorissen et al.](https://www.ncbi.nlm.nih.gov/pubmed/18486855) (2008). Performance is evidently more differentiated when the classifiers are trained with the TCGA-LIHC dataset, with VLDA yielding the second lowest median classification error. The non-naïve Bayes classifiers did not exhibit superior classification error rates than their naïve Bayes counterparts. This may imply that the true between-variable covariances in both datasets are weak.

The total computational time required for a 5-fold CV repetition is presented in Table 2. The two VaDA variants compute at least 200 times faster than the other classifiers. Similar to the simulation study results, we found non-naïve Bayes classifiers (Rflda(EHC) and FADA) to be much slower when compared to naïve Bayes classifiers.

| Datasets | Summary stats. | VLDA | VQDA | Dlda (EHC) | RFlda (EHC) | FADA | NSC | penLDA-L1 |
|----------|----------------|------|------|------------|-------------|------|-----|----------|
|          | Mean           | 7.19 | 8.32 | 12067.46   | 15826.15    | 30478.45 | 1998.25 | 2450.89 |
|          | SD             | (0.27) | (0.40) | (698.38)   | (788.26)    | (2447.08) | (60.44) | (96.70) |
| LIHC     | Mean           | 11.45 | 13.36 | 3106.23    | 3131.93     | 131630.45 | 2783.55 | 1410.32 |
|          | SD             | (0.82) | (0.75) | (353.85)   | (467.73)    | (14329.72) | (120.61) | (40.49) |

Table 2: Total computation time (ms) required for 5-fold cross validation in genomics datasets.

### 7 Conclusion

We have proposed a classifier that integrates two common objectives in high dimensional data analysis variable selection and classification. The Bayesian framework of the classifier lends a two fold-advantage to the classifier, provided priors are chosen according to the recommended settings in this paper. Firstly, it leads us to a variable
Figure 3: 5-fold CV classification errors for genomics datasets (50 reps)

A VQDA derivations

In the VQDA setting ($\sigma^2_{j1} \neq \sigma^2_{j0}$) the posterior distribution of $\gamma_j$ given $D$ and $x_{n+1}$ may be expressed as

$$p(\gamma_j \mid D, x_{n+1}) = \frac{p(\gamma_j, D, x_{n+1})}{p(D, x_{n+1})}.$$ 

By letting $h \to \infty$, the marginal likelihood of the data in the denominator is of the same form as equation (9) with the exception that

$$\theta_1 = (\mu_1, \mu_0, \mu, \sigma^2_1, \sigma^2_0, \sigma^2, \rho_y, \rho_\gamma),$$

and

$$\lambda_{LRT}(\bar{x}_j, x_{n+1}, y, y_{n+1}) = (n_1 + 1) \log(\hat{\sigma}^2_{j1}) - (n_1 + y_{n+1}) \log(\hat{\sigma}^2_{j1}) - (n_0 + 1 - y_{n+1}) \log(\hat{\sigma}^2_{j0}),$$

where

$$\hat{\sigma}^2_{j1} = \frac{1}{n_1 + y_{n+1}} \left[ \|y^T (\bar{x}_j - \hat{\mu}_{j1}1)\|^2 + y_{n+1}(x_{n+1,j} - \hat{\mu}_{j1})^2 \right],$$

$$\hat{\sigma}^2_{j0} = \frac{1}{n_0 + 1 - y_{n+1}} \left[ \|(1 - y)^T (\bar{x}_j - \hat{\mu}_{j0}1)\|^2 + (1 - y_{n+1})(x_{n+1,j} - \hat{\mu}_{j0})^2 \right].$$
Table 3 Iterative scheme for obtaining the parameters in the optimal densities $q(\gamma, y_{n+1}, \ldots, y_{n+m})$ in VQDA

Require: For each $j$, initialise $w_j^{(0)}$ with a number in $[0, 1]$.

while $||w^{(t)} - w^{(t-1)}||^2 > \epsilon$ do

At iteration $t$,

1: $\eta_j^{(t)} \leftarrow \log(1^T w_{j-1}^{(t-1)} + 1) - \log\{p - 1^T w_{j-1}^{(t-1)} - 1 + b_j(r, \kappa)\} + \frac{1}{2} \log(\frac{n_j m_j}{2})$

$+ \xi(\frac{n_j}{2}) + \xi(\frac{m_j}{2}) - \frac{3}{2} \log(n + 1) + \frac{1}{2}(n + 1) \log(\tilde{\sigma}_j^2) - \frac{n_j}{2} \log(\tilde{\sigma}_j^2) - \frac{m_j}{2} \log(\tilde{\sigma}_j^2)$

2: $w_j^{(t)} \leftarrow \expit(\eta_j^{(t)})$

Upon convergence of $w$, compute for $i = 1, \ldots, m$

3: $\tilde{y}_j \leftarrow \expit\left[\log\left(\frac{n_j}{m_j}\right) + 1^T w\left\{\log \Gamma\left(\frac{n_j + 1}{2}\right) - \log \Gamma\left(\frac{n_j}{2}\right) + \log \Gamma\left(\frac{m_j + 1}{2}\right) - \log \Gamma\left(\frac{m_j}{2}\right)\right\}

+ \frac{1}{2} w^T\left\{\log \phi(x_{n+i}; \tilde{\mu}_j, \tilde{\sigma}_j^2) - \log \phi(x_{n+i}; \tilde{\mu}_0, \tilde{\sigma}_0^2)\right\}\right]$

and the $j$th entry of $\lambda_{\text{Bayes}}$ is (as $h \to \infty$

$$
\lambda_{\text{Bayes}}(\tilde{x}_j, x_{n+1}, y, y_{n+1}) \rightarrow \lambda_{\text{LRT}}(\tilde{x}_j, x_{n+1}, y, y_{n+1}) + \log(n_1 + y_{n+1}) + \log(n_0 + 1 - y_{n+1})

- \log(2) - 3 \log(n + 1) - 2\xi\{(n_1 + y_{n+1})/2\} + 2\xi\{(n_1 + y_{n+1})/2\}

+ 2\xi\{(n_0 + 1 - y_{n+1})/2\},

= \lambda_{\text{LRT}}(\tilde{x}_j, x_{n+1}, y, y_{n+1}) - 2 \log(n + 1) + O(n_0^{-1} + n_1^{-1}),
$$

where $\xi(x) = \log \Gamma(x) + x - x \log(x) - \frac{1}{2} \log(2\pi)$.

Since the calculation of the marginal likelihood involves a combinatorial sum over $2^{p+1}$ binary combinations, exact Bayesian inference is also computationally impractical in the VQDA setting.

Similar to VLDA, we will use RCVB to approximate the posterior $p(\gamma, y_{n+1} | x, x_{n+1}, y)$ by

$$
q(y_{n+1}, \gamma) = q(y_{n+1}) \prod_{j=1}^{p} q_j(\gamma_j).
$$

This yields the approximate posterior for $\gamma_j$ as

$$
q_j(\gamma_j) \propto \int \exp\left[\mathcal{E}_{-q_j} \{\log p(D, x_{n+1}, y_{n+1}, \gamma, \theta_j)\}\right] d\theta_j,

\propto \exp\left[\mathcal{E}_{-q_j} \left\{\log \mathcal{B}(a_\gamma + 1^T \gamma, b_\gamma + p - 1^T \gamma)\right\} + \frac{1}{2} \mathcal{E}_{-q_j} \left\{\lambda_{\text{Bayes}}(\tilde{x}_j, x_{n+1}, y, y_{n+1})\right\}\right].
$$

For a sufficiently large $n$, we can avoid the need to evaluate the expectation $\mathcal{E}_{-q_j} \left\{\lambda_{\text{Bayes}}(\tilde{x}_j, x_{n+1}, y, y_{n+1})\right\}$.
by applying Taylor’s expansion to obtain the approximation

\[ E \cdot q_j \log(a_j + 1^T \gamma_{-j}) \approx \log(a_j + 1^T w_{-j}), \]

\[ E \cdot q_j \log(b_j + p - 1^T \gamma_{-j} - 1) \approx \log(b_j + p - 1^T w_{-j} - 1), \]

\[ \tilde{\sigma}_j^2 \approx \frac{1}{m_1} ||\mathbf{y}^T (\tilde{x}_j - \tilde{\mu}_j 1) ||^2, \]

\[ \tilde{\sigma}_j^2 \approx \frac{1}{m} ||(1 - \mathbf{y})^T \{\tilde{x}_j - \tilde{\mu}_j 0\} ||^2, \]

\[ \tilde{\mu}_j \approx \frac{1}{m} \mathbf{y}^T \tilde{x}_j, \]

\[ \tilde{\mu}_{j0} \approx \frac{1}{m_0} (1 - \mathbf{y})^T \tilde{x}_j, \]

\[ \tilde{\mu}_j \approx \frac{1}{n} 1^T \tilde{x}_j, \] (16)

and, similar to \( \text{VLDA} \), \( \lambda_{\text{Bayes}} \) does not depend on the new observation \((x_{n+1}, y_{n+1})\). By using the approximation in (16), we have

\[ w_j = \frac{q_j(\gamma_j = 1)}{q_j(\gamma_j = 1) + q_j(\gamma_j = 0)}, \]

\[ \approx \exp \left[ \log(a_j + 1^T w_{-j}) - \log(b_j + p - 1^T w_{-j} - 1) + \frac{1}{2} \log(\frac{n \mu}{\xi}) + \xi(\frac{n}{\phi}) \right. \]

\[ + \left. \xi(\frac{n}{\phi}) - \frac{1}{2} \log(n + 1) + \frac{1}{2} \lambda_{\text{LRT}}(\tilde{x}_j, y_{n+1}) \right], \]

To obtain the approximate density for \( y_{n+1} \), we integrate analytically over \( \theta_1 \) to obtain

\[ q(y_{n+1}) \propto \int \exp \left[ E \cdot p \{ \log p(D, x_{n+1}, y_{n+1}, \gamma, \theta_1) \} \right] d\theta_1, \]

\[ \propto \exp \left[ \log \mathcal{B}(a_y + n_1 + y_{n+1}, b_y + n_0 + 1 - y_{n+1}) + 1^T w \left\{ \log \Gamma(\frac{n+y_{n+1}}{2}) + \log \Gamma(\frac{n+1-y_{n+1}}{2}) \right\} \right. \]

\[ + \left. \frac{1}{2} w^T \left\{ \log \phi(x_{n+1}; \tilde{\mu}_1, \tilde{\sigma}_1^2) - \log \phi(x_{n+1}; \tilde{\mu}_0, \tilde{\sigma}_0^2) \right\} \right], \]

where the \( j \)th element of the \( p \times 1 \) vector \( \phi(x_{n+1}; \tilde{\mu}_j, \tilde{\sigma}_j^2) \) is the Gaussian density

\[ \phi(x_{n+1,j}; \tilde{\mu}_j, \tilde{\sigma}_j^2), \]

and the log prefix denotes an element-wise log of a vector.

In the general case with \( m \) new observations, we may apply Taylor’s expansion results from (16) to compute the approximate classification probability for \( y_{n+1} \) as

\[ \hat{y}_i = \frac{q(y_{n+i} = 1)}{q(y_{n+i} = 1) + q(y_{n+i} = 0)}, \]

\[ \approx \exp \left[ \log \left( \frac{n_0}{n_1} \right) + 1^T w \left\{ \log \Gamma(\frac{n_1+1}{2}) - \log \Gamma(\frac{n_1}{2}) + \log \Gamma(\frac{n_0+1}{2}) - \log \Gamma(\frac{n_0}{2}) \right\} \right. \]

\[ + \left. \frac{1}{2} w^T \left\{ \log \phi(x_{n+i}; \tilde{\mu}_1, \tilde{\sigma}_1^2) - \log \phi(x_{n+i}; \tilde{\mu}_0, \tilde{\sigma}_0^2) \right\} \right]. \]

The RCVB algorithm for \( \text{VLQDA} \) may be found in Table 3.

**B Supplementary material to section 5**

**B.1 Proof of Theorem 1**

*Proof.* Recall that

\[ E^{(1)} = e_0^{(1)} + e_1^{(1)}, \]

21
where $e_0^{(1)}$ is the sum of Type I errors after the first RCVB cycle and $e_1^{(1)}$ is the sum of Type II errors. We shall show separately that each of the two sums converge in probability to 0.

We begin by showing that

$$ e_0^{(1)} = o_p(1). \tag{17} $$

From the main paper, recall that

$$ w_j^{(t)} = \exp \left[ \frac{\lambda(\bar{X}_l)}{2} - \frac{1}{2} \log(n + 1) + \log \left\{ a_{\gamma} + 1^T w_{-j}^{(r-1)} \right\} - \log \left\{ b_{\gamma} + p - 1^T w_{-j}^{(r-1)} - 1 \right\} \right]. $$

Following the expression for $w_j^{(t)}$ in main paper and writing $\lambda = \lambda_{\text{LRT}}$, the LHS is

$$ e_0^{(1)} \leq \sum_{l \in \mathcal{J}_0} \exp \left\{ \frac{\lambda(\bar{X}_l)}{2} - \frac{1}{2} \log(n + 1) + \log \left( \frac{n}{n + 1} \right) - \log \left( b_{\gamma} + p - 1 \right) \right\}, \tag{18} $$

where the first line follows from the initialisation $1^T w^{(0)}_{-j} \leq p - 1$. The inequality in (18) follows by applying the summation over each term. We examine the convergence of the term $\max_{l \in \mathcal{J}_0} \lambda(\bar{X}_l)$. Following Result 1, we have

$$ \max_{l \in \mathcal{J}_0} \lambda(\bar{X}_l) = (n + 1) \log \left\{ 1 + \frac{\max_{l \in \mathcal{J}_0} F_l}{n - 2} \right\}, $$

where each $F_l$ are independent, identically distributed as the F-distribution with 1 and $n - 2$ degrees of freedom which we shall denote as $\mathcal{F}_{1,n-2}$. This allows us to re-write the bound in (18) as

$$ \exp \left[ \frac{1}{2} (n + 1) \log \left\{ 1 + \frac{M_n}{n - 2} \right\} - \frac{\kappa(n + 1)}{\log(n + 1)} \right], \tag{19} $$

where $M_n = \max_{l \in \mathcal{J}_0} F_l$ is the maximum among $p_n$ random variables that follow $\mathcal{F}_{1,n-2}$ and we have substituted the hyperparameter setting for $b_{\gamma} = p_n \exp[\kappa(n + 1)/(\log(n + 1))^r]/\sqrt{n + 1}$ for some $r < 1$ and $\kappa > 0$ from Section 2 of the main paper.

To prove (17), we need to show that the exponential term in (19) converges in probability to 0. This follows by applying Result 7 to obtain

$$ \frac{1}{2} (n + 1) \log \left\{ 1 + \frac{M_n}{n - 2} \right\} - \frac{\kappa(n + 1)}{\log(n + 1)} \overset{p}{\rightarrow} -\infty \tag{20} $$

and hence our proof of (17) is completed.

Next, we show that $e_1^{(1)} = o_p(1)$. We begin by bounding

$$ e_1^{(1)} \leq \sum_{l \in \mathcal{J}_1} \exp \left\{ -\frac{\lambda(\bar{X}_l)}{2} + \frac{1}{2} \log(n + 1) + \log(b_{\gamma} + p - 1) \right\}, \tag{21} $$

To further bound (21), we need a lower bound for $F_l/(n - 2)$ that does not depend on $l$. From Result 1 we
may compute it as follows.

\[
\frac{F_1}{n-2} = \exp\left[\log\left(n_1(\hat{\mu}_{11} - \hat{\mu})^2 + n_0(\hat{\mu}_{00} - \hat{\mu})^2\right) - \log\left((n + 1)\hat{\sigma}_{11}^2\right)\right],
\]

\[
= \exp\left[2\log(d_l) + \log\left(n_1\left(\frac{a_n}{n}\right)^2 + n_0\left(\frac{a_n}{n}\right)^2\right) - \log\left((n + 1)\hat{\sigma}_{11}^2\right)\right],
\]

\[
\geq \exp\left[2\log\left(d_{(1)}\right) + \log\left(\frac{(d_{(1)})^2}{n}\right) + \left(\frac{v_n}{n}\right)^2 - \log\left(V_n\right) + \log(n - 2) - \log\left(\max_{i\in\mathcal{J}_1} \sigma_{i1}^2\right)\right],
\]

\[
= B_n,
\]

where

\[
d_l = |\hat{\mu}_{11} - \hat{\mu}|, \quad d_{(1)} = \min_{i\in\mathcal{J}_1} d_l \quad \text{and} \quad V_n = (n + 1) \sum_{i\in\mathcal{J}_1} \hat{\sigma}_{i1}^2.
\]

Since each pair of \((\tilde{X}_j, \tilde{X}_l)\) are independent for \(j \neq l\) and that each

\[
(n + 1)\hat{\sigma}_{11}^2/\sigma_{11}^2 = \frac{1}{\sigma_{11}^2} \left[\|y \odot \tilde{X}_j - \hat{\mu}_{j1}1\|^2 + \|(1 - y) \odot \tilde{X}_j - \hat{\mu}_{j0}1\|^2\right] \sim \chi^2_{n-2},
\]

where \(\odot\) is the Hadamard product, this allows us to compute the exact distribution of \(V_n\) as

\[
V_n \sim \chi^2_{p_1n(n-2)}.
\]

Furthermore, condition 1 and Result 8 implies that \(d_{(1)}\) converges in probability to a positive constant. Consequently, the lower bound \(B_n\) also converges in probability to a positive constant. By substituting this back into (21), the upper bound for \(e_1^{(1)}\) may be simplified as

\[
p_{1n} \exp\left[-(n + 1)\left\{\frac{1}{2} \log\left(1 + B_n\right) - A_n\right\}\right],
\]

where

\[
A_n = (n + 1)^{-1} \left\{2\log(p_n) - \frac{1}{2} \log(n + 1) + \frac{n(n + 1)}{\{\log(n + 1)\}^r}\right\}.
\]

Following condition 3, observe that \(A_n \rightarrow 0\). This completes our proof.

\[\square\]

### B.2 Supporting results

In this section, we shall provide details of the supporting results required for Theorem 1. Here, all sample moment estimators used correspond to their Taylor’s series approximation.

We begin by deriving the exact distribution of the likelihood ratio statistic, \(\lambda_{LRT}(\tilde{X}_j)\). This will be carried out by expressing the statistic as an increasing function of a F-statistic. The result is a first step to analyse the asymptotic behaviour of the maximum of likelihood ratio statistics in (18) among the truly non-discriminative variables. It is also used to derive lower bounds of likelihood ratio statistics in (21) among the truly discriminative variables.

**Result 1.** For \(1 \leq j \leq p_n\), we have

\[
\lambda_{LRT}(\tilde{X}_j) = (n + 1) \log\left\{1 + \frac{F_j}{n-2}\right\},
\]

where

\[
F_j = \hat{\sigma}_{j1}^{-2} \left(1 - \frac{3}{n+1}\right) \left\{n_1(\hat{\mu}_{j1} - \hat{\mu})^2 + n_0(\hat{\mu}_{j0} - \hat{\mu})^2\right\}.
\]
Proof. From the expression for $\lambda_{\text{LRT}}(\tilde{X}_j)$ in the main paper, we have

$$
\lambda_{\text{LRT}}(\tilde{X}_j) = (n + 1) \log \left( \frac{\hat{\sigma}_j^2}{\hat{\sigma}_j^1} \right),
$$

$$
= (n + 1) \log \left( 1 + \frac{\hat{\sigma}_j^2 - \hat{\sigma}_j^1}{\hat{\sigma}_j^1} \right),
$$

$$
= (n + 1) \log \left( 1 + \frac{F_j}{n - 2} \right).
$$

The last equality holds by observing that $(n + 1)\left\{ \hat{\sigma}_j^2 - \hat{\sigma}_j^1 \right\}$ is the difference between total sum of squares and residual sum of squares of a one-way ANOVA model with 2 groups (see Wu and Hamada [2009]).

Next, we will present intermediate results that lead us to the convergence in Equation (20). The first result is Hölder’s inequality and is used to bound the rate at which $E(M_n)$ increases.

**Result 2.** Let $X$ and $Y$ be two random variables, $s$ and $t$ be two constants. If $s,t \in [1, \infty]$ with $\frac{1}{s} + \frac{1}{t} = 1$ then

$$
E|XY| \leq \|X\|_s \|Y\|_t.
$$

Here, $\|X\|_s = (E|X|^s)^{1/s}$ for $s \in [1, \infty]$; $\|X\|_\infty = \inf\{m : P(|X| > m) = 0\}$.

Using Hölder’s inequality, the bound of $E(M_n)$ may be expressed as a function of the $s$-th order moment of the F-distribution with 1 and $n - 2$ degrees of freedom. The general formula of its moments are given in the next result.

**Result 3.** Let $F$ be a random variable that follows the F-distribution with 1 and $n - 2$ degrees of freedom. For any $s$ such that $s/2 < n - 2$, the $s$-th order moment is

$$
E(F^s) = \exp \left\{ s \log(n - 2) + \log \Gamma(s + \frac{1}{2}) + \log \Gamma(\frac{n - 2}{2} - s) - \Gamma(\frac{1}{2}) - \log \Gamma(\frac{n - 2}{2}) \right\}.
$$

Using the previous two results, the bound for $E(M_n)$ may be expressed as a function of beta functions and ratio of gamma functions. As such, the following results are required for us to proceed with their asymptotic analysis. Result 4 can be found in Graham and Knuth [1994], whereas Result 5 follows from the Stirling series of expansion of a log-gamma function.

**Result 4.** For any $n > 0$, the ratio of gamma functions

$$
\frac{\Gamma(n + 1/2)}{\Gamma(n)} = \sqrt{n} \left( 1 - \frac{1}{8n} + \frac{1}{128n^2} - \frac{5}{1024n^3} + \frac{21}{32768n^4} + \ldots \right),
$$

$$
= O(n^{1/2}).
$$

**Result 5.** Let $B$ denote the beta function and

$$
g(n) := \log B(a_n, b_n),
$$

where

$$
a_n := \frac{n - 3}{2(\log(n))^u} + \frac{1}{2} \quad \text{and} \quad b_n = \frac{n - 2}{2} - \frac{n - 3}{2(\log(n))^u},
$$

for any constant $0 < u < 1$. The function

$$
n^{-1} g(n) \{\log(n)\}^u \to -\infty
$$

as $n \to \infty$.  

24
Proof. We shall calculate an upper bound for \( g(n) \) and show that it converges to \(-\infty\) when multiplied by 
\( n^{-1}\{\log(n)\}^n \). Following the Stirling’s series expansion for the log-gamma function, i.e.
\[
\log \Gamma(x) = x \log(x) - x - \frac{1}{2} \log \left( \frac{x}{2\pi} \right) + o(1),
\]
we have the following asymptotic result
\[
g(n) \sim a_n \log\left\{ \frac{a_n}{(a_n + b_n)} \right\} + b_n \log\left\{ \frac{b_n}{(a_n + b_n)} \right\} + \frac{1}{2} \log\left\{ a_n^{-1} + b_n^{-1} \right\} + \frac{1}{2} \log(2\pi),
\]
\[
\leq a_n \log\left\{ \frac{a_n}{(a_n + b_n)} \right\} + b_n \log\left\{ \frac{b_n}{(a_n + b_n)} \right\},
\]
\[
= -\frac{a_n}{2} \log(n) - b_n \log\left\{ 1 + O\{\log(n)\}^{-u} \right\} + o(1).
\]
By multiplying both sides by \( n^{-1}\{\log(n)\}^n \), the equation becomes
\[
n^{-1} g(n) \{\log(n)\}^n \leq -u \log \log(n) - 1 + o(1),
\]
which converges to \(-\infty\) as \( n \) diverges.

We are now ready to demonstrate the required convergence in Equation (20). This result involves a sequence of maximums of iid F-statistics that is described in the following two-dimensional array with infinite number of rows and columns:

| Row number | Column 1 | Column 2 | Column 3 | ... | Sequence, \( \{M_n\} \) |
|------------|----------|----------|----------|-----|--------------------------|
| Row 3      | \( F_{3,1} \) | \( F_{3,2} \) | \( F_{3,3} \) | \( \ldots \) | \( \max_{1 \leq j \leq p_n} F_{3,j} \) |
| Row 4      | \( F_{4,1} \) | \( F_{4,2} \) | \( F_{4,3} \) | \( \ldots \) | \( \max_{1 \leq j \leq p_n} F_{4,j} \) |
| Row 5      | \( F_{5,1} \) | \( F_{5,2} \) | \( F_{5,3} \) | \( \ldots \) | \( \max_{1 \leq j \leq p_n} F_{5,j} \) |
| \( \vdots \) | \( \ddots \) | \( \ddots \) | \( \ddots \) | \( \vdots \) | \( \vdots \) |
| Row \( n \) | \( F_{n,1} \) | \( F_{n,2} \) | \( F_{n,3} \) | \( \ldots \) | \( \max_{1 \leq j \leq p_n} F_{n,j} \) |
| \( \vdots \) | \( \ddots \) | \( \ddots \) | \( \ddots \) | \( \vdots \) | \( \vdots \) |

The F-statistics in each row of the array are iid random variables that follow \( F_{1,n-2} \), where \( n \) corresponds to the row index; the maximums are over the first \( p_n \) F-statistics where \( p_n \) is an increasing sequence such that \( p_n \leq \tilde{p}_n = \exp\{(n + 1)/\log(n + 1)\} \). We shall establish convergence properties for \( \{M_n\}_{n=3}^{\infty} \) in the next two results. In Result 6, we derive a non-trivial bound on their respective expectations that gives the rate at which the expectation increases. This bound is subsequently used to show the required convergence in equation (20) through Result 7.

**Result 6.** Suppose \( \{M_n\}_{n=3}^{\infty} \) is a sequence of random variables such that
\[
M_n = \max_{1 \leq j \leq p_n} F_{n,j},
\]
where \( \{p_n\} \) is an increasing sequence and \( \{F_{n,j}\}_{j=1}^{\infty} \) are iid \( F_{1,n-2} \) random variables. If each \( p_n \leq \tilde{p}_n \), then
\[
\mathbb{E}(M_n) \leq \exp\{\log(n - 2) + \tilde{m}_n(u)\},
\]
where
\[
\tilde{m}_n(u) = -u \log \log(n) - 1 + o(1).
\]
for any \( u < 1 \).
Proof. By using the pdf of the first order statistic, we express the expectation of $M_n$ as an integral

$$
\mathbb{E}(M_n) = p_{0n} \int_0^\infty x \frac{dG(x)}{dx} \left( \mathbb{P}(X \leq x) \right)^{p_{0n}-1} dx,
$$

$$
= p_{0n} \mathbb{E} \left( XG(X)^{p_{0n}-1} \right),
$$

$$
\leq p_{0n} \mathbb{E}(X^{c_n})^{1/c_n} \mathbb{E}(U^{d_n})^{d_n-1/d_n},
$$

where the last line follows from an application of Hölder’s inequality (see Result 2), $X$ is a $F_{1,n-2}$ random variable with cdf $G$, the random variable $U$ follows the standard uniform distribution and the fixed sequences are defined as

$$
c_n := \frac{1}{2} (n-3)(\log(n))^{-u}, \quad d_n := 1/(1-c^{-1}).
$$

for any $u < 1$. By further substituting the expression for moments of an F-distribution (see Result 3), the bound becomes

$$
\mathbb{E}(M_n) \leq \exp \left[ \log(n-2) + (1-d^{-1}) \log \tilde{p}_n + c^{-1} \log \mathbb{E}(X^{c_n}) + d^{-1} \log \left\{ d_n(1-\tilde{p}_n^{-1}) + \tilde{p}_n^{-1} \right\} \right],
$$

$$
= \exp \left[ \log(n-2) + c^{-1} \left\{ \log \tilde{p}_n + \log \mathcal{B}(a_n, b_n) + \log \Gamma \left\{ (n-1)/2 \right\} - \log \Gamma \left\{ n/2 - 1 \right\} \right\} 
\right. 
$$

$$
\left. + d^{-1} \log \left\{ d_n(1-\tilde{p}_n^{-1}) + \tilde{p}_n^{-1} \right\} \right],
$$

where the fixed sequences $a_n$ and $b_n$ have been defined in Result 3. To simplify components of the exponent term-by-term at the RHS, we make use of the asymptotics in Results 4, 5 and 6 to obtain

$$
c^{-1} \log \tilde{p}_n = O \left( (\log(n))^{-u-1} \right), \quad d^{-1} \log \left\{ d_n(1-\tilde{p}_n^{-1}) + \tilde{p}_n^{-1} \right\} = o(1),
$$

$$
c^{-1} \left[ \log \Gamma \left\{ (n-1)/2 \right\} - \log \Gamma \left\{ n/2 - 1 \right\} \right] = O \left( n^{-1}(\log(n))^{-u+1} \right),
$$

and

$$
c^{-1} \log \mathcal{B}(a_n, b_n) = -u \log \log(n) - 1 + o(1).
$$

The expression for $\tilde{m}_n$ follows by considering that all terms except $c^{-1} \log \mathcal{B}(a_n, b_n)$ converges to 0. \hfill \square

**Result 7.** Suppose $\{M_n\}_{n=3}^\infty$ is a sequence of random variables such that

$$
M_n = \max_{1 \leq j \leq p_{0n}} F_{n,j},
$$

where $\{p_{0n}\}$ is an increasing sequence and $\{F_{n,j}\}_{j=1}^\infty$ are iid $F_{1,n-2}$ random variables. If each $p_{0n} \leq \tilde{p}_n$, then

$$
\left\{ (\log(n))^{r} \log \left\{ 1 + \frac{M_n}{n-2} \right\} \right\} \overset{p}= 0.
$$

for any $r < 1$.

**Proof.** Since the expression on the LHS is non-negative, we may apply Markov’s inequality to obtain the bound

$$
\mathbb{P} \left[ \left\{ (\log(n))^{r} \log \left\{ 1 + \frac{M_n}{n-2} \right\} \right\} > \epsilon \right] \leq \epsilon^{-1} \mathbb{E} \left[ \left\{ (\log(n))^{r} \log \left\{ 1 + \frac{M_n}{n-2} \right\} \right\} \right].
$$

Hence, it suffices for us to bound the expectation on the right and show that the bound converges to 0.

By Jensen’s inequality, we have

$$
\mathbb{E} \left[ (\log(n))^{r} \log \left\{ 1 + \frac{M_n}{n-2} \right\} \right] \leq (\log(n))^{r} \log \left\{ 1 + \mathbb{E} \left( \frac{M_n}{n-2} \right) \right\} \leq (\log(n))^{r} \mathbb{E} \left( \frac{M_n}{n-2} \right)
$$

for any $r < 1$. \hfill \square
where the last line follows from the inequality $\log(1 + z) \leq z$ for any $z > -1$. With reference to Result 6 we can always find a constant $u \in (r, 1)$ such that the expectation $E \{M_n/(n - 2)\}$ on the RHS may be bounded by the expression

$$E \left( \frac{M_n}{n - 2} \right) \leq \exp \{\tilde{m}_n(u)\},$$

$$\leq \exp \{-u \log \log(n) + o(1)\}.$$  

By multiplying $\{\log(n)\}^r$ to both sides of the equation, we have

$$\{\log(n)\}^r E \left( \frac{M_n}{n - 2} \right) \leq \exp \{-u - r \log \log(n) + o(1)\}$$

and hence our result follows by noting that $r < u$. \[\Box\]

The next result is used to justify the convergence of the bound in (22) to a positive constant. This requires us to show that the minimum absolute differences between the group-conditional sample means converges in probability to a positive constant among the truly discriminative variables.

**Result 8.** Consider a finite sequence of independent random variables $\{d_j\}_{j=1}^{\mathcal{C}}$ such that each $d_j = \delta_j + o_p(1)$ for a positive constant $\delta_j$. Let $p_{1n}$ denote a fixed, increasing sequence in $n$ that is bounded above by $C$ and

$$d_{n,(1)} = \min_{1 \leq j \leq p_{1n}} d_j$$

as the minimum of the first $p_{1n}$ of the random sequence. If $E(d_j) = m_j > 0$ and $\text{Var}(d_j) < \infty$, then

$$d_{n,(1)} = \delta + o_p(1),$$

where $\delta$ is a positive constant.

**Proof.** The result is equivalent to showing that $P(d_{n,(1)} > 0) \rightarrow 1$. We can arrive at this required result by bounding

$$P(d_{n,(1)} > 0) \geq P(\min_{1 \leq j \leq C} d_j > 0),$$

$$= \prod_{j=1}^{C} P(d_j > 0).$$

Since each $d_j \overset{p}{\rightarrow} \delta_j > 0$, our result holds by taking limits with respect to $n$ on both sides of the above inequality. \[\Box\]

**References**

Ahdesmäki, M., Strimmer, K., 2010. Feature selection in omics prediction problems using CAT score and false discovery rate control. The Annals of Applied Statistics 4 (1), 503–519.

Benjamini, Y., Daniel, Y., 2001. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29 (4), 1165—1188.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57 (1), 289—300.
Bickel, P. J., Levina, E., 2004. Some theory for Fisher’s linear discriminant function, ‘naïve Bayes’ and some alternatives when there are many more variables than observations. Bernoulli 10 (6), 989–1010.

Blei, D. M., Jordan, M. I., 2006. Variational inference for Dirichlet processes. Bayesian Analysis 1 (1), 121–144.

Blei, D. M., Kucukelbir, A., McAuliffe, J. D., 2017. Variational inference: A review for statisticians. Journal of American Statistical Association 112 (518), 859–877.

Bonferroni, C. E., 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Cai, T., Liu, W., 2011. A direct estimation approach to sparse linear discriminant analysis. Journal of the American Statistical Association 106 (496), 1566–1577.

Chen, Y., Feng, J., 2014. Efficient method for Moore-Penrose inverse problems involving symmetric structrue based on group theory. Journal of Computing in Civil Engineering 28 (2), 182–190.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Mining 10 (35), 1–17.

Clemmensen, L., Witten, D., Hastie, T., Ersboll, B., 2011. Sparse discriminant analysis. Technometrics 53 (4), 406–413.

Courrieu, P., 2005. Fast computation of Moore-Penrose inverse matrices. Neural Information Processing - Letters and Reviews 8 (2), 25–29.

Craig-Shapiro, R., Kuhn, M., Xiong, C., Pickering, E. H., Liu, J., Misko, T. P., Perrin, R. J., Bales, K. R., Soares, H., Fagan, A. M., David, M. H., 2011. Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer’s disease diagnosis and prognosis. PLoS ONE, e18850.

Donoho, D., Jin, J., 2004. Higher criticism for detecting sparse heterogeneous mixtures. The Annals of Statistics 32 (3), 962—994.

Donoho, D., Jin, J., 2008. Higher criticism thresholding. optimal feature selection when useful features are rare and weak. Proceedings of the National Academy of Sciences 105 (39), 14790–14795.

Dudoit, S., Fridyland, J., Speed, T. P., 2002. Comparison of discrimination methods for classification of tumours using gene expression data. Journal of American Statistical Association 97 (457), 77–87.

Erickson, B. J., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., Rieger-Christ, K., Lemmerman, J., 2016. Radiology Data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma [TCGA-LIHC] collection. The Cancer Imaging Archive.

Fan, J., Fan, Y., 2008. High-dimensional classification using features annealed independence rules. The Annals of Statistics 36 (6), 2605–2637.

Fernández-Delgado, M., Cernadas, E., Barro, S., 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning 15, 3133–3181.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (2), 179–188.
Friedman, J. H., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84 (405), 165–175.

Friguet, C., Kloareg, M., Causeur, D., 2009. A factor model approach to multiple testing under dependence. Journal of the American Statistical Association 104 (488), 1406–1415.

Graham, R. L., Knuth, D. E., 1994. Concrete mathematics: a foundation for computer science. 2nd edition. Addison-Wesley.

Hall, P., Pham, T., Wand, M. P., Wang, S. S. J., 2011. Asymptotic normality and valid inference for Gaussian variational approximation. The Annals of Statistics 39 (5), 2502–2532.

Hastie, T., Tibshirani, R., Narasimhan, B., Chu, G., 2014. pamr: Prediction analysis for microarrays. R package version 1.55.

Jorissen, R. N., Lipton, L., Gibbs, P., Chapman, M., Desai, J., Jones, I. T., Yeatman, T. J., East, P., Tomlinson, I. P., Verspaget, H. W., Aaltonen, L. A., Kruhoffer, M., Orntoft, T. F., Andersen, C. L., Sieber, O. M., 2008. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. Clinical Cancer Research 14 (24), 8061–8069.

Liu, J. J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L., Ling, X. B., 2005. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 21 (11), 2691—2697.

Luts, J., Ormerod, J. T., 2014. Mean field variational Bayesian inference for support vector machine classification. Computational Statistics and Data Analysis 73, 163–176.

Matthews, B. W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405 (2), 442–451.

Ormerod, J. T., Stewart, M., Yu, W., Romanes, S., 2017. Bayesian hypothesis test with diffused priors: Can we have our cake and eat it too? ArXiv.

Ormerod, J. T., Wand, M. P., 2010. Explaining variational approximations. The American Statistician 64 (2), 140–153.

Pedro, D. S. A., 2011. Two group classification with high-dimensional correlated data: a factor model approach. Computational Statistics and Data Analysis 55 (11), 2975–2990.

Pedro, D. S. A., 2015. HiDimDA: High dimensional discriminant analysis. R package version 0.2-4.

Perthame, E., Friguet, C., Causeur, D., 2016. Stability of feature selection in classification issues for high-dimensional correlated data. Statistics and Computing 26 (4), 783–796.

Perthame, E., Friguet, C., Causeur, D., 2018. FADA: Variable Selection for Supervised Classification in High Dimension. R package version 1.3.3.

Safo, S. E., Ahn, J., 2016. General sparse multi-class linear discriminant analysis. Computational Statistics and Data Analysis 99, 81–90.

Shaffer, J. P., 1995. Multiple hypothesis testing. Annual Review of Psychology 46, 561—584.
Shao, J., Wang, Y., Deng, X., Wang, S., 2011. Sparse linear discriminant analysis with applications to high dimensional data. Annals of Statistics 39 (2), 1241–1265.

Storey, J. D., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Annals of Statistics 31 (6), 2013—2035.

Teh, Y. W., Newman, D., Welling, M., 2007. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems. Vol. 19. MIT Press, pp. 1353–1360.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science 18 (1), 104–117.

van der Maaten, L., Hinton, G., 2017. Visualising data using t-sne. Journal of Machine Learning Research 9, 2579–2605.

Wang, Y., Blei, D., 2018. Frequentist consistency of variational bayes. Journal of the American Statistical Association 9, 1–15.

Witten, D., 2011. Classification and clustering of sequencing data using a Poisson model. Annals of Applied Statistics 5 (4), 2493–2518.

Witten, D., 2015. penalizedLDA: Penalized Classification using Fisher’s Linear Discriminant. R package version 1.1.

Witten, D., Tibshirani, R., 2011. Penalized classification using Fisher’s linear discriminant. Journal of Royal Statistical Society Series B 73 (5), 754–772.

Wu, C. F. J., Hamada, M. S., 2009. Experiments: planning, analysis, and optimization. 2nd edition. Wiley Series in Probability and Statistics. Wiley.

Zhang, A., Zhou, H., 2017. Theoretical and computational guarantees of mean field variational inference for community detection. ArXiv.