Statistical Inference in Classification of High-Dimensional Gaussian Mixture

Hanwen Huang
Department of Biostatistics, Data Science and Epidemiology
Medical College of Georgia, Augusta University, Augusta, GA, 30912
hhuang1@augusta.edu

Peng Zeng
Department of Mathematics & Statistics
Auburn University, Auburn, AL 36849
zengpen@auburn.edu

April 16, 2025

Abstract

We consider the classification problem of a high-dimensional mixture of two Gaussians with general covariance matrices. Using the replica method from statistical physics, we investigate the asymptotic behavior of a broad class of regularized convex classifiers in the limit where both the sample size n and the dimension p approach infinity while their ratio $\alpha = n/p$ remains fixed. This approach contrasts with traditional large-sample theory in statistics, which examines asymptotic behavior as $n \to \infty$ with p fixed. A key advantage of this asymptotic regime is that it provides precise quantitative guidelines for designing machine learning systems when both p and n are large but finite. Our focus is on the generalization error and variable selection properties of the estimators. Specifically, based on the distributional limit of the classifier, we construct a de-biased estimator to perform variable selection through an appropriate hypothesis testing procedure. Using L_1-regularized logistic regression as an example, we conduct extensive computational experiments to verify that our analytical findings align with numerical simulations in finite-sized systems. Additionally, we explore the influence of the covariance structure on the performance of the de-biased estimator.

Keywords— De-biased estimator; Generalization error; Logistic regression; L_1-regularization; Variable selection.

1 Introduction

It is well known that, with the advancement of computer technology, modern statistical problems are increasingly high dimensional, that is, the number of parameters p is large. Examples abound in genetic molecular measurements (where numerous gene-level features are recorded), chemometrics (high-dimensional spectral data), and medical image analysis (where 3D shapes are represented by high-dimensional vectors), among others. At the same time, the proliferation of large-scale data collection, particularly via the Internet, requires consideration of cases where the sample size n is also large. Consequently, developing efficient inference tools to address challenges in large-scale data, where both n and p are enormous, is a crucial task in statistical modeling. However, many existing statistical theories and methods, originally developed for small-data problems, struggle to scale effectively to such high-dimensional settings.

In this paper, we focus on classification methods that fall within a general regularization framework, formulated as the minimization of penalized loss functions. This framework encompasses many commonly used classification methods, such as logistic regression and support vector machines (SVM), as special cases. The statistical inference of this class of methods in low-dimensional settings has been extensively studied in the literature (see, e.g., Blanchard et al. (2008); Wang et al. (2019)). To address high-dimensional problems, where the number of parameters can even exceed the number of samples, sparse regularization techniques, such as L_1-penalty, are commonly employed. However, sparse regularization comes at a cost: the distributions of sparse estimators are typically intractable. As a result, building an inference procedure to quantify the uncertainty of the estimated parameters is very challenging. This contrasts with classical statistics, where exact distributions are available or can be approximated using large-sample asymptotic theorems.

Recent rapid advances in statistical theory regarding the asymptotic performance of many classic machine learning algorithms in the limits of both large n and large p have shed light on this issue. Particularly noteworthy is the considerable effort devoted to establishing asymptotic results for regularized convex classification methods under the assumption that n and p grow at the same rate, i.e., $n/p \to \alpha > 0$. For example, Huang (2017); Mai and Couillet (2018) derived the asymptotic results of SVM under Gaussian mixture models, where data is assumed to be generated from a spiked Gaussian mixture distribution with two components, one for each class. In the same setting, Mai et al. (2019); Huang and Yang (2021) studied L_2-regularized logistic regression and general margin-based classification methods, respectively. Montanari et al. (2019) analyzed the hard-margin SVM under the single Gaussian model, in which data is assumed to be generated from a single Gaussian distribution. Deng et al. (2019) examined unregularized logistic regression under two-component Gaussian mixture models. The sharp asymptotics for unregularized logistic regression were studied in Candès et al. (2020) under the single Gaussian model. Gerace et al. (2020) investigated the classification error for L_2-regularized logistic regression for a single Gaussian model with a two-layer neural network covariance structure. Analogous results for Gaussian mixture models with standard Gaussian components were provided in Mignacco et al. (2020). The sharp asymptotics of generic convex generalized linear models were studied in Kabashima (2008); Takahashi and Kabashima (2022); Gerbelot et al. (2023) for rotationally invariant Gaussian data and in Loureiro et al. (2021) for block-correlated Gaussian data. Multiclass classification for Gaussian mixture models was also examined in Loureiro et al. (2021).

Although much of the existing literature focuses on analyzing the generalization error of classifiers, the impact of regularization on variable selection remains largely unexplored. Variable selection plays a crucial role in real-world applications, particularly in medical research. For example, machine learning models are often developed to classify whether a patient has a specific type of cancer based on gene expression data (Verhaak et al., 2010). However, each patient is represented by thousands of genes, only a small subset of which are truly relevant for distinguishing cancerous from non-cancerous cases. A well-known drawback of standard classification methods is their susceptibility to performance degradation when all available genes are included, as many may be irrelevant or redundant (Hastie et al., 2009). In fact, Fan and Fan (2008) has shown that in high-dimensional settings, using all features for classification can result in performance as poor as random guessing due to noise accumulation. Therefore, variable selection is essential for improving classification accuracy. Additionally, it enables doctors and biologists to better understand which genes—or groups of genes—are associated with cancer. To address this issue, various methods have been proposed. In particular, a unified approach that simultaneously performs variable selection and prediction using appropriate sparsity regularization, such as L_1-regularization, has been shown to achieve superior performance.

The goal of this paper is to construct statistical inference procedures for the L_1-regularized convex classification methods in high-dimensional settings, that is, to calculate confidence intervals and p-values for parameters estimated from the model. Our results are based on the asymptotic behavior of the estimators in the limit of both $n \to \infty$ and $p \to \infty$ at a fixed rate $n/p \to \alpha$ under two-component Gaussian mixture models. We derive the analytical results using the replica method developed in statistical mechanics. All analytical results are confirmed by numerical experiments on finite-size systems and thus our formulas are verified to be correct.

Note that some of the results in the aforementioned literature have been rigorously established under the Gaussian assumption. As summarized by Huang and Yang (2021), rigorous analytical methods include convex random geometry (Candès et al., 2020), random matrix theory (Dobriban and Wager, 2018), message-passing algorithms (Bayati and Montanari, 2011; Berthier et al., 2020; Loureiro et al., 2021), the convex Gaussian min-max theorem (Montanari et al., 2019; Mignacco et al., 2020; Deng et al., 2019), and interpolation techniques (Barbier and Macris, 2018). Most rigorous work to date has focused on i.i.d. randomness, corresponding to the case of standard Gaussian design. In contrast, the present study considers a mixture of two Gaussian components with an arbitrary covariance structure. Although deriving a rigorous proof for our results remains an open problem, we provide numerical evidence through simulations on moderate system sizes, demonstrating that the theoretical formula holds in the high-dimensional limit.

The remainder of this paper is organized as follows. In Section 2, we present the asymptotic results in the joint limit of large p and n for general L_1-regularized convex classification methods. In Section 3, using L_1-regularized logistic regression as an example, we conduct numerical studies comparing the theoretical results with Monte Carlo simulations on finite-size systems under different experimental settings. The final section concludes the paper. The derivation of the main analytical results, along with the development of the computing algorithm for implementing the corresponding statistical inference procedures, is provided in the Appendix.

2 De-biased estimator of penalized classification

In binary classification, we are typically given a random sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i \in \{1, -1\}$ denotes the categorical label and $\mathbf{x}_i \in \mathbb{R}^p$ represents the input covariates. The goal of linear classification is to estimate a vector $\mathbf{w} \in \mathbb{R}^p$ such that $\mathrm{sign}(\mathbf{x}^T\mathbf{w})$ can be used to predict output labels for future observations based only on their input covariates $\mathbf{x}$. This paper focuses on regularized convex classification methods that can be formulated within the following *loss* + *penalty* framework

$$\hat{\mathbf{w}} \quad = \quad \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{p}}\right) + \sum_{j=1}^{p} J_\lambda(w_j) \right\}, \tag{1}$$

where $V(u)$ is the convex loss function, and $J_\lambda(w) = \lambda|w|$ represents the L_1-regularization term with regularization parameter λ. The general requirements for the loss function are that it is convex, decreasing, and satisfying $V(u) \to 0$ as $u \to \infty$. Many commonly used classification techniques fit this regularization framework. Examples include penalized logistic regression (PLR; Lin et al. (2000)) and support vector machine (SVM; Vapnik (1995)). The loss functions associated with these methods are:

$$\begin{aligned} \text{PLR}: &\quad V(u) = \log[1+\exp(-u)], \\ \text{SVM}: &\quad V(u) = (1-u)_+. \end{aligned}$$

Beyond these methods, many other classification techniques also fit within this regularization framework. Examples include distance-weighted discrimination (DWD; Marron et al. (2007)), the unified machine with large margin (Liu et al., 2011), AdaBoost in Boosting (Freund and Schapire, 1997; Friedman et al., 2000), the import vector machine (IVM; Zhu and Hastie (2005)), and ψ-learning (Shen et al., 2003).

Our goal is to study the high-dimensional limiting behavior of $\hat{\mathbf{w}}$. Similar to the argument presented in Javanmard and Montanari (2013); van de Geer et al. (2014) regarding estimators for generalized linear models, the idea behind our construction is based on the following straightforward heuristic derivation. Define $\sum_{i=1}^n V\left(y_i \mathbf{x}_i^T \mathbf{w}/\sqrt{p}\right) = \mathcal{L}(\mathbf{w})$ and $\sum_{j=1}^p J_\lambda(w_j) = \lambda \mathcal{R}(\mathbf{w})$, then $\hat{\mathbf{w}}$ is given by

$$\hat{\mathbf{w}} \;=\; \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left\{\mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w})\right\}.$$

Using the Karush–Kuhn–Tucker (KKT) conditions, we obtain

$$\nabla\mathcal{L}(\hat{\mathbf{w}}) + \lambda\hat{\boldsymbol{\kappa}} = 0, \tag{2}$$

where the vector $\hat{\boldsymbol{\kappa}}$ arises from the subgradient of $\|\mathbf{w}\|_1$, with $\hat{\kappa}_j = \text{sign}(\hat{w}_j)$ if $\hat{w}_j \neq 0$ and 0 otherwise. Let $\mathbf{w}_0 \in \mathbb{R}^p$ denote the true parameter value, defined as the minimizer of the population loss, i.e.

$$\mathbf{w}_0 = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \mathbb{E}[V(y\mathbf{x}^T\mathbf{w}/\sqrt{p})]. \tag{3}$$

Applying a Taylor expansion of $\mathcal{L}(\hat{\mathbf{w}})$ around $\mathbf{w}_0$, (2) can be approximated by

$$\nabla\mathcal{L}(\mathbf{w}_0) + \nabla^2\mathcal{L}(\mathbf{w}_0)(\hat{\mathbf{w}} - \mathbf{w}_0) + \lambda\hat{\boldsymbol{\kappa}} \approx 0.$$

Approximate $\nabla^2\mathcal{L}(\mathbf{w}_0) \approx \hat{\boldsymbol{\Sigma}}$, we obtain

$$\hat{\mathbf{w}} - \hat{\boldsymbol{\Sigma}}^{-1}\nabla\mathcal{L}(\hat{\mathbf{w}}) \approx \mathbf{w}_0 - \hat{\boldsymbol{\Sigma}}^{-1}\nabla\mathcal{L}(\mathbf{w}_0). \tag{4}$$

From (3), we have $\nabla\mathcal{L}(\mathbf{w}_0) \to \mathbb{E}[\nabla V(y\mathbf{x}^T\mathbf{w}_0/\sqrt{p})] = 0$. Thus, $\hat{\mathbf{w}} - \hat{\boldsymbol{\Sigma}}^{-1}\nabla\mathcal{L}(\hat{\mathbf{w}})$ can reasonably be considered as a consistent estimator for the population parameter $\mathbf{w}_0$. Consequently, adding an extra term $\hat{\boldsymbol{\Sigma}}^{-1}\nabla\mathcal{L}(\hat{\mathbf{w}})$ to the original estimator $\hat{\mathbf{w}}$ corrects the shrinkage effect of L_1-regularization.

To derive an exact formulation for the asymptotic behavior of the classification method (1), we employ the replica method, a technique from statistical physics used to analyze the properties of the system in the limit $n, p \to \infty$ with a fixed ratio $\alpha = n/p$. This type of asymptotic analysis is highly valuable as it provides not only bounds but also sharp predictions for the limiting joint distribution of the estimated and true parameters in a model. From this joint distribution, various computations can be performed to obtain precise predictions for key quantities such as the mean squared error and the misclassification error rate. The replica method is a powerful tool for the theoretical analysis of high-dimensional problems. Although heuristic in nature and involving certain mathematical subtleties (Mézard and Montanari, 2009), it has been successfully applied to a wide range of challenging problems in various domains (Parisi, 1979; Talagrand, 2003).

Assume that each training data point $(\mathbf{x}_i, y_i)$, for $i = 1, \cdots, n$, is an independent random vector drawn from a joint distribution function $p(\mathbf{x}, y)$. Conditional on $y = +1, -1$, the feature vector $\mathbf{x}$ follows multivariate normal distributions $p(\mathbf{x}|y = +1)$, $p(\mathbf{x}|y = -1)$ with mean $\boldsymbol{\mu}_+, \boldsymbol{\mu}_-$ and covariance $\boldsymbol{\Sigma}_+, \boldsymbol{\Sigma}_-$, respectively. Without loss of generality, we assume that $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \boldsymbol{\mu}$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ and that $\boldsymbol{\Sigma}_\pm$ are $p \times p$ positive definite matrices. In this setting, the data are generated from a mixture of two multivariate Gaussian distributions with different means and covariance matrices. The strength of the signal can be characterized by $\mu = \|\boldsymbol{\mu}\|$. Here, $\boldsymbol{\Sigma}_\pm$ can be any finite positive definite matrix, making our model quite general. For simplicity, we assume $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_- = \boldsymbol{\Sigma}$. For vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, we define $\|\mathbf{u}\|^2_{\boldsymbol{\Sigma}} = \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}$ and $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=1}^p u_j v_j$. The following result characterizes the exact limiting distribution of the solution $\hat{\mathbf{w}}$ to (1) in this asymptotic setting.

Result 1. Define two random vectors

$$\begin{aligned} \bar{\mathbf{w}} &= \hat{\mathbf{w}} - \frac{1}{\sqrt{p}\zeta} \sum_{i=1}^n y_i V' \left(\frac{y_i \mathbf{x}_i^T \hat{\mathbf{w}}}{\sqrt{p}} \right) \boldsymbol{\Sigma}^{-1} \mathbf{x}_i, & (5) \\ \tilde{\mathbf{w}} &= \sqrt{p} R_0 \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}} / \zeta + \tau \boldsymbol{\Sigma}^{-1/2} \mathbf{z}, & (6) \end{aligned}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$, $V'(\cdot)$ is the derivative of the convex loss function with respect to its argument, $\hat{\mathbf{w}}$ is the minimizer of (1), $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$, $\tau = \sqrt{\zeta_0}/\zeta$, and ζ_0, ζ, R_0 can be solved from the following set of nonlinear equations:

$$\begin{aligned} \zeta_0 &= \frac{\alpha}{q^2} \mathbb{E}(\hat{u}_\epsilon - R\mu - \sqrt{q_0}\epsilon)^2, \\ \zeta &= -\frac{\alpha}{q\sqrt{q_0}} \mathbb{E}[(\hat{u}_\epsilon - R\mu - \sqrt{q_0}\epsilon)\epsilon], \\ R_0 &= \frac{\alpha\mu}{q} \mathbb{E}(\hat{u}_\epsilon - R\mu - \sqrt{q_0}\epsilon), \qquad (7) \\ q_0 &= \frac{1}{p} \mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \boldsymbol{\Sigma} \hat{\mathbf{w}}_{\mathbf{z}}), \\ q &= \frac{1}{p\sqrt{\zeta_0}} \mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \boldsymbol{\Sigma}^{1/2} \mathbf{z}), \\ R &= \frac{1}{\sqrt{p}} \mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \hat{\boldsymbol{\mu}}), \end{aligned}$$

where the first three expectations are with respect to $\epsilon \sim N(0, 1)$, the last three expectations are with respect to $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$, and

$$\begin{aligned} \hat{u}_\epsilon &= \arg\min_{u \in \mathbb{R}} \left[V(u) + \frac{(u - R\mu - \sqrt{q_0}\epsilon)^2}{2q} \right], & (8) \\ \hat{\mathbf{w}}_{\mathbf{z}} &= \arg\min_{\mathbf{w} \in \mathbb{R}^p} \left[\frac{\zeta}{2} \|\mathbf{w}\|^2_{\boldsymbol{\Sigma}} - \langle \sqrt{\zeta}_0 \boldsymbol{\Sigma}^{1/2} \mathbf{z} + \sqrt{p} R_0 \hat{\boldsymbol{\mu}}, \mathbf{w} \rangle + \sum_{j=1}^p J_\lambda(w_j) \right]. & (9) \end{aligned}$$

Then $\bar{\mathbf{w}}$ and $\tilde{\mathbf{w}}$ follow the same distribution asymptotically in the limit of $n, p \to \infty$ with fixed $\alpha = n/p$.

The derivation of **Result 1** is provided in Section A.1 using the replica method. The six parameters $\zeta_0, \zeta, R_0, q_0, q, R$, defined in (7), represent order parameters that characterize the degree of similarity or correlation between different "replicas" of the system. For example, q_0 relates to the length of the estimation vector $\hat{\mathbf{w}}$, R describes the overlap between $\hat{\mathbf{w}}$ and the signal $\boldsymbol{\mu}$, and q corresponds to the overlap between $\hat{\mathbf{w}}$ and the random noise $\mathbf{z}$.

The two random vectors $\bar{\mathbf{w}}$ and $\tilde{\mathbf{w}}$ defined in (5) and (6) correspond to the left-hand side and right-hand side of (4), respectively. Particularly, $\bar{\mathbf{w}}$ can be interpreted as the de-biased estimator of the penalized convex classification method (1), providing a foundation for statistical inference. The analogous de-biased estimator for LASSO was constructed in Javanmard and Montanari (2014b) using the replica method in the context of the standard linear regression with general Gaussian random design. This was generalized to deterministic design in Javanmard and Montanari (2014a). An extension to the rotation-invariant design matrix was made in Takahashi and Kabashima (2018); Na et al. (2023), where the replica method was combined with the TAP equations, i.e., the fixed point equation of the approximate message-passing iteration.

The random vector, $\tilde{\mathbf{w}}$ defined in (6) consists of two terms: the first corresponds to the true signal $\boldsymbol{\mu}$, while the second represents noise arising from high-dimensional effects. Notably, as $\alpha \to \infty$, we have $\tau \to 0$ according to (7), causing the signal term in (6) to dominate over the random noise term, thereby recovering the traditional large-sample result. Conversely, as $\alpha \to 0$, we have $\tau \to \infty$, meaning that $\tilde{\mathbf{w}}$ is entirely determined by random noise. This aligns with the high-dimensional, low-sample-size results described in Hall et al. (2005).

It is important to note that this paper focuses on theoretical analysis. In practice, $\boldsymbol{\Sigma}$ is typically unknown and must be estimated. Let the prediction accuracy of the classifier $\hat{\mathbf{w}}$ be defined as $Pres_{\hat{\mathbf{w}}} = \mathbb{E}[\mathrm{I}(\mathrm{sign}(\mathbf{x}^T\hat{\mathbf{w}}) = y)]$, where $\mathrm{I}(\cdot)$ is the indicator function, and the expectation is taken with respect to a fresh sample $(\mathbf{x}, y)$ that is independent of the training data. The generalization error of $\hat{\mathbf{w}}$ is then given by $Err_{\hat{\mathbf{w}}} = 1 - Pres_{\hat{\mathbf{w}}}$. Both $Pres_{\hat{\mathbf{w}}}$ and $Err_{\hat{\mathbf{w}}}$ serve as measures of the predictive performance of a model, and we will present $Pres_{\hat{\mathbf{w}}}$ in our numerical studies. The distributional limit of the estimator $\hat{\mathbf{w}}$, as derived in **Result 1**, facilitates the computation of $Pres_{\hat{\mathbf{w}}}$, as detailed in the following result.

Result 2. Under the conditions of **Result 1**, the limiting distribution of $\hat{\mathbf{w}}$ leads to the asymptotic precision

$$Pres_{\hat{\mathbf{w}}} = P(y\mathbf{x}^T\hat{\mathbf{w}} \geq 0) \to \Phi\left(\frac{R\mu}{\sqrt{q_0}}\right), \tag{10}$$

where the parameters R, q_0 are determined from the set of nonlinear equations (7) and Φ is the CDF of the standard normal distribution.

The derivation of **Result 2** is provided in Section A.2. From (10), it is evident that precision increases with both the magnitude of the signal, $\mu = \|\boldsymbol{\mu}\|$, and the overlap between $\hat{\mathbf{w}}$ and $\boldsymbol{\mu}$, as supported by the theoretical arguments in Engel and Van den Broeck (2001); Hertz et al. (1991) and our numerical studies in Figures 1, 2, and 9. The precision does not explicitly depend on α; however, since the order parameters R and q_0 depend on α, it also increases with α, as demonstrated in our numerical studies in Section 3.

Toward variable selection for the penalized convex classification method (1), we consider the population loss $\mathbb{E}[V(y\mathbf{x}^T\mathbf{w}/\sqrt{p})]$. This definition $\mathbf{w}_0$ in (3) has a strong connection to the Bayes rule, which is theoretically optimal if the underlying distribution is known. The Bayes rule is given by $\mathrm{sign}(\mathbf{x}^T\mathbf{w}_{\mathrm{Bayes}})$ with $\mathbf{w}_{\mathrm{Bayes}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \mathbb{E}[\mathrm{I}\{\mathrm{sign}(\mathbf{x}^T\mathbf{w}) \neq y\}]$. The Bayes rule is unattainable if we assume that we have no knowledge of the high-dimensional conditional density $p(\mathbf{x}|y)$. Note that $\mathbf{w}_{\mathrm{Bayes}}$ and $\mathbf{w}_0$ are equivalent to each other in the important special case of Fisher linear discriminant analysis. In more general settings, $\mathbf{w}_{\mathrm{Bayes}}$ and $\mathbf{w}_0$ may not be the same. The minimizer from the population loss could be a reasonable target for inference in many applications (Zhang et al., 2016; Lu et al., 2016).

If the conditional distribution $p(\mathbf{x}|y)$ is multivariate normal, $\mathbf{w}_0$ is proportional to $\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$. In general settings, $\mathbf{w}_0$ is approximately proportional to $\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$ which is in the same direction as the first term on the right-hand side of (6). The basic intuition of this observation is that, according to equations (5) and (6), the marginal distribution of $\bar{w}_j - cw_{0,j}$ is expected to be asymptotically $N(0, \tau^2(\mathbf{\Sigma}^{-1})_{jj})$ for $j = 1, \cdots, p$, where c is a constant. Assume that the true parameter $\mathbf{w}_0$ is sparse, then variable selection can be achieved using a series of hypothesis testing procedures. Here, we consider testing the null hypothesis $H_{0j} : w_{0,j} = 0$ for $j = 1, \cdots, p$ and assigning the p-values for these tests. Rejecting H_{0j} is equivalent to stating that $w_{0j} \neq 0$. The decision rule for j-th hypothesis can be based on the p-value $p_j = 2\left[1 - \Phi\left(\left|\bar{w}_j \big/ \left\{\tau\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{jj}}\right\}\right|\right)\right]$, where $\Phi(x)$ is the standard Gaussian CDF. The confidence interval for the j-th coefficient can be estimated as

$$\left[\bar{w}_j - \Phi(1-\delta/2)\tau\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{jj}},\ \bar{w}_j + \Phi(1+\delta/2)\tau\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{jj}}\right], \tag{11}$$

where δ is the prespecified significance level.

3 Simulation Results

This section presents simulation studies for L_1-regularized logistic regression (PLR) to validate the findings discussed in the previous sections. Unlike the existing literature, which usually assumes low correlation among predictors, we generate synthetic data including both low correlation and high correlation among predictors. The L_1-regularized logistic regression model is fitted using the R package `glmnet`.

3.1 Precision rates

This subsection compares the empirical precision rates computed by Monte Carlo and those computed from the asymptotic result based on **Result 2**. We set $p = 1000$ and $n = p\alpha$, where $\alpha = 0.5$. The response variable y takes values in $\{-1, +1\}$ with equal probabilities for each. The predictors $\mathbf{x} \mid y$ are simulated from $N_p(y\boldsymbol{\mu}, \mathbf{\Sigma})$ for various configurations of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. We choose $\boldsymbol{\mu} = a\mathbf{\Sigma}\mathbf{w}_0$, where $\mathbf{w}_0$ is a sparse vector with components equal to 1 or 0 with sparsity level $\epsilon = 0.01, 0.05, 0.1$, and a is selected to ensure that the length of $\boldsymbol{\mu}$ is 2. Set $\mathbf{\Sigma} = \sigma^2\mathbf{C}$, where $\sigma^2 = 2.0$. We consider four different correlation structures for $\mathbf{C}$: (1) IID, $\mathbf{C} = \mathbf{I}_p$, the identity matrix; (2) block diagonal, $\mathbf{C}$ is a block diagonal matrix with blocks $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$; (3) AR1, $\mathbf{C} = (0.8^{|i-j|})$ representing an autoregressive model of order 1; (4) banded, $\mathbf{C}$ is a banded correlation matrix where the diagonal elements are $C_{ii} = 1$ and the off-diagonal elements are $C_{ij} = 0.4$ if $i \neq j$ and $|i - j| \leq 2$ and $C_{ij} = 0$ otherwise.

The precision rate of a classifier is the probability of correctly classifying a new observation. Let $\hat{\mathbf{w}}$ be the fitted coefficient from a simulated dataset. The corresponding precision rate is defined as $P(\tilde{y}\tilde{\mathbf{x}}^T\hat{\mathbf{w}} > 0)$ for a new observation $(\tilde{y}, \tilde{\mathbf{x}})$. In practice, we randomly simulate $(\tilde{y}, \tilde{\mathbf{x}})$ using the same mechanism as the original training data of the same size and approximate the accurate probability of classification by a proportion of samples. As $n, p \to \infty$, the asymptotic precision rate is given by $\Phi(R\|\boldsymbol{\mu}\|/\sqrt{q_0})$ in (10), where $\Phi(\cdot)$ is the CDF of the standard normal distribution, R and q_0 are saddle point parameters evaluated according to the algorithm described in Appendix A.3.

Figure 1 compares the theoretical precision rates with the empirical results in different correlation structures and sparsity levels. For each setting, the empirical precision rates are computed by averaging over 500 replicates with different random seeds. The model is fitted using various tuning parameters, $\log\lambda = -3, -2.5, \cdots, 0.5$. The three lines represent the precision rates computed from the asymptotic results at different sparsity levels, while the error bars denote the confidence intervals of the mean precision rates calculated from Monte Carlo simulations. The figure illustrates a close alignment between the empirical findings and the theoretical predictions. Precision increases with sparsity for a fixed covariance structure. This is because sparsity generally leads to a reduction in the effective dimension, meaning that even if a system has a high-dimensional representation, the number of truly independent or influential degrees of freedom is much smaller. For the same sparsity level, precision decreases as the covariance structure becomes more complex. Specifically, the IID case yields the highest precision, while the AR(1) case results in the lowest precision. The precision rates for the block diagonal and banded cases fall in between. Plots for other configurations show similar consistency but are not included here due to space limitations.

Figure 1 shows that the precision rate depends on the choice of λ, as confirmed by the implicit dependence of the asymptotic precision in (10) on λ. In practice, λ is selected to maximize the precision rate. Figure 1 suggests that the optimal λ is approximately $e^{-0.5}$ or e^0, depending on the sparsity level and the correlation structure. To illustrate the dependence of the precision rate at the optimal λ on α, we consider the IID correlation structure and simulate data for $\alpha = 0.3, \ldots, 1.5$ following the previously described mechanism. Figure 2 presents the asymptotic precision at the optimal λ as a function of α, where the three lines correspond to different sparsity levels $\epsilon = 0.01, 0.05, 0.1$. The error bars are the 95% confidence intervals of the precision rate calculated by 500 replicates. The precision rate increases as the true model becomes sparser and α increases, both of which correspond to scenarios where n becomes much larger relative to the effective dimension, that is, the number of relevant variables.

3.2 Variable selection

This subsection evaluates the performance of variable selection through confidence intervals constructed using the formulas proposed in Section 2. Synthetic data are generated following the same procedure as in Section 3.1. For a typical dataset, the histogram of the estimated coefficients $\hat{\mathbf{w}}$ is displayed in the left plot of Figure 3, which corresponds to the AR1 case with $\epsilon = 0.1$ and $\log\lambda = -2$. The spike indicates that most components of $\hat{\mathbf{w}}$ are zero, which is an effect of the LASSO penalty. The right plot of Figure 3 shows the histogram of the de-biased estimate $\bar{\mathbf{w}}$. The histogram shows a slightly right-skewed distribution resulting from two closely positioned clusters. Note that the mean of $\bar{\mathbf{w}}$ is $\sqrt{p}R_0\boldsymbol{\Sigma}^{-1}\hat{\mu}/\zeta$, which corresponds to the first term in (6) and takes two distinct values. The first cluster corresponds to zero coefficients, and the second cluster corresponds to the nonzero coefficients. Both clusters follow normal distributions, as suggested in **Result 1**.

We further construct individual 95% confidence intervals for components of $\bar{\mathbf{w}}$ following the formula (11) derived in Section 2. Ideally, we expected 95% of these confidence intervals to include the corresponding true parameters if we choose the significance level $\delta = 0.05$. Consider four correlation structures as explained in Section 3.1. Each box plot in Figure 4 summarizes the empirical confidence levels based on 500 replicates for a scenario with a different sparsity level $\epsilon = 0.01, 0.05, 0.1$ at different values of tuning parameter $\log \lambda = -3, -2.5, \ldots, 0.5$. In these graphs, the horizontal line represents the nominal confidence level of 95%. These findings confirm the validity of the formulas for constructing confidence intervals in various choices of λ. Although plots for other configurations demonstrate similar consistency, they are not displayed here to save space; nevertheless, they are available upon request.

Furthermore, we assess the power of the procedure as the probability of correctly identifying significant variables when their corresponding true values are nonzero. The theoretical power is derived using **Result 1**, while the empirical power is estimated as the proportion of non-zero components whose 95% confidence intervals do not include zero. Figure 5 displays the theoretical powers as lines and 95% confidence intervals of the mean empirical powers based on 500 replicates as error bars. The close alignment between the theoretical and empirical powers indicates the validity of the theory. The results show that the power increases with a larger value of λ, which demonstrates a pattern similar to the precision rate in Figure 1. Furthermore, the sparsity level of $\mathbf{w}_0$ also influences the power; specifically, the power decreases as $\mathbf{w}_0$ becomes denser.

The performance of a variable selection procedure can be visualized using a ROC curve, which shows the trade-off between true positive and false positive rates. Figure 6 shows the ROC curves for the IID and AR1 correlation structures for different choices of λ and sparsity levels. The performance of variable selection improves as the true model becomes sparser and the correlation among predictors decreases. Different choices of λ may lead to different performance of variable selection. The area under the curve (AUC) serves as a numeric measure of the performance of variable selection. Figure 7 presents the 95% confidence intervals for AUC at different sparsity levels $\epsilon = 0.01, 0.05, 0.1$ for different correlation structures. The performance of variable selection improves as the true model becomes sparser, since LASSO excels in sparse settings. The relationship between AUC and λ is similar to that between power and λ in Figure 5.

An interesting question is whether the performance of variable selection is related to the variance of the debiased estimator, or equivalently, the width of confidence intervals. The formula (11) suggests that the width of a confidence interval is $2\Phi(1-\delta/2)\tau\sqrt{(\mathbf{\Sigma}^{-1})_{jj}}$, which depends on λ through τ. Therefore, we examine the relationship between τ and λ for four correlation structures in Figure 8. The three lines in the figure represent different sparsity levels: $\epsilon = 0.01, 0.05, 0.1$. Figure 8 shows that the dependence of τ on λ follows similar patterns across different sparsity levels and covariance structures. In particular, under the AR1 setting, the three curves are almost identical across different sparsity levels. The value of τ, or the width of confidence intervals, is minimized around $\log \lambda = 0$ or 0.5, depending on the sparsity levels and correlation structures. Notice that this optimal λ is larger than the optimal λ that maximizes the precision rate in Section 3.1. It is consistent with an observation in regression that the optimal λ for variable selection is usually larger than the optimal λ for prediction (Friedman et al., 2010). Figure 9 shows the precision rate at the optimal λ that minimizes the width of the confidence intervals as a function of α. It exhibits a similar pattern to Figure 2, which presents the precision rate at the optimal λ that maximizes the precision rate.

4 Summary

This paper focuses on learning Gaussian mixture data using L_1-regularized classification methods. We study the asymptotic behavior of the estimators in the framework in which $p, n \to \infty$ while

$n/p \to \alpha$ is fixed. We first derive the limiting distribution of the regularized convex classifiers with an arbitrarily covariance structure. Then we obtain the generalization error of the classifiers and further propose a de-biased estimator for classification, which allows us to perform variable selection through an appropriate hypothesis testing procedure. Using L_1-regularized logistic regression as an example, we conduct extensive computational experiments to confirm that our theoretical predictions are consistent with simulation results. Our next step is to implement the current framework for another commonly used classification method: the support vector machine.

Our analysis is based on the replica method, which is not yet fully rigorous. Another future direction of our research is to provide a rigorous justification for the results derived in this paper. So far, the rigorous work in this area has mainly focused on i.i.d. randomness. The rigorous results for general covariance structure are more challenging, and one possible solution is to use Gordon's Gaussian minmax inequalities (Gordon, 1985).

Acknowledgments

The authors thank the editor, associate editor, and three referees for their helpful comments and suggestions, which have significantly improved the presentation.

Appendix

A.1 Derivation of Result 1

This appendix outlines the replica calculation leading to **Result 1**. Our derivation follows from a Boltzmann formulation of the optimization problem in (1), followed by a replica analysis inspired by the statistical physics toolbox of disordered systems. We present only the main steps. For a general introduction to the replica method and its motivation, we refer to Mézard et al. (1987); Mézard and Montanari (2009).

Our main strategy for deriving that the two vectors $\bar{\mathbf{w}}$ and $\tilde{\mathbf{w}}$, defined in (5) and (6), have the same asymptotic distribution is to show that

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \tilde{g}(\bar{w}_j) = \lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \tilde{g}(\tilde{w}_j) \tag{A-1}$$

for a complete set of functions $\tilde{g} : \mathbb{R} \to \mathbb{R}$. To achieve this, we first introduce in (A-5) an appropriately defined free energy function $\mathcal{F}(s)$ with an external field s. We then derive both sides of (A-1) by applying the large deviation analysis and replica analysis, respectively, to $\left.\frac{d\mathcal{F}(s)}{ds}\right|_{s\to 0+}$.

Denote $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T$, $\mathbf{y} = (y_1, \cdots, y_n)^T$. Recall that we are considering the regularized classification of the form

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{p}}\right) + \sum_{j=1}^{p} J_\lambda(w_j) \right\}. \tag{A-2}$$

Let us introduce the conjugate vector $\mathbf{v}$ corresponding to $\mathbf{w}$ and define their Boltzmann distribution as

$$P(\mathbf{w},\mathbf{v};\beta,s) = \frac{1}{Z_p(\beta,s)}\exp\left\{-\beta\left[\sum_{i=1}^{n}V\left(\frac{y_i\mathbf{x}_i^T(\mathbf{w}+s\tilde{d}\mathbf{\Sigma}^{-1}\mathbf{v})}{\sqrt{p}}\right)\right.\right.$$
$$\left.\left.+\sum_{j=1}^{p}\{J_\lambda(w_j)-s(g(v_j)-v_jw_j)\}\right]\right\}, \qquad \text{(A-3)}$$

where $\tilde{d}\in\mathbb{R}$ will be defined below, $\beta{>}0$ is a 'temperature' parameter, $s{>}0$, and $g:\mathbb{R}\to\mathbb{R}$ is a continuous, strictly convex function. The conjugate vector $\mathbf{v}$ is introduced to facilitate the Lagrangian-dual optimization, allowing us to obtain $\tilde{g}(u)\equiv\max_{x\in\mathbb{R}}[ux-g(x)]$. The partition function $Z_p(\beta,s)$ in (A-3) is defined as

$$Z_p(\beta,s) = \int\exp\left\{-\beta\left[\sum_{i=1}^{n}V\left(\frac{y_i\mathbf{x}_i^T(\mathbf{w}+s\tilde{d}\mathbf{\Sigma}^{-1}\mathbf{v})}{\sqrt{p}}\right)\right.\right.$$
$$\left.\left.+\sum_{j=1}^{p}\{J_\lambda(w_j)-s(g(v_j)-v_jw_j)\}\right]\right\}d\mathbf{w}d\mathbf{v}. \qquad \text{(A-4)}$$

In the low-temperature and small-s limit, i.e., as $\beta\to\infty$ and $s\to0+$, $Z_p(\beta,s)$ exists and is dominated by the values of $\hat{\mathbf{w}}$, the solution of (A-2).

Within the replica method, it is assumed that the limits $p\to\infty$, $\beta\to\infty$ exist almost surely for the quantity $(p\beta)^{-1}\log Z_p(\beta,s)$, and that the order of the limits can be exchanged. We therefore define the free energy

$$\mathcal{F}(s) = -\lim_{\beta\to\infty}\lim_{p\to\infty}\frac{1}{p\beta}\log Z_p(\beta,s) = -\lim_{p\to\infty}\lim_{\beta\to\infty}\frac{1}{p\beta}\log Z_p(\beta,s). \qquad \text{(A-5)}$$

In other words, $\mathcal{F}(s)$ is the exponential growth rate of $Z_p(\beta,s)$.

We assume that the derivative of $\mathcal{F}(s)$ as $s\to0+$ can be obtained by differentiating inside the limit. This condition holds, for example, if the cost function is strongly convex as $s\to0+$. We get

$$\left.\frac{d\mathcal{F}}{ds}\right|_{s\to0+} = \lim_{\beta,p\to\infty,s\to0+}\mathbb{E}_{P(\mathbf{w},\mathbf{v};\beta,s)}\frac{1}{p}\left\{\sum_{j=1}^{p}\left[g(v_j)-v_j\left\{w_j-\sum_{i=1}^{n}V'\left(\frac{y_i\mathbf{x}_i^T\mathbf{w}}{\sqrt{p}}\right)\frac{\tilde{d}y_i(\mathbf{x}_i^T\mathbf{\Sigma}^{-1})_j}{\sqrt{p}}\right\}\right]\right\}.$$

At $\beta\to\infty$, $s\to0+$, $\mathbf{w}$ concentrates at $\hat{\mathbf{w}}$, the integral with respect to $\mathbf{v}$ concentrates at

$$\min_{v_j}\left[g(v_j)-v_j\left\{\hat{w}_j-\sum_{i=1}^{n}V'\left(\frac{y_i\mathbf{x}_i^T\hat{\mathbf{w}}}{\sqrt{p}}\right)\frac{\tilde{d}y_i(\mathbf{x}_i^T\mathbf{\Sigma}^{-1})_j}{\sqrt{p}}\right\}\right],$$

for $j=1,\cdots,p$, which leads to the left hand side of (A-1)

$$\left.\frac{d\mathcal{F}}{ds}\right|_{s\to0+} = -\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\tilde{g}(\bar{w}_j), \qquad \text{(A-6)}$$

where

$$\bar{w}_j = \hat{w}_j-\sum_{i=1}^{n}V'\left(\frac{y_i\mathbf{x}_i^T\hat{\mathbf{w}}}{\sqrt{p}}\right)\frac{\tilde{d}y_i(\mathbf{x}_i^T\mathbf{\Sigma}^{-1})_j}{\sqrt{p}}. \qquad \text{(A-7)}$$

Hence, by computing $\frac{d\mathcal{F}}{ds}\big|_{s\to 0+}$ for a complete set of functions $\tilde{g}$, we get access to the corresponding limit quantities (A-6), and hence, via standard weak convergence arguments, to the empirical distribution of $\bar{w}_j$.

Next, we apply replica analysis to $\frac{d\mathcal{F}}{ds}\big|_{s\to 0+}$ to obtain the right-hand side of (A-1). We assume that $p^{-1}\log Z_p(\beta, s)$ is tightly concentrated around its expectation, allowing $\mathcal{F}(s)$ to be evaluated by computing

$$\mathcal{F}(s) = -\lim_{p\to\infty}\lim_{\beta\to\infty}\frac{1}{p\beta}\mathbb{E}\log Z_p(\beta, s), \tag{A-8}$$

where the expectation is taken with respect to the distribution of the training data $\mathbf{X}$ and $\mathbf{y}$.

In order to evaluate the integration of a log function, we make use of the replica method based on the identity

$$\log Z = \lim_{k\to 0}\frac{\partial Z^k}{\partial k} = \lim_{k\to 0}\frac{\partial}{\partial k}\log Z^k, \tag{A-9}$$

and rewrite (A-8) as

$$\mathcal{F}(s) = -\lim_{\beta\to\infty}\lim_{p\to\infty}\frac{1}{p\beta}\lim_{k\to 0}\frac{\partial}{\partial k}\Xi_k(\beta), \tag{A-10}$$

where

$$\Xi_k(\beta) = \langle\{Z_\beta(\beta, s)\}^k\rangle_{\mathbf{X},\mathbf{y}} = \int\{Z_\beta(\beta, s)\}^k\prod_{i=1}^n P(\mathbf{x}_i, y_i)d\mathbf{x}_i dy_i. \tag{A-11}$$

Equation (A-10) can be derived by using the fact that $\lim_{k\to 0}\Xi_k(\beta) = 1$ and exchanging the order of the averaging and the differentiation with respect to k. In the replica method, we will first evaluate $\Xi_k(\beta)$ for the integer k and then apply it to the real k and take the limit of $k \to 0$.

For the integer k, to represent $\{Z_\beta(\mathbf{X},\mathbf{y})\}^k$ in the integrand of (A-11), we use the identity.

$$\left(\int f(x)\nu(dx)\right)^k = \int f(x_1)\cdots f(x_k)\nu(dx_1)\cdots\nu(dx_k),$$

where $\nu(dx)$ denotes the measure over $x \in \mathbb{R}$. We obtain

$$\begin{aligned}\{Z_\beta(\beta, s)\}^k &= \prod_{a=1}^k\left(\int\exp\left\{-\beta\left[\sum_{i=1}^n V\left(\frac{y_i\mathbf{x}_i^T(\mathbf{w}^a + s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}^a)}{\sqrt{p}}\right)\right.\right.\right.\\ &\qquad\left.\left.\left. + \sum_{j=1}^p\left\{J_\lambda(w_j^a) + s(g(v_j^a) - v_j^a w_j^a)\right\}\right]\right\}d\mathbf{w}^a d\mathbf{v}^a\right),\end{aligned}$$

where we have introduced replicated parameters

$$\mathbf{w}^a \equiv [w_1^a, \cdots, w_p^a]^T \quad \text{and} \quad \mathbf{v}^a \equiv [v_1^a, \cdots, v_p^a]^T, \text{ for } a = 1, \cdots, k.$$

Exchanging the order of the two limits $p \to \infty$ and $k \to 0$ in (A-10), we have

$$\mathcal{F}(s) = -\lim_{\beta\to\infty}\frac{1}{\beta}\lim_{k\to 0}\frac{\partial}{\partial k}\left(\lim_{p\to\infty}\frac{1}{p}\Xi_k(\beta)\right). \tag{A-12}$$

Define the measure $\nu(d\mathbf{w})$ over $\mathbf{w} \in \mathbb{R}^p$ as follows

$$\nu(d\mathbf{w}) \;=\; \left[\int \exp\left\{-\beta\left[J_\lambda(\mathbf{w} - s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}) + s\sum_{j=1}^{p}\Big\{g(v_j) - v_j w_j + s\tilde{d} v_j(\boldsymbol{\Sigma}^{-1}\mathbf{v})_j\Big\}\right]\right\} d\mathbf{v}\right] d\mathbf{w}.$$

Similarly, define the measure $\nu_+(d\mathbf{x})$ and $\nu_-(d\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^p$ as

$$\nu_+(d\mathbf{x}) = P(\mathbf{x}|y = +1)d\mathbf{x} \quad \text{and} \quad \nu_-(d\mathbf{x}) = P(\mathbf{x}|y = -1)d\mathbf{x}.$$

In order to carry out the calculation of $\Xi_k(\beta)$, we let $\nu^k(d\mathbf{w}) \equiv \nu(d\mathbf{w}^1) \times \cdots \times \nu(d\mathbf{w}^k)$ be a measure over $(\mathbb{R}^p)^k$, with $\mathbf{w}^1, \cdots, \mathbf{w}^k \in \mathbb{R}^p$. Analogously $\nu^n(d\mathbf{x}) \equiv \nu(d\mathbf{x}_1) \times \cdots \times \nu(d\mathbf{x}_n)$ with $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$, $\nu^n(dy) \equiv \nu(dy_1) \times \cdots \times \nu(dy_n)$ with $y_1, \cdots, y_n \in \{-1, 1\}$. With these notations and the change of variable $\mathbf{w}^a + s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}^a \to \mathbf{w}^a$, we can rewrite (A-11) as

$$\begin{aligned}
\Xi_k(\beta) &= \int \exp\left\{-\beta\sum_{i=1}^{n}\sum_{a=1}^{k} V\left(\frac{y_i\mathbf{x}_i^T\mathbf{w}^a}{\sqrt{p}}\right)\right\}\nu^k(d\mathbf{w})\nu^n(d\mathbf{y})\nu^n(d\mathbf{x}) \\
&= \int \left[\int \exp\left\{-\beta\sum_{a=1}^{k} V\left(\frac{\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\right\}\nu_+(d\mathbf{x})\right]^{n_+} \times \\
&\qquad \left[\int \exp\left\{-\beta\sum_{a=1}^{k} V\left(\frac{-\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\right\}\nu_-(d\mathbf{x})\right]^{n_-} \nu^k(d\mathbf{w}) \\
&= \int \exp\{p(\alpha_+ \log I_+ + \alpha_- \log I_-)\}\nu^k(d\mathbf{w}), \qquad \text{(A-13)}
\end{aligned}$$

where $\alpha_\pm = n_\pm/p$ and

$$I_\pm \;=\; \int \exp\left\{-\beta\sum_{a=1}^{k} V\left(\frac{\pm\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\right\}\nu_\pm(d\mathbf{x}). \qquad \text{(A-14)}$$

Notice that above we used the fact that the integral over $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in (\mathbb{R}^p)^n$ factors into n_+ integrals over $(\mathbb{R})^p$ with measure $\nu_+(d\mathbf{x})$ and n_- integrals over $(\mathbb{R})^p$ with measure $\nu_-(d\mathbf{x})$. Next, we introduce the integration variables $du^a, d\tilde{u}^a$ for $1 \le a \le k$. Letting $\nu^k(du) = du^1 \cdots du^k$ and $\nu^k(d\tilde{u}) = d\tilde{u}^1 \cdots d\tilde{u}^k$, we obtain

$$\begin{aligned}
I_\pm &= \int \exp\left\{-\beta\sum_{a=1}^{k} V(u^a)\right\}\prod_{a=1}^{k}\delta\left(u^a \mp \frac{\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\nu^k(du)\nu_\pm(d\mathbf{x}) \\
&= \int \exp\left\{-\beta\sum_{a=1}^{k} V(u^a)\right\}\exp\left\{\sum_{a=1}^{k} i\sqrt{p}\left(u^a \mp \frac{\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\tilde{u}^a\right\}\nu^k(du)\nu^k(d\tilde{u})\nu_\pm(d\mathbf{x}) \\
&= \int \exp\left\{-\beta\sum_{a=1}^{k} V(u^a) + i\sqrt{p}\sum_{a=1}^{k}\left(u^a \mp \frac{\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\tilde{u}^a\right\}\nu_\pm(d\mathbf{x})\nu^k(du)\nu^k(d\tilde{u}) \\
&= \int \exp\Bigg\{-\beta\sum_{a=1}^{k} V(u^a) + i\sqrt{p}\sum_{a=1}^{k} u^a\tilde{u}^a - \frac{1}{2}\sum_{ab}(\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b\tilde{u}^a\tilde{u}^b \\
&\qquad\qquad -i\sum_{a=1}^{k}(\mathbf{w}^a)^T\boldsymbol{\mu}\tilde{u}^a\Bigg\}\nu^k(du)\nu^k(d\tilde{u}). \qquad \text{(A-15)}
\end{aligned}$$

Note that conditional on $y = \pm 1$, $\mathbf{x}$ follows multivariate distributions with mean $\pm\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In deriving (A-15), we have used the fact that the low-dimensional marginals of $\mathbf{x}$ can be approximated by a Gaussian distribution based on the multivariate central limit theorem.

From (A-15) we have $I_+ = I_- \equiv I$. Thus, applying (A-15) to (A-13) we obtain

$$\Xi_k(\beta) \quad = \quad \int \exp\{p\alpha \log I\}\nu^k(d\mathbf{w}).$$

Now we introduce integration variables $Q_{ab}, \tilde{Q}_{ab}$ and R^a, R_0^a associated with $(\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b/p$ and $(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}/\sqrt{p}$ respectively for $1 \le a, b \le k$ where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\mu$. Denote $\mathbf{Q} \equiv (Q_{ab})_{1\le a,b\le k}$, $\tilde{\mathbf{Q}} \equiv (\tilde{Q}_{ab})_{1\le a,b\le k}$, $\mathbf{R} \equiv (R^a)_{1\le a\le k}$, and $\mathbf{R}_0 \equiv (R_0^a)_{1\le a\le k}$. Note that constant factors can be applied to the integration variables, and we choose convenient factors for later calculations. Letting $d\mathbf{Q} \equiv \prod_{a,b} dQ_{ab}$, $d\tilde{\mathbf{Q}} \equiv \prod_{a,b} d\tilde{Q}_{ab}$, $d\mathbf{R} \equiv \prod_a dR^a$, and $d\mathbf{R}_0 \equiv \prod_a dR_0^a$, we obtain

$$\begin{aligned}
\Xi_k(\beta) \quad &= \quad \int \exp\{p\alpha \log I\}\nu^k(d\mathbf{w}) \\
&\qquad \int \exp\left\{i\beta p\left(\sum_{ab}\left[Q_{ab} - \frac{(\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b}{p}\right]\tilde{Q}_{ab} + \sum_a\left[R^a - \frac{(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}}{\sqrt{p}}\right]R_0^a\right)\right\} d\mathbf{Q}d\tilde{\mathbf{Q}}d\mathbf{R}d\mathbf{R}_0 \\
&= \quad \int \exp\left\{-p\left[-\alpha\log I - i\beta\left\{\sum_{ab}\left[Q_{ab} - \frac{(\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b}{p}\right]\tilde{Q}_{ab} + \sum_a\left[R^a - \frac{(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}}{\sqrt{p}}\right]R_0^a\right\}\right]\Bigg]\right\} \\
&\qquad \nu^k(d\mathbf{w})d\mathbf{Q}d\tilde{\mathbf{Q}}d\mathbf{R}d\mathbf{R}_0 \\
&= \quad \int \exp\left\{-p\left[-\log I - i\beta\left(\sum_{ab} Q_{ab}\tilde{Q}_{ab} + \sum_a R^aR_0^a\right)\right]\Bigg]\right\} \\
&\qquad \left[\int \exp\left\{-i\beta\sum_{ab}\tilde{Q}_{ab}(\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b - i\beta\sqrt{p}\sum_a R_0^a(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}\right\}\nu^k(d\mathbf{w})\right] d\mathbf{Q}d\tilde{\mathbf{Q}}d\mathbf{R}d\mathbf{R}_0,
\end{aligned}$$

which can be rewritten as

$$\Xi_k(\beta) \quad = \quad \int \exp\left\{-p\mathcal{S}_k(\mathbf{Q}, \tilde{\mathbf{Q}}, \mathbf{R}, \mathbf{R}_0)\right\} d\mathbf{Q}d\tilde{\mathbf{Q}}d\mathbf{R}d\mathbf{R}_0, \tag{A-16}$$

where

$$\mathcal{S}_k(\mathbf{Q}, \tilde{\mathbf{Q}}, \mathbf{R}, \mathbf{R}_0) \quad = \quad -i\beta\left(\sum_{ab} Q_{ab}\tilde{Q}_{ab} + \sum_a R^aR_0^a\right) - \frac{1}{p}\log\xi(\tilde{\mathbf{Q}}, \mathbf{R}_0) - \hat{\xi}(\mathbf{Q}, \mathbf{R}), \tag{A-17}$$

where

$$\begin{aligned}
\xi(\tilde{\mathbf{Q}}, \mathbf{R}_0) \quad &= \quad \int \exp\left\{-i\beta\sum_{ab}\tilde{Q}_{ab}(\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b - i\beta\sum_a\sqrt{p}R_0^a(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}\right\}\nu^k(d\mathbf{w}), \\
\hat{\xi}(\mathbf{Q}, \mathbf{R}) \quad &= \quad \alpha\log\hat{I},
\end{aligned} \tag{A-18}$$

where $\hat{I}$ can be obtained from (A-15) as

$$\begin{aligned}
\hat{I} \quad = \quad \int \exp\Bigg\{&-\beta\sum_{a=1}^k V(u^a) + i\sqrt{p}\sum_{a=1}^k u^a\tilde{u}^a \\
&-\frac{p}{2}\sum_{ab} Q_{ab}\tilde{u}^a\tilde{u}^b - i\sqrt{p}\sum_{a=1}^k R^a\mu\tilde{u}^a\Bigg\}\nu^k(du)\nu^k(d\tilde{u}).
\end{aligned} \tag{A-19}$$

Now we apply steepest descent method to the remaining integration. According to Varadhan's proposition (Tanaka, 2002), only the saddle points of the exponent of the integrand contribute to the integration in the limit of $p \to \infty$. We next use the saddle point method in (A-16) to obtain

$$-\lim_{p\to\infty}\frac{1}{p}\Xi_k(\beta) \quad = \quad \mathcal{S}_k(\mathbf{Q}^\star, \tilde{\mathbf{Q}}^\star, \mathbf{R}^\star, \mathbf{R}_0^\star),$$

where $\mathbf{Q}^\star, \tilde{\mathbf{Q}}^\star, \mathbf{R}^\star, \mathbf{R}_0^\star$ are the saddle point locations. Looking for saddle-points over all the entire space is in general difficult to perform. We assume replica symmetry for saddle-points such that they are invariant under exchange of any two replica indices a and b, where $a \neq b$. Under this symmetry assumption, the space is greatly reduced and the exponent of the integrand can be explicitly evaluated. The replica symmetry is also motivated by the fact that $\mathcal{S}_k(\mathbf{Q}^\star, \tilde{\mathbf{Q}}^\star, \mathbf{R}^\star, \mathbf{R}_0^\star)$ is indeed left unchanged by such change of variables. This is equivalent to postulating that $R^a = R$, $R_0^a = iR_0$,

$$(Q_{ab})^\star = \left\{\begin{array}{ll} q_1 & \text{if a=b} \\ q_0 & \text{otherwise} \end{array}\right., \quad \text{and} \quad (\tilde{Q}_{ab})^\star = \left\{\begin{array}{ll} i\frac{\beta\zeta_1}{2} & \text{if a=b} \\ i\frac{\beta\zeta_0}{2} & \text{otherwise} \end{array}\right., \tag{A-20}$$

where the factor $i\beta/2$ is for future convenience. The next step consists of substituting the above expressions for $\mathbf{Q}^\star, \tilde{\mathbf{Q}}^\star, \mathbf{R}^\star, \mathbf{R}_0^\star$ in $\mathcal{S}_k(\mathbf{Q}^\star, \tilde{\mathbf{Q}}^\star, \mathbf{R}^\star, \mathbf{R}_0^\star)$ and then taking the limit $k \to 0$.

We will separately consider each term of $\mathcal{S}_k(\mathbf{Q}^\star, \tilde{\mathbf{Q}}^\star, \mathbb{R}^\star, R_0^\star)$. Let us begin with the first term in (A-17).

$$-i\beta\left(\sum_{ab} Q_{ab}\tilde{Q}_{ab} + \sum_a R^a R_0^a\right) \quad = \quad \frac{k\beta^2}{2}(\zeta_1 q_1 - \zeta_0 q_0) + k\beta R R_0 + o(k). \tag{A-21}$$

Taking the limit $\beta \to \infty$, the analysis of the saddle point parameters $q_0, q_1, \zeta_0, \zeta_1$ shows that q_0, q_1 has the same limit as $q_1 - q_0 = (q/\beta) + o(\beta^{-1})$ and ζ_0, ζ_1 has the same limit as $\zeta_1 - \zeta_0 = (-\zeta/\beta) + o(\beta^{-1})$. We have

$$\begin{aligned} -i\beta\left(\sum_{ab} Q_{ab}\tilde{Q}_{ab} + \sum_a R^a R_0^a\right) \quad &= \quad \frac{k\beta^2}{2}((\zeta_0 - \zeta/\beta)(q_0 + q/\beta) - \zeta_0 q_0) + k\beta R R_0 \\ &= \quad \frac{k\beta}{2}(\zeta_0 q - \zeta q_0) + k\beta R R_0. \end{aligned} \tag{A-22}$$

Next consider the second term $\log \xi(\tilde{\mathbf{Q}}, \mathbf{R}_0)$ in (A-17). For p-vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ and $p \times p$ matrix $\boldsymbol{\Sigma}$, introducing the notation $\|\mathbf{v}\|_{\boldsymbol{\Sigma}}^2 \equiv \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$ and $\langle \mathbf{u}, \mathbf{v}\rangle \equiv \sum_{j=1}^p u_j v_j / p$, we have

$$\begin{aligned} \xi(\tilde{\mathbf{Q}}, \mathbf{R}_0) \quad &= \quad \int \exp\left\{\frac{\beta^2}{2}(\zeta_1 - \zeta_0)\sum_{a=1}^k \|\mathbf{w}^a\|_{\boldsymbol{\Sigma}}^2 + \frac{\beta^2\zeta_0}{2}\sum_{a,b=1}^k (\mathbf{w}^a)^T\boldsymbol{\Sigma}\mathbf{w}^b \right. \\ &\qquad \left. + \beta\sqrt{p}\sum_{a=1}^k R_0(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}\right\}\nu^k(d\mathbf{w}) \\ &= \quad \mathbb{E}\int \exp\left\{\frac{\beta^2}{2}(\zeta_1 - \zeta_0)\sum_{a=1}^k \|\mathbf{w}^a\|_{\boldsymbol{\Sigma}}^2 + \beta\sqrt{\zeta_0}\sum_{a=1}^k (\mathbf{w}^a)^T\boldsymbol{\Sigma}^{1/2}\mathbf{z} \right. \\ &\qquad \left. + \beta\sqrt{p}\sum_{a=1}^k R_0(\mathbf{w}^a)^T\hat{\boldsymbol{\mu}}\right\}\nu^k(d\mathbf{w}), \end{aligned} \tag{A-23}$$

where expectation is with respect to $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$. Notice that, given $\mathbf{z} \in \mathbb{R}^p$, the integrals over $\mathbf{w}^1, \cdots, \mathbf{w}^k$ factorize, whence

$$\begin{aligned} \xi(\tilde{\mathbf{Q}}, \mathbf{R}_0) &= \mathbb{E}\left\{\left[\int \exp\left\{\frac{\beta^2}{2}(\zeta_1 - \zeta_0)\|\mathbf{w}\|^2_{\boldsymbol{\Sigma}} + \beta\sqrt{\zeta_0}\mathbf{w}^T\boldsymbol{\Sigma}^{1/2}\mathbf{z} \right.\right.\right. \\ &\qquad \left.\left.\left. + \beta\sqrt{p}R_0\mathbf{w}^T\hat{\boldsymbol{\mu}}\right\}\nu(d\mathbf{w})\right]^k\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} \log\xi(\tilde{\mathbf{Q}}, \mathbf{R}_0) &= \log\mathbb{E}\left\{\left[\int \exp\left\{-\beta\frac{\zeta}{2}\|\mathbf{w}\|^2_{\boldsymbol{\Sigma}} + \beta\sqrt{\zeta_0}\mathbf{w}^T\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \beta\sqrt{p}R_0\mathbf{w}^T\hat{\boldsymbol{\mu}}\right\}\nu(d\mathbf{w})\right]^k\right\} \\ &= \mathbb{E}\min_{\mathbf{w},\mathbf{v}\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\|\mathbf{w}\|^2_{\boldsymbol{\Sigma}} - \left\langle\sqrt{\zeta_0}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \sqrt{p}R_0\hat{\boldsymbol{\mu}}, \mathbf{w}\right\rangle \right. \\ &\qquad \left. + \sum_{j=1}^p\left[J_\lambda(w_j - s\tilde{d}(\boldsymbol{\Sigma}^{-1}\mathbf{v})_j) + s\left\{g(v_j) - v_jw_j + s\tilde{d}v_j(\boldsymbol{\Sigma}^{-1}\mathbf{v})_j\right\}\right]\right\} \end{aligned}$$

After changing the variable $\mathbf{w} - s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v} \to \mathbf{w}$, we obtain

$$\begin{aligned} \log\xi(\tilde{\mathbf{Q}}, \mathbf{R}_0) &= -k\beta\mathbb{E}\min_{\mathbf{w},\mathbf{v}\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\|\mathbf{w} + s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}\|^2_{\boldsymbol{\Sigma}} - \left\langle\sqrt{\zeta_0}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \sqrt{p}R_0\hat{\boldsymbol{\mu}}, \mathbf{w} + s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}\right\rangle \right. \\ &\qquad \left. + \sum_{j=1}^p\left[J_\lambda(w_j) + s\left\{g(v_j) - v_jw_j\right\}\right]\right\}. \qquad \text{(A-24)} \end{aligned}$$

Finally, we consider the third term in (A-17). After integration over $\nu^k(d\tilde{u})$, (A-19) becomes

$$\begin{aligned} \hat{I} &= \int \exp\left\{-\beta\sum_{a=1}^k V\left(u^a\right) - \frac{1}{2}\sum_{ab}(u^a - R\mu)(\mathbf{Q}^{-1})_{ab}(u^b - R\mu) \right. \\ &\qquad \left. -\frac{1}{2}\log\det\mathbf{Q}\right\}\nu^k(du). \qquad \text{(A-25)} \end{aligned}$$

Similarly, using (A-20), we obtain

$$\begin{aligned} \sum_{ab}(u^a - R\mu)(\mathbf{Q}^{-1})_{ab}(u^b - R\mu) &= \frac{\beta\sum_a(u^a - R\mu)^2}{q} - \frac{\beta^2 q_0\{\sum_a(u^a - R\mu)\}^2}{q^2}, \\ \log\det\mathbf{Q} &= \log\left[(q_1 - q_0)^k\left(1 + \frac{kq_0}{q_1 - q_0}\right)\right] = \frac{k\beta q_0}{q}, \end{aligned}$$

where we retain only the leading order terms. Therefore, (A-25) becomes

$$
\begin{aligned}
\hat{I} &= \int \exp\left\{-\beta \sum_{a=1}^{k} V\left(u^a\right) - \frac{\beta \sum_a (u^a - R\mu)^2}{2q} + \frac{\beta^2 q_0 \{\sum_a (u^a - R\mu)\}^2}{2q^2} \right. \\
&\qquad \left. - \frac{1}{2}\frac{k\beta q_0}{q}\right\} \nu^k(du) \\
&= \int Dz \int \exp\left\{-\beta \sum_{a=1}^{k} V\left(u^a\right) - \frac{\beta \sum_a (u^a - R\mu)^2}{2q} + \frac{\beta \sum_a (u^a - R\mu)\sqrt{q_0} z}{q} \right. \\
&\qquad \left. - \frac{1}{2}\frac{k\beta q_0}{q}\right\} \nu^k(du) \\
&= \exp\left(-\frac{k\beta q_0}{2q}\right) \int Dz \left(\int \exp\left\{-\beta V(u) - \frac{\beta(u - R\mu - \sqrt{q_0} z)^2}{2q} + \frac{\beta q_0 z^2}{2q}\right\} du\right)^k,
\end{aligned}
$$

where the expectation $Dz = \int \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$. Substituting this expression into (A-18), we obtain

$$
\begin{aligned}
\hat{\xi}(\mathbf{Q}, \mathbf{R}) &= \alpha \log \hat{I} = -\frac{k\alpha\beta q_0}{2q} + \alpha \log \int Dz \exp(k \log G) \\
&= -\frac{k\beta\alpha q_0}{2q} + \alpha \log\left(1 + k \int Dz \log G\right) \\
&= -\frac{k\beta\alpha q_0}{2q} + k\alpha \int Dz \log G, \qquad \text{(A-26)}
\end{aligned}
$$

where

$$
\begin{aligned}
\log G &= \log \int \exp\left\{-\beta V(u) - \frac{\beta(u - R\mu - \sqrt{q_0} z)^2}{2q} + \frac{\beta q_0 z^2}{2q}\right\} du \\
&= -\beta \min_u \left[V(u) + \frac{(u - R\mu - \sqrt{q_0} z)^2}{2q} - \frac{q_0 z^2}{2q}\right]
\end{aligned}
$$

in the limit of $\beta \to \infty$. Substituting this expression into (A-26), we obtain

$$
\hat{\xi}(\mathbf{Q}, \mathbf{R}) = -k\alpha\beta E\left\{\min_u \left[V(u) + \frac{(u - R\mu - \sqrt{q_0} z)^2}{2q}\right]\right\}, \qquad \text{(A-27)}
$$

where the expectation is with respect to $z \sim N(0, 1)$.

Putting (A-22), (A-24), and (A-27) together into (A-16) and then into (A-10), we obtain

$$
\begin{aligned}
\mathcal{F}(s) &= \frac{\zeta_0 q - \zeta q_0}{2} + RR_0 + \alpha \mathrm{E} \min_{u \in \mathbb{R}} \left\{V(u) + \frac{\left(u - R\mu - \sqrt{q_0} z\right)^2}{2q}\right\} \\
&\quad + \frac{1}{p} \mathrm{E} \min_{\mathbf{w}, \mathbf{v} \in \mathbb{R}^p} \left\{\frac{\zeta}{2} \|\mathbf{w} + s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}\|_{\boldsymbol{\Sigma}}^2 - \left\langle \sqrt{\zeta_0} \boldsymbol{\Sigma}^{1/2} \mathbf{z} + \sqrt{p} R_0 \hat{\boldsymbol{\mu}}, \mathbf{w} + s\tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v} \right\rangle \right. \\
&\quad \left. + \sum_{j=1}^{p} \left[J_\lambda(w_j) + s\left\{g(v_j) - v_j w_j\right\}\right]\right\}, \qquad \text{(A-28)}
\end{aligned}
$$

where the expectations are with respect to $z \sim N(0,1)$, and $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$, with z and $\mathbf{z}$ independent from each other.

$$\begin{aligned} & \frac{d\mathcal{F}}{ds}(s=0) \\ = & \lim_{p\to\infty}\frac{1}{p}\min_{\mathbf{v}\in\mathbb{R}^p} E\left\{\zeta\tilde{d}\hat{\mathbf{w}}_{\mathbf{z}}^T\mathbf{v} - \left\langle \sqrt{\zeta_0}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \sqrt{p}R_0\hat{\boldsymbol{\mu}}, \tilde{d}\boldsymbol{\Sigma}^{-1}\mathbf{v}\right\rangle + \sum_{j=1}^{p}\left[g(v_j) - v_j\hat{w}_j\right]\right\}, \end{aligned}$$

where

$$\hat{\mathbf{w}}_{\mathbf{z}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\left\|\mathbf{w} - \frac{\sqrt{\zeta_0}}{\zeta}\boldsymbol{\Sigma}^{-1/2}\mathbf{z} - \frac{\sqrt{p}R_0\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}}{\zeta}\right\|_{\boldsymbol{\Sigma}}^2 + \sum_{j=1}^{p}J_\lambda(w_j)\right\}. \qquad \text{(A-29)}$$

At this point, we choose $\tilde{d} = 1/\zeta$. Minimizing over $\mathbf{v}$ (recall that $\tilde{g}(x) = \max_{u\in\mathbb{R}}[ux - g(u)]$, we get

$$\begin{aligned} & \frac{d\mathcal{F}}{ds}(s=0) \\ = & \lim_{p\to\infty}\frac{1}{p}\min_{\mathbf{v}\in\mathbb{R}^p} E\left\{\sum_{j=1}^{p}\left[g(v_j, w_{0j}) - v_j(\sqrt{\zeta_0}\boldsymbol{\Sigma}^{-1/2}\mathbf{z} + \sqrt{p}R_0\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}})_j/\zeta\right]\right\} \\ = & -\lim_{p\to\infty}\frac{1}{p}E\left\{\sum_{j=1}^{p}\tilde{g}\left((\sqrt{\zeta_0}\boldsymbol{\Sigma}^{-1/2}\mathbf{z} + \sqrt{p}R_0\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}})_j/\zeta\right)\right\}. \end{aligned}$$

Renaming $\zeta_0 = \zeta^2\tau^2$, we get out final expression for $\frac{d\mathcal{F}}{ds}(s=0)$ as

$$\frac{d\mathcal{F}}{ds}(s=0) = -\lim_{p\to\infty}\frac{1}{p}E\left\{\sum_{j=1}^{p}\tilde{g}\left((\tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z})_j + \sqrt{p}R_0(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}})_j/\zeta\right)\right\}. \qquad \text{(A-30)}$$

Compared with (A-6), this shows that the distribution limit of $\bar{\mathbf{w}}$, defined in (A-7), is the same as $\tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z} + \sqrt{p}R_0\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}/\zeta$.

Here, $\zeta, \zeta_0, q, q_0, R, R_0$ are the order parameters which can be determined from the saddle-point equations of $\mathcal{F}(s=0)$. Define the functions F, G, and H as

$$\begin{aligned} F &= \mathbb{E}_z\left(\hat{u} - R\mu - \sqrt{q_0}z\right), \\ G &= \mathbb{E}_z\left\{\left(\hat{u} - R\mu - \sqrt{q_0}z\right)z\right\}, \\ H &= \mathbb{E}_z\left\{\left(\hat{u} - R\mu - \sqrt{q_0}z\right)^2\right\}, \end{aligned}$$

where

$$\hat{u} = \arg\min_{u\in\mathbb{R}}\left\{V(u) + \frac{(u - R\mu - \sqrt{q_0}z)^2}{2q}\right\}.$$

Then all the order parameters can be determined by the following saddle-point equations derived from (A-28):

$$\begin{aligned}
\xi_0 &= \frac{\alpha}{q^2} H, && \text{(A-31)}\\
\xi &= \frac{\alpha G}{\sqrt{q_0} q}, && \text{(A-32)}\\
q_0 &= \frac{1}{p} \mathbb{E}_z \|\hat{\mathbf{w}}_{\mathbf{z}}\|_{\mathbf{\Sigma}}^2, && \text{(A-33)}\\
q &= \frac{1}{p\sqrt{\zeta_0}} \mathbb{E} \left\langle \mathbf{\Sigma}^{1/2}\mathbf{z}, \hat{\mathbf{w}}_{\mathbf{z}} \right\rangle && \text{(A-34)}\\
R &= \frac{1}{\sqrt{p}} \mathbb{E}_z \langle \hat{\boldsymbol{\mu}}, \hat{\mathbf{w}}_{\mathbf{z}} \rangle, && \text{(A-35)}\\
R_0 &= \frac{\alpha \mu}{q}, && \text{(A-36)}
\end{aligned}$$

where $\hat{\mathbf{w}}$ is solved from (A-29). The result in (A-28) is for the general penalty function $J_\lambda(w)$. For quadratic penalty $J_\lambda(w) = \lambda w^2$, we get the closed form limiting distribution of $\hat{\mathbf{w}}$ as

$$\hat{\mathbf{w}}_{\mathbf{z}} = (\xi \mathbf{\Sigma} + \lambda \mathbf{I}_p)^{-1} \left(\sqrt{\xi_0} \mathbf{\Sigma}^{1/2} \mathbf{z} + \sqrt{p} R_0 \hat{\boldsymbol{\mu}} \right). \quad \text{(A-37)}$$

For L_1-penalty $J_\lambda(w) = \lambda |w|$, (A-29) becomes a LASSO type of regression problem which only has a closed form solution in the i.i.d. situation, i.e. $\mathbf{\Sigma} = \mathbf{I}_p$. For general $\mathbf{\Sigma}$, we rely on a numerical algorithm to solve $\hat{\mathbf{w}}_{\mathbf{z}}$.

A.2 Derivation of Result 2

The prediction accuracy for binary classification is given by

$$\mathbb{E}\mathrm{I}(y\mathbf{x}^T\hat{\mathbf{w}} \geq 0) = P(y=1)\mathbb{E}\{\mathrm{I}(\mathbf{x}^T\hat{\mathbf{w}} \geq 0)|y=1\} + P(y=-1)\mathbb{E}\{\mathrm{I}(\mathbf{x}^T\hat{\mathbf{w}} \leq 0)|y=-1\}.$$

Conditional on $y = 1$, $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$. Thus, $\mathbf{x}^T\hat{\mathbf{w}} \sim N(\hat{\mathbf{w}}^T\boldsymbol{\mu}, \hat{\mathbf{w}}^T\mathbf{\Sigma}\hat{\mathbf{w}})$ and $\mathbb{E}\{\mathrm{I}(\mathbf{x}^T\hat{\mathbf{w}} \geq 0)|y=1\} = \Phi(\delta)$, where $\delta = \frac{\hat{\mathbf{w}}^T\boldsymbol{\mu}}{\sqrt{\hat{\mathbf{w}}^T\mathbf{\Sigma}\hat{\mathbf{w}}}}$. Similar we can derive that conditional on $y = -1$, $\mathbb{E}\{\mathrm{I}(\mathbf{x}^T\hat{\mathbf{w}} \leq 0)|y = -1\} = \Phi(\delta)$. Therefore $\mathbb{E}\mathrm{I}(y\mathbf{x}^T\hat{\mathbf{w}} \geq 0) = \Phi(\delta)$, and from (A-33) and (A-35), we get $\delta = \frac{R\mu}{\sqrt{q_0}}$.

A.3 Numerical Implementation

The parameters ζ_0, ζ, R_0, q_0, q, and R, which are determined by the nonlinear equations in (7) of Proposition 1, play an important role in characterizing the properties of $\hat{\mathbf{w}}$. This subsection discusses the algorithm used to calculate these parameters. An implementation is available at the GitHub repository (`https://github.com/pzengauburn/classification-lasso`).

For the L_1-penalty, (7) does not have a closed-form solution, so we use an iterative numerical algorithm to solve it. The parameters can be partitioned into two groups: $\{\zeta_0, \zeta, R_0\}$ and $\{q_0, q, R\}$, with one group updated while the other is held fixed in each iteration. Starting with an initial set of values $\{\zeta_0^{(0)}, \zeta^{(0)}, R_0^{(0)}\}$, we iteratively compute these parameters as follows for $k = 1, 2, \ldots$ until convergence.

- For given $\{\zeta_0^{(k)}, \zeta^{(k)}, R_0^{(k)}\}$, calculate $\{q_0^{(k+1)}, q^{(k+1)}, R^{(k+1)}\}$ by

  $$\begin{aligned} q_0^{(k+1)} &= \frac{1}{p}\mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \boldsymbol{\Sigma} \hat{\mathbf{w}}_{\mathbf{z}}), \\ q^{(k+1)} &= \frac{1}{p\sqrt{\zeta_0^{(k)}}}\mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \boldsymbol{\Sigma}^{1/2} \mathbf{z}), \\ R^{(k+1)} &= \frac{1}{\sqrt{p}}\mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \hat{\boldsymbol{\mu}}), \end{aligned}$$

  where $\hat{\mathbf{w}}_{\mathbf{z}}$ is given in (9) which depends on $\{\zeta_0^{(k)}, \zeta^{(k)}, R_0^{(k)}\}$. It requires high-dimensional numerical integration to compute these expectations because there is no explicit expression for $\hat{\mathbf{w}}_{\mathbf{z}}$ when $J_\lambda(w) = \lambda|w|$ and $\boldsymbol{\Sigma}$ are general. In our implementation, we use Monte Carlo integration to compute the expectation. For a general $\boldsymbol{\Sigma}$, (9) is essentially a LASSO problem that can be efficiently solved by the coordinate descent algorithm (Wu and Lange, 2008).

- For given $\{q_0^{(k+1)}, q^{(k+1)}, R^{(k+1)}\}$, calculate $\{\zeta_0^{(k+1)}, \zeta^{(k+1)}, R_0^{(k+1)}\}$ by

  $$\begin{aligned} \zeta_0^{(k+1)} &= \frac{\alpha}{{q^{(k+1)}}^2}\mathbb{E}\left(\hat{u}_\epsilon - R^{(k+1)}\mu - \sqrt{q_0^{(k+1)}}\epsilon\right)^2, \\ \zeta^{(k+1)} &= -\frac{\alpha}{q^{(k+1)}\sqrt{q_0^{(k+1)}}}\mathbb{E}\left[\left(\hat{u}_\epsilon - R^{(k+1)}\mu - \sqrt{q_0^{(k+1)}}\epsilon\right)\epsilon\right], \\ R_0^{(k+1)} &= \frac{\alpha\mu}{q^{(k+1)}}\mathbb{E}\left(\hat{u}_\epsilon - R^{(k+1)}\mu - \sqrt{q_0^{(k+1)}}\epsilon\right), \end{aligned}$$

  where $\hat{u}_\epsilon$ is given in (8). The crucial step in this iteration involves evaluating the integrations $\mathbb{E}(\hat{u}_\epsilon^2)$, $\mathbb{E}(\hat{u}_\epsilon \epsilon)$, and $\mathbb{E}(\hat{u}_\epsilon)$. Since there is no explicit expression for $\hat{u}_\epsilon$ in many applications, these expectations are computed numerically. Because $\hat{u}_\epsilon$ is univariate, numerical integration is shown to be efficient and straightforward. Taken logistic regression as an example which is considered one of the most popular and successful classification methods available on the shelf, $\hat{u}_\epsilon$ can be obtained as the the root of equation

  $$\frac{1}{e^u + 1} = \frac{u - R^{(k+1)}\mu - \sqrt{q_0^{(k+1)}}\epsilon}{q^{(k+1)}},$$

  which is obtained by setting the first-order derivative of the objective function in (8) to be zero.

In practice, convergence can be slow when values oscillate between regions. To accelerate convergence, damping can be applied to update the parameter as a linear combination of its current value and the newly computed one. For example, q_0 is updated as

$$q_0^{(k+1)} = sq_0^{(k)} + (1-s)\frac{1}{p}\mathbb{E}(\hat{\mathbf{w}}_{\mathbf{z}}^T \boldsymbol{\Sigma} \hat{\mathbf{w}}_{\mathbf{z}}),$$

where $s \in [0, 1]$. A similar damping approach can be applied to other parameters. Choosing $s = 0.5$ generally produces good performance. Increase s if oscillations persist, while gradually decrease s if convergence is too slow.

References

Barbier, J. and N. Macris (2018). The adaptive interpolation method: A simple scheme to prove replica formulas in bayesian inference.

Bayati, M. and A. Montanari (2011). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory 58*(4), 1997–2017.

Berthier, R., A. Montanari, and P.-M. Nguyen (2020). State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA 9*(1), 33–79.

Blanchard, G., O. Bousquet, and P. Massart (2008). Statistical performance of support vector machines. *The Annals of Statistics 36*(2), 489 – 531.

Candès, E. J., P. Sur, et al. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics 48*(1), 27–42.

Deng, Z., A. Kammoun, and C. Thrampoulidis (2019). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA.*

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics 46*(1), 247–279.

Engel, A. and C. Van den Broeck (2001). *Statistical Mechanics of Learning.* Cambridge University Press.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Annals of statistics 36*(6), 2605.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119 – 139.

Friedman, J., T. Hastie, and R. Tibshirani (2000, 04). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics 28*(2), 337–407.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*, 1–22.

Gerace, F., B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová (2020). Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR.

Gerbelot, C., A. Abbara, and F. Krzakala (2023). Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *IEEE Transactions on Information Theory 69*(3), 1824–1852.

Gordon, Y. (1985). Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics 50*, 265–289.

Hall, P., J. S. Marron, and A. Neeman (2005, 05). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(3), 427–444.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer.

Hertz, J., R. G. Palmer, and A. S. Krogh (1991). *Introduction to the Theory of Neural Computation* (1st ed.). Perseus Publishing.

Huang, H. (2017). Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research 18*, 45:1–45:21.

Huang, H. and Q. Yang (2021, nov). Large dimensional analysis of general margin based classification methods. *Journal of Statistical Mechanics: Theory and Experiment 2021*(11), 113401.

Javanmard, A. and A. Montanari (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*(82), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Kabashima, Y. (2008). Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *Journal of Physics: Conference Series 95*(1), 012001.

Lin, X., G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein (2000). Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv. *The Annals of Statistics 28*(6), 1570–1600.

Liu, Y., H. H. Zhang, and Y. Wu (2011). Soft or hard classification? large margin unified machines. *Journal of the American Statistical Association 106*, 166–177.

Loureiro, B., C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mézard, and L. Zdeborová (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Loureiro, B., G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborova (2021). Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems (M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, eds.)*, Volume 34, pp. 10144–10157. Curran Associates, Inc.

Lu, S., Y. Liu, L. Yin, and K. Zhang (2016). Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a–n/a.

Mai, X. and R. Couillet (2018). Statistical analysis and improvement of large dimensional svm. *private communication*.

Mai, X., Z. Liao, and R. Couillet (2019, May). A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3357–3361.

Marron, J. S., M. Todd, and J. Ahn (2007). Distance-weighted discrimination. *Journal of the*

American Statistical Association 102, 1267–1271.

Mézard, M. and A. Montanari (2009). *Information, Physics, and Computation.* Oxford Graduate Texts. OUP Oxford.

Mézard, M., G. Parisi, and M. A. Virasoro (1987). *Spin Glass Theory and Beyond.* Singapore: World Scientific.

Mignacco, F., F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova (2020). The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International Conference on Machine Learning*, pp. 6874–6883. PMLR.

Montanari, A., F. Ruan, Y. Sohn, and J. Yan (2019). The generalization error of maxmargin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.

Na, S., T. Huang, Y. Liu, T. Takahashi, Y. Kabashima, and X. Wang (2023). Compressed sensing radar detectors under the row-orthogonal design model: A statistical mechanics perspective. *IEEE Transactions on Signal Processing 71*, 2668–2682.

Parisi, G. (1979, Dec). Infinite number of order parameters for spin-glasses. *Phys. Rev. Lett. 43*, 1754–1756.

Shen, X., G. C. Tseng, X. Zhang, and W. H. Wong (2003). On ψ-learning. *Journal of the American Statistical Association 98*(463), 724–734.

Takahashi, T. and Y. Kabashima (2018). A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment 2018*(7), 073405.

Takahashi, T. and Y. Kabashima (2022). Macroscopic analysis of vector approximate message passing in a model-mismatched setting. *IEEE Transactions on Information Theory 68*(8), 5579–5600.

Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models.* Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg.

Tanaka, T. (2002). A statistical-mechanics approach to large-system analysis of cdma multiuser detectors. *Information Theory, IEEE Transactions on 48*(11), 2888–2910.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014, 03). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*, 1166 – 1202.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* New York, NY: Springer.

Verhaak, R., K. Hoadley, E. Purdom, V. Wang, Y. Qi, M. Wilkerson, C. Miller, L. Ding, T. Golub, J. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. Feiler, J. Hodgson, C. James, J. Sarkaria, C. Brennan, A. Kahn, P. Spellman, R. Wilson, T. Speed, J. Gray, M. Meyerson, G. Getz, C. Perou, and D. Hayes (2010, January). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell 17*(1), 98–110.

Wang, X., Z. Yang, X. Chen, and W. Liu (2019). Distributed inference for linear support vector machine. *Journal of Machine Learning Research 20*(113), 1–41.

Wu, T. T. and K. Lange (2008, March). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics 2*(1).

Zhang, X., Y. Wu, L. Wang, and R. Li (2016). Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(1), 53–76.

Zhu, J. and T. Hastie (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics 14*(1), 185–205.

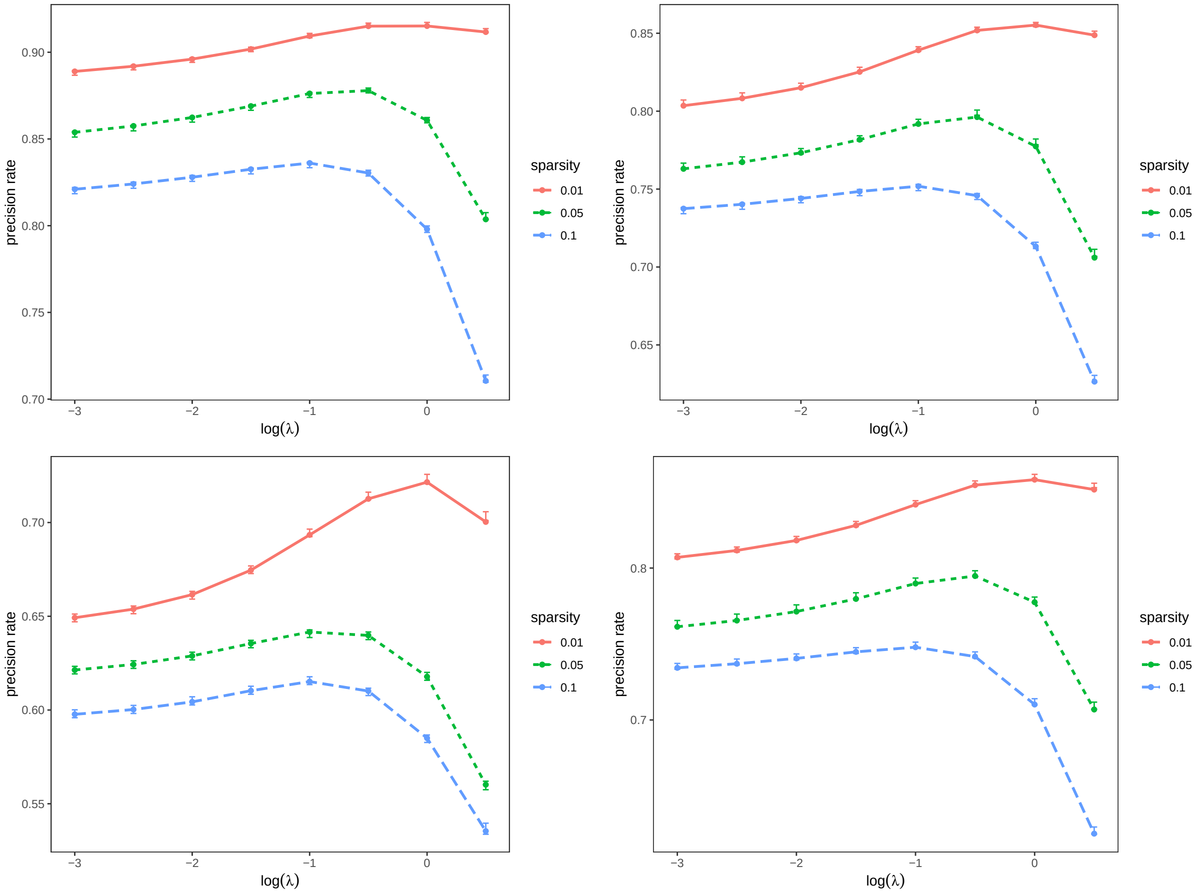

Figure 1: Comparison between theoretical and empirical precision rates for four different correlation structures: IID (top-left), block (top-right), AR1 (bottom-left), and banded (bottom-right). In each plot, the three lines are the theoretical precision rates at different sparsity levels $\epsilon = 0.01, 0.05, 0.1$. The error bars are the 95% confidence intervals of the mean precision rates based on 500 replicates.

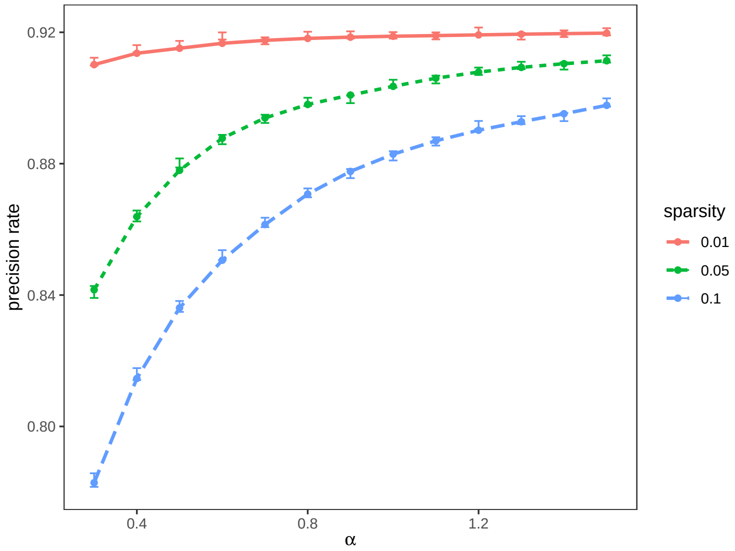

Figure 2: The precision rate at the optimal λ as a function of α under an IID correlation structure. The three lines are the asymptotic precision rates at different sparsity levels $\epsilon = 0.01, 0.05, 0.1$. The error bars are the 95% confidence intervals of the mean precision rates based on 500 replicates.

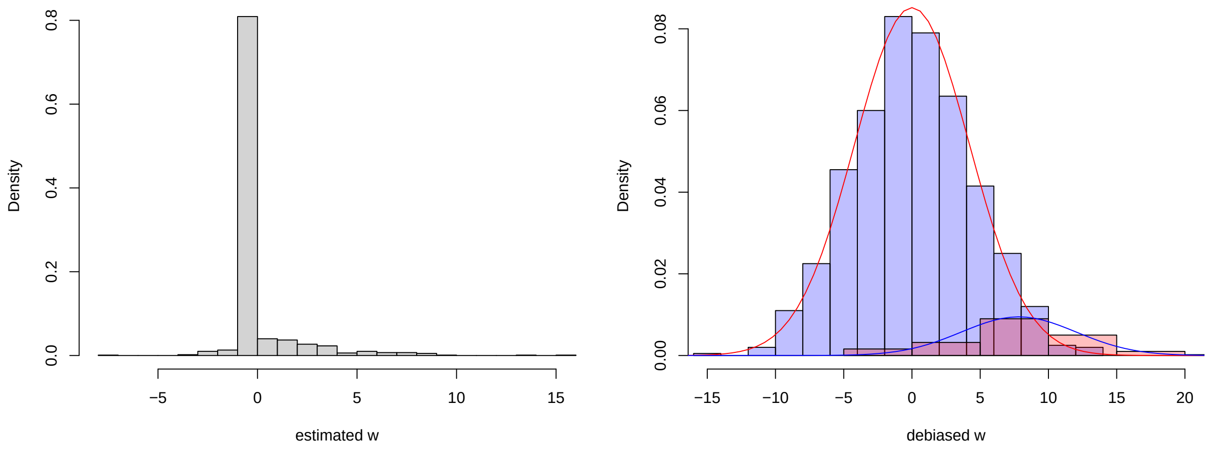

Figure 3: Histograms of the components of PLR estimator $\hat{\mathbf{w}}$ (left) and the corresponding de-biased estimator $\bar{\mathbf{w}}$ (right) for a typical dataset. In the right plot, the curves represent the asymptotic normal densities for the zero and nonzero components.

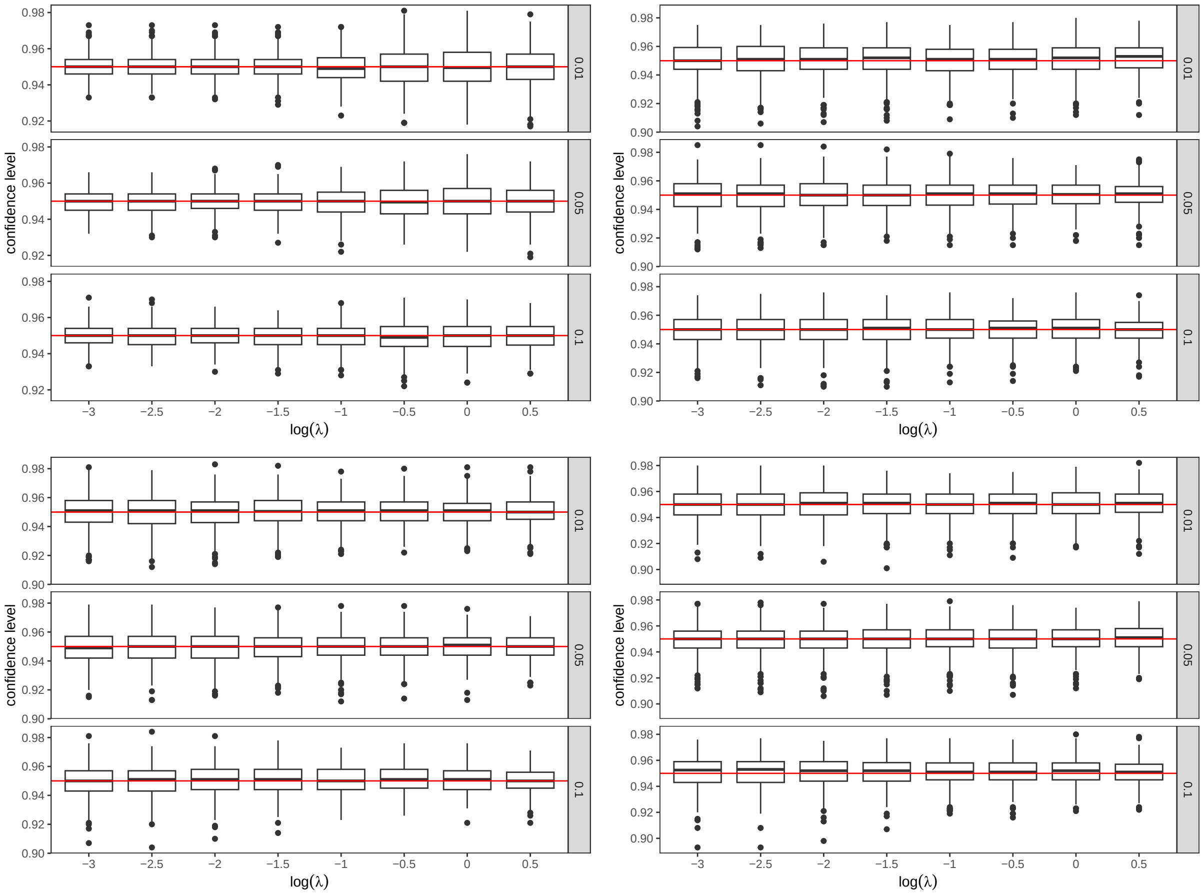

Figure 4: Boxplots of empirical confidence levels based on 500 replicates for four correlation structures: IID (top-left), block (top-right), AR1 (bottom-left), and banded (bottom-right). In each plot, the horizontal line indicates the nominal confidence level 0.95. Different rows of plots represent different sparsity levels, $\epsilon = 0.01, 0.05, 0.1$.

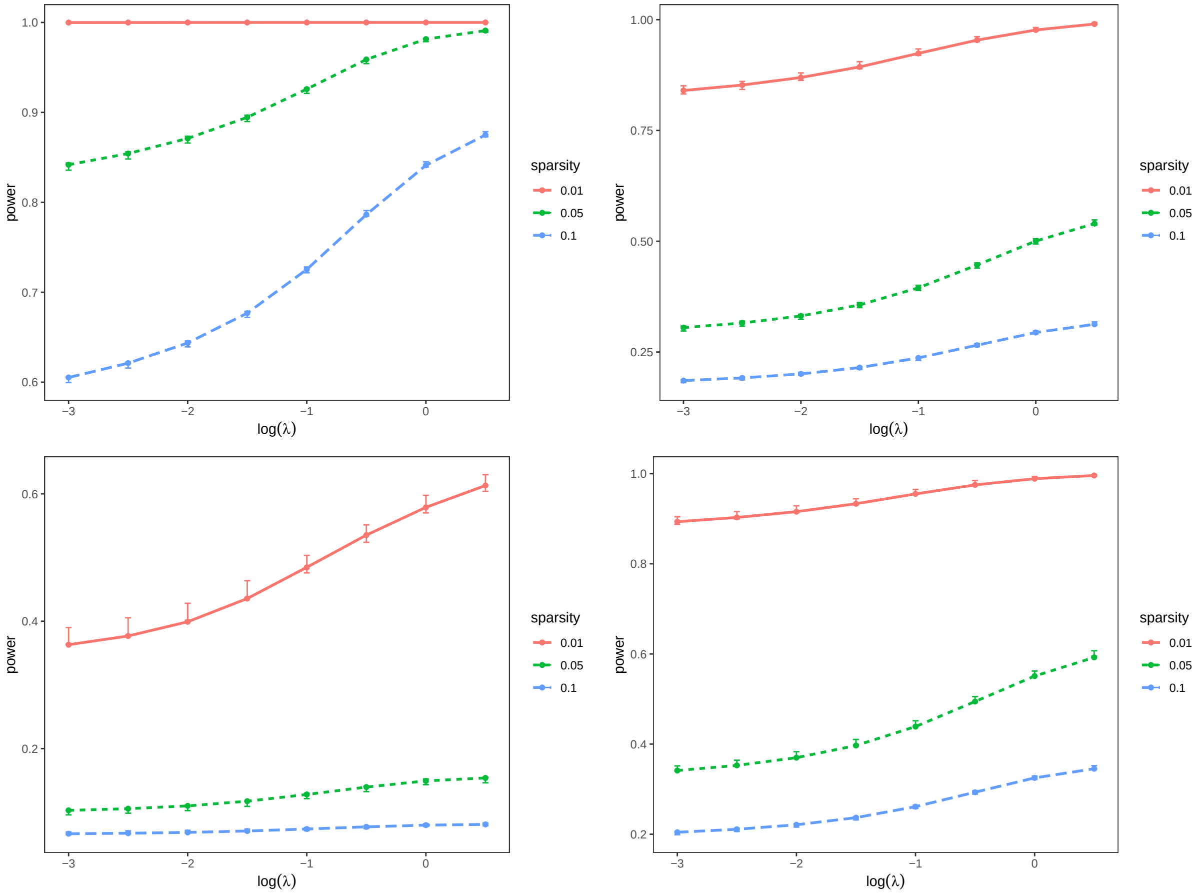

Figure 5: Comparison between theoretical and empirical powers of hypothesis testing for four different correlation structures: IID (top-left), block (top-right), AR1 (bottom-left), and banded (bottom-right). In each plot, the three lines are the theoretical powers at different sparsity levels $\epsilon = 0.01, 0.05, 0.1$. The error bars are the 95% confidence intervals of the mean powers based on 500 replicates.

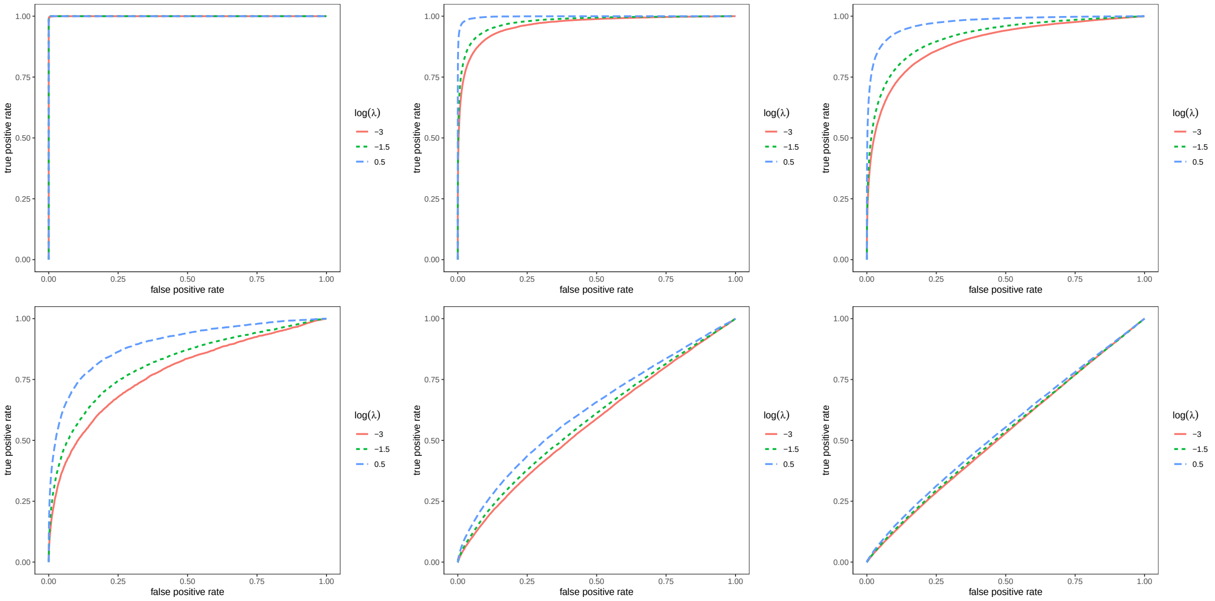

Figure 6: The ROC curve for the IID (top row) and AR1 (bottom row) correlation structures for different sparsity level $\epsilon = 0.01$ (left column), 0.05 (middle column), and 0.1 (right column). The three lines in each plot correspond to $\log \lambda = -3, -1.5, 0.5$, respectively.

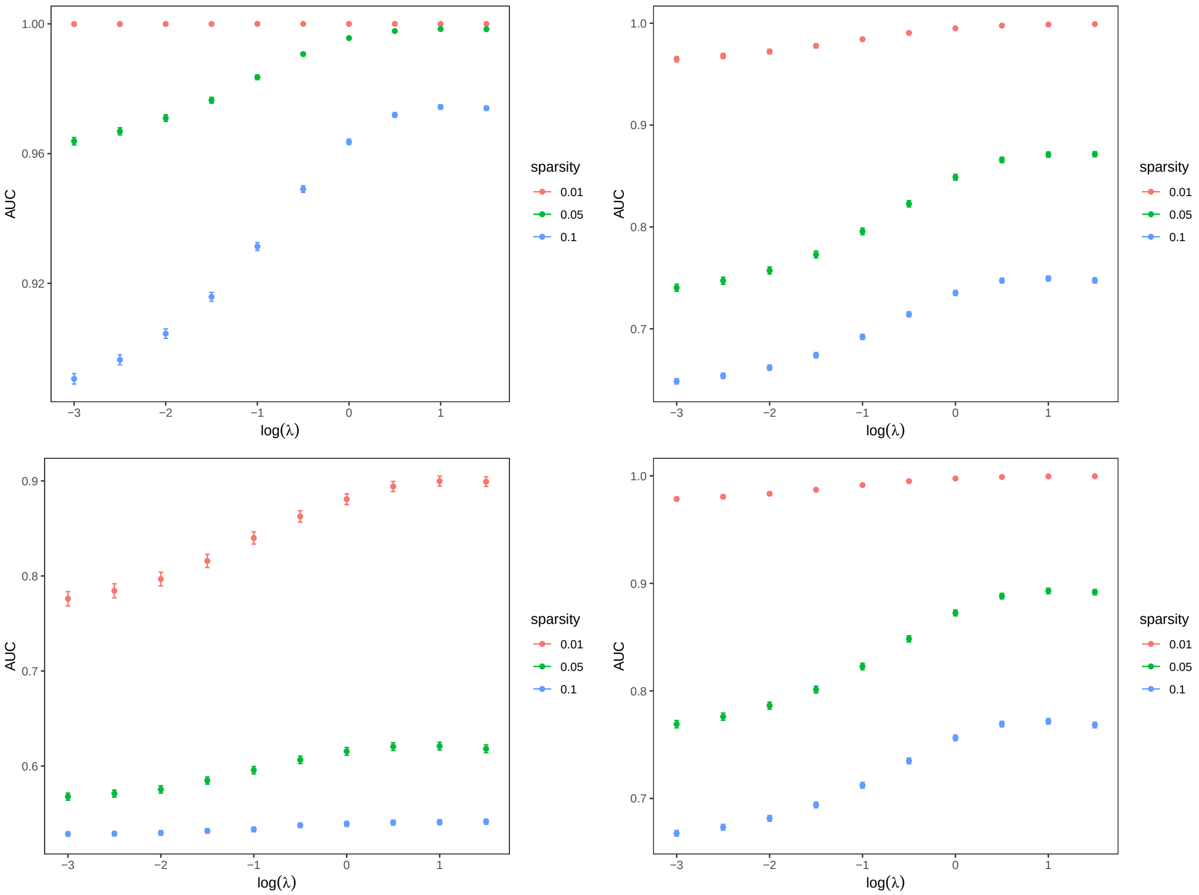

Figure 7: Empirical AUC for four different correlation structures: IID (top-left), block (top-right), AR1 (bottom-left), and banded (bottom-right). The error bars are the 95% confidence intervals of the mean AUC based on 500 replicates at different sparsity levels $\epsilon = 0.01, 0.05, 0.1$.

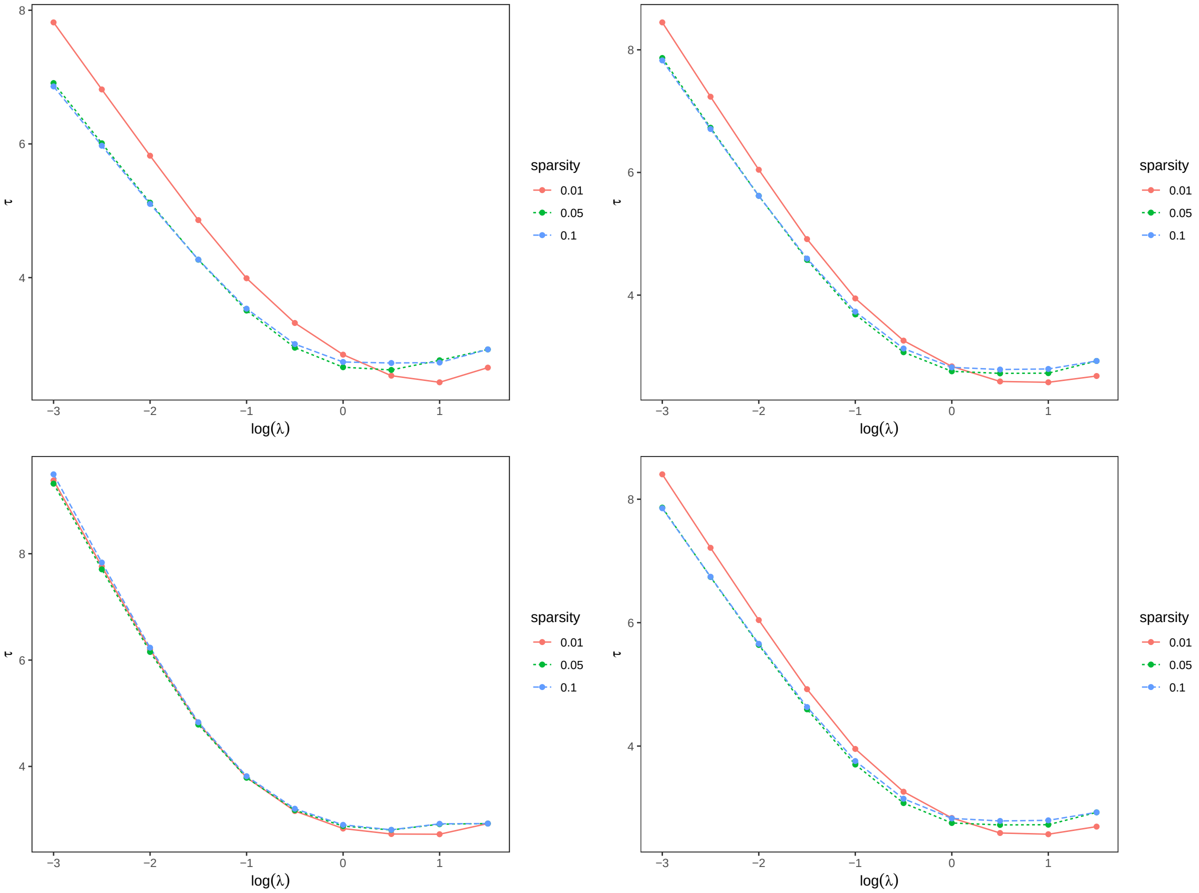

Figure 8: τ as a function of $\log(\lambda)$ for four different correlation structures: IID (top-left), block (top-right), AR1 (bottom-left), and banded (bottom-right). The three lines represent different sparsity levels $\epsilon = 0.01, 0.05, 0.1$.

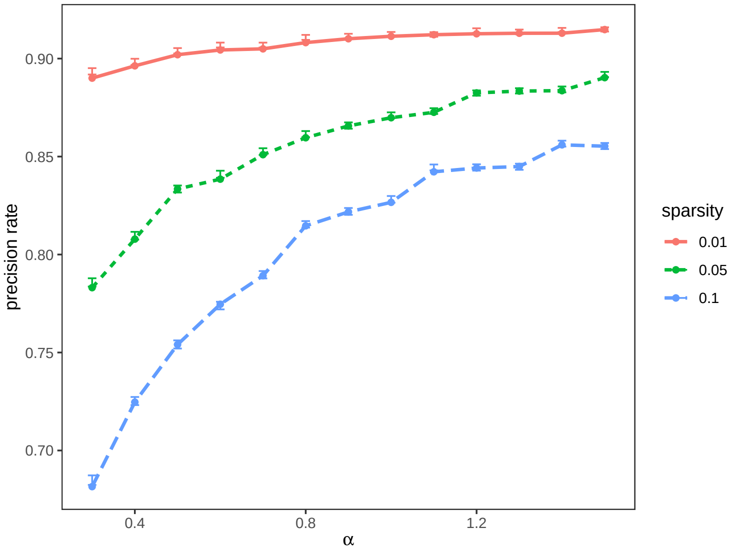

Figure 9: The precision rate at the optimal λ that minimizes τ as a function of α under an IID correlation structure. The three lines are the asymptotic precision rates at different sparsity levels $\epsilon = 0.01, 0.05, 0.1$. The error bars are the 95% confidence intervals of the mean precision rates based on 500 replicates.