A Naïve Bayes Regularized Logistic Regression Model for Classification

Yi Tan[a], Ben Sherwood[b], Prakash P. Shenoy[b]

[a] *The University of Alabama in Huntsville, College of Business, Huntsville, 35899, AL, USA*
[b] *University of Kansas, School of Business, Lawrence, 66045, KS, USA*

Abstract

To reduce the estimator's variance and prevent overfitting, regularization techniques have attracted great interest from the statistics and machine learning communities. Most existing regularized methods rely on the sparsity assumption that a model with fewer parameters predicts better than one with many parameters. This assumption works particularly well in high-dimensional problems. However, the sparsity assumption may not be necessary when the number of predictors is relatively small compared to the number of training instances. This paper argues that shrinking the coefficients towards a low-variance data-driven estimate could be a better regularization strategy for such situations. For classification problems, we propose a naïve Bayes regularized logistic regression (NBRLR) that shrinks the logistic regression coefficients toward the naïve Bayes estimate to provide a reduction in variance. Our approach is primarily motivated by the fact that naïve Bayes is functionally equivalent to logistic regression if naïve Bayes' conditional independence assumption holds. Under standard conditions, we prove the consistency of the NBRLR estimator. Extensive simulation and empirical experimental results show that NBRLR is a competitive alternative to various state-of-the-art classifiers.

Keywords:
Classification, Regularization Method, Logistic Regression, Naïve Bayes, Data-driven Shrinkage

1. Introduction

Logistic regression (LR) is widely used in machine learning for classification problems. It is a discriminative classifier that directly learns the class variable's conditional probability given the predictors without assuming anything about the distribution of the predictors. As per Ng and Jordan (2002), LR converges to the best

linear classifier when the training sample size, n, goes to infinity by producing the smallest bias and therefore is highly preferred amongst linear classifiers when the training sample size is large. However, when the training sample is limited, or there is a large number of parameters, p, to be estimated, regularization is required to avoid overfitting. Many regularized methods have been proposed to improve prediction error in regression frameworks, including lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and LARS (Efron et al., 2004). These estimators rely largely on the sparsity assumption, i.e., only a small proportion of predictors are likely to be informative. Thus, they work particularly well in high-dimensional problems, i.e., p is relatively large compared to n.

A good regularization strategy should be shrinking the regression coefficients towards the values close to the truth. One limitation of these approaches is that, in practice, the sparsity assumption is often violated. Especially when p is relatively small compared to n, predictors are less likely to be irrelevant to the class variable and thus tend to be influential. Shrinking the coefficients of influential predictors towards zero introduces bias and causes the regression estimates to be suboptimal. As a result, traditional sparsity-enforced approaches may not perform well. Also, in this scenario, there tends to be less multicollinearity among predictors. This limits the benefit of ridge regression (Hoerl and Kennard, 1970), which is motivated by dealing with multicollinearity, not sparsity. We argue that when p is relatively small compared to n, a better regularization strategy is to shrink the coefficients towards a low-variance data-driven estimate.

It has been shown that naïve Bayes (NB), a probabilistic classifier with an equivalent functional form compared to LR, tends to have a lower variance than LR (Zaidi et al., 2013, 2014). NB is a generative classifier that learns the predictors' and class variables' joint probability distribution. It infers the posterior probability of a class label given data using Bayes rules, assuming that the predictors are mutually conditionally independent of each other given the class variable. This assumption is mostly motivated by the need to learn a smaller number of parameters from high-dimensional data to overcome overfitting. Consequently, NB performs surprisingly well, even against other more sophisticated classifiers, especially when the training set size is small (Domingos and Pazzani, 1996; Hand and Yu, 2001).

In this paper, we propose a naïve Bayes regularized logistic regression model (NBRLR) for classification problems, which uses regularization to shrink the estimates of an LR model towards the NB estimate. As LR and NB form a well-known discriminative-generative pair, our work adds to the literature exploring hybrid models that take advantage of both approaches. Such models can be placed into two categories. The first category comprises two-stage approaches, which train the model

generatively with the NB model in one stage while training the model discriminatively with the LR model in the other stage. Raina et al. (2003) and Fujino et al. (2007) investigate a class of hybrid model for supervised learning in the context of text classification problems that are partly generative and partly discriminative. Specifically, they allow different partitions of the predictors into subgroups, each modeled under the NB assumption based on domain knowledge. Then these sub-generative models are combined with weight parameters that are determined discriminatively. Our study differs because we do not require prior domain knowledge to fit a model.

Kang and Tian (2006) introduce a restricted class of Bayesian network classifiers using LR as the discriminative component and NB as the generative component. Tan and Shenoy (2020) examine the construction of such hybrid models, i.e., to decide whether a given predictor should be assigned to the LR part or the NB part. Specifically, they develop a metric to compare models, which uses conditional independence as a proxy for model bias and training sample size as a proxy for variance. The weakness of this method is that it serves as a selection mechanism; a predictor is classified as an NB or LR predictor with no middle ground. Our proposed method is a shrinkage approach, which is more stable to small perturbations of data changes, and may improve the prediction accuracy.

Our work belongs to the second category, which uses the maximum likelihood parameterization of NB to pre-condition the discriminative search of LR. Zaidi et al. (2013, 2014) discuss a weighted variant of NB with predictor weights selected by minimizing either the negative conditional log-likelihood or the mean squared error rather than based on measures of predictiveness. Their strategy can also be viewed as using weights to alleviate the predictor independence assumption of NB. The resulting weighted NB model is equivalent to LR but computationally much more efficient. Zaidi et al. (2016) introduces *accelerated logistic regression* for training LR with high-order predictors. The proposed method significantly improves efficiency and reduces LR's bias, making it particularly useful for large datasets. In these papers, authors search for the optimal feature weights of the weighted NB by maximizing discriminative scores. Our work differs in that we estimate the LR coefficients by a penalized likelihood, with coefficients being shrunk towards the NB estimates.

An outline of the remainder of the paper is as follows. Section 2 compares the LR and the NB models and describes our method for cases where the predictors can be either categorical or continuous. We also provide theoretical results, including the consistency of our estimator. Section 3 presents the coordinate descent algorithm we use. Section 4 includes simulation results to show how our estimator performance is affected by the number of training instances, the number of predictors, and the dependence among predictors, under four simulation settings. Section 5 provides em-

pirical results from experiments using thirty datasets from the UCI Machine Learning Repository or Kaggle. Finally, in Section 6, we summarize and conclude.

2. Naïve Bayes Penalized Logistic Regression

2.1. Logistic Regression

In this study, we consider the independent and identically distributed samples $\{y_i, \mathbf{x}_i\}_{i=1}^n$, with $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$ and $y_i \in \{0, 1\}$. We consider cases where the predictors can be either categorical or continuous. We assume that the categorical predictors are all binary for the simplicity of exposition. The proposed method can be easily extended to non-binary cases by recoding the categorical predictors with dummies.

LR is a discriminative classifier that directly learns the conditional probability $P(y_i \mid \mathbf{x}_i)$ by assuming the form:

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}^*)}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}^*)}. \tag{1}$$

where $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^T \in \mathbb{R}^{p+1}$ with β_0^* being the intercept, and β_j^* being the coefficient corresponding the the j^{th} predictor. In LR, $\boldsymbol{\beta}^*$ is estimated by maximizing the conditional likelihood as:

$$\hat{\boldsymbol{\beta}}_{LR} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\max} \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}).$$

LR is a well-known low-bias, high-variance estimator. As shown by Ng and Jordan (2002), LR is the best linear classifier asymptotically. However, LR estimates can have very large variances when the sample size is relatively small compared to the number of predictors. In the cases of perfect fits, they can be infinitely large. One advantage of regularization techniques, for example lasso (Tibshirani, 1996), is they increase the stability of the estimates.

The lasso estimator for logistic regression is defined as

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \left[-\frac{1}{n} l\left(\boldsymbol{\beta}\right) + \frac{\lambda}{n} \sum_{j=1}^{p} |\beta_j| \right],$$

where $l\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i' \boldsymbol{\beta} - \log\left[1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})\right]\}$ is the log-likelihood function. Lasso is particularly useful in high-dimensional situations by assuming sparsity of the true $\boldsymbol{\beta}^*$. This assumption is primarily driven by the "*bet on sparsity*" principle (Hastie et al., 2001):

"*Use a procedure that does well in sparse problems since no procedure does well in dense problems.*"

Our intuition is that a good regularization strategy is to shrink the regression coefficient toward the values close to the truth. When p is relatively small compared to n, the sparsity assumption may not be necessary. Specifically, the predictors are less likely to be irrelevant to the class variable and tend to be influential. Shrinking the coefficients of influential predictors towards zero introduces bias and makes the regression estimates suboptimal. Unlike lasso, ridge regression is another common regularization method that does not assume sparsity. Ridge estimators reduce the variance caused by correlated predictors but at the cost of introducing bias to the estimator. However, such sacrifice may not be worth it when p is relatively small compared to n, as there tends to be less multicollinearity among predictors. This paper proposes a model-based approach for balancing the bias-variance tradeoff by shrinking $\boldsymbol{\beta}$ towards the NB estimate instead of zero. In some settings, NB can be preferred to LR because of the low variance in NB estimates. In the following, we present the equivalent functional forms of LR and NB, which, along with the small variance in NB estimates, motivates our decision to shrink LR coefficients towards the NB estimates.

2.2. Logistic Regression versus Naïve Bayes

Naïve Bayes (NB) is a simple and effective supervised classification model based on applying Bayes' rule with the strong assumption of conditional independence, i.e., predictors are conditionally independent of each other given the class variable. Using Bayes rule, NB can be expressed as:

$$P(y_i = 1|\mathbf{x}_i) = \frac{P(y_i = 1)\prod_{j=1}^{p} P(x_{ij}|y_i = 1)}{\sum_{\tilde{y}=0}^{1} P(y_i = \tilde{y})\prod_{j=1}^{p} P(x_{ij}|y_i = \tilde{y})}. \tag{2}$$

The conditional independence assumption reduces the complexity of an NB model. Therefore NB exhibits low variance and performs surprisingly well when the training set size is small (Domingos and Pazzani, 1996; Hand and Yu, 2001). However, the conditional independence assumption rarely holds in practice. Any violation of the assumption will result in a bias, making NB estimates suboptimal. Accordingly, NB is a low-variance high-bias classifier compared to LR (Ng and Jordan, 2002).

Many papers have explored hybrid models that combine LR and NB into one model to take advantage of both approaches. One category of methods is to fit the model in two stages, a generative stage where we fit an NB model and a discriminative stage where we fit an LR model. Methods in the second category use

the maximum likelihood parameterization of NB to pre-condition the discriminative search of LR (Zaidi et al., 2013, 2014, 2016). Our method belongs to the second category. Instead of searching for the optimal feature weights of the weighted NB by maximizing discriminative scores, we estimate the LR coefficients by a penalized likelihood with coefficients being shrunk towards the NB estimates.

2.3. Naïve Bayes Regularized Logistic Regression

Our approach is primarily motivated by the fact that LR and NB converge toward the identical classifier assuming that NB's conditional independence assumption holds. Specifically, we rewrite the parametric form of $P(y_i = 1|\mathbf{x}_i)$ of NB as

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= \frac{P(y_i = 1)\prod_{j=1}^{p} P(x_{ij}|y_i = 1)}{\sum_{y=0}^{1} P(y_i = y)\prod_{j=1}^{p} P(x_{ij}|y_i = y)} \\ &= \frac{\exp\left[\log\frac{P(y_i=1)}{P(y_i=0)} + \sum_{j=1}^{p}\log\frac{P(x_{ij}|y_i=1)}{P(x_{ij}|y_i=0)}\right]}{1+\exp\left[\log\frac{P(y_i=1)}{P(y_i=0)} + \sum_{j=1}^{p}\log\frac{P(x_{ij}|y_i=1)}{P(x_{ij}|y_i=0)}\right]}. \end{aligned} \tag{3}$$

When a predictor X_j is binary, we have $P(x_{ij}|y_i = \tilde{y}) = P(x_{ij} = 1|y_i = \tilde{y})^{x_{ij}} \cdot P(x_{ij} = 0|y_i = \tilde{y})^{1-x_{ij}}$. Define $G_{ij}(a) = P(x_{ij} = 0|y_i = a)$, the summation term can be expanded as

$$\begin{aligned} &\log\frac{P(x_{ij}|y_i = 1)}{P(x_{ij}|y_i = 0)} = \\ &\log\frac{G_{ij}(1)}{G_{ij}(0)} + \log\frac{P(x_{ij} = 1|y_i = 1)/G_{ij}(1)}{P(x_{ij} = 1|y_i = 0)/G_{ij}(0)}x_{ij}. \end{aligned}$$

When a predictor X_j is continuous, $P(x_{ij}|y_i)$ represents the corresponding conditional density of X_j. One common assumption for each continuous predictor X_j of a NB model is that, $x_{ij} \mid y_i = 1 \sim N(u_{j1}, \sigma_j^2)$ and $x_{ij} \mid y_i = 0 \sim N(u_{j0}, \sigma_j^2)$ (Mitchell, 1997). Note that the standard deviations σ_j varies from predictor to predictor but does not depend on the value of y_i. Then, the summation term can be expanded as

$$\log\frac{P(x_{ij}|y_i = 1)}{P(x_{ij}|y_i = 0)} = \frac{u_{j0}^2 - u_{j1}^2}{2\sigma_j^2} + \frac{u_{j1} - u_{j0}}{\sigma_j^2}x_{ij}.$$

Suppose the predictor x_{ij} is categorical for $j \in \{1, \ldots, q\}$, and continuous for $j \in \{q+1, \ldots, p\}$. We get the direct equivalence between LR and NB by substituting the two expressions above back into Eq. (3):

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(z_i)}{1+\exp(z_i)}, \tag{4}$$

where $z_i = \log \frac{P(y_i=1)}{P(y_i=0)} + \sum_{j=1}^{q} \log \frac{G_{ij}(1)}{G_{ij}(0)} + \sum_{j=q+1}^{p} \frac{u_{j0}^2 - u_{j1}^2}{2\sigma_j^2} + \sum_{j=1}^{q} \log \frac{P(x_{ij}=1|y_i=1)/G_{ij}(1)}{P(x_{ij}=1|y_i=0)/G_{ij}(0)} x_{ij} + \sum_{j=q+1}^{p} \frac{u_{j1}-u_{j0}}{\sigma_j^2} x_{ij}$.

Next, we define the NBRLR model for classification problems. Following the conventional regularization methods set-up, we assume that $\{x_{ij}\}_{i=1}^{n}$ are standardized so that $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for all j. Let $\{\hat{\eta}_0, \hat{\eta}_j\}$ denote the naïve Bayes estimate of the model defined as

$$\begin{aligned}
\hat{\eta}_0 &= \log \tfrac{\hat{P}(y_i=1)}{\hat{P}(y_i=0)} + \textstyle\sum_{j=1}^{q} \log \frac{\hat{G}_{ij}(1)}{\hat{G}_{ij}(0)} + \sum_{j=q+1}^{p} \frac{\hat{u}_{j0}^2 - \hat{u}_{j1}^2}{2\hat{\sigma}_j^2}, \\
\hat{\eta}_j &= \log \tfrac{\hat{P}(x_{ij}=1|y_i=1)/\hat{G}_{ij}(1)}{\hat{P}(x_{ij}=1|y_i=0)/\hat{G}_{ij}(0)} x_{ij} \quad for \quad j \in \{1, \ldots, q\}, \\
\hat{\eta}_j &= \tfrac{\hat{u}_{j1}-\hat{u}_{j0}}{\hat{\sigma}_j^2} \qquad\qquad\qquad for \quad j \in \{q+1, \ldots, p\}.
\end{aligned}$$

where $\hat{P}()$, $\hat{G}()$, $\hat{\rho}$, $\hat{u}_{j0}$, $\hat{u}_{j1}$ and $\hat{\sigma}_j$ are the corresponding fitted parameter values from training data.

The NBRLR estimator $\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}}$ is defined by

$$\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} - \frac{1}{n} l(\boldsymbol{\beta}) + \frac{\lambda}{n} \sum_{j=0}^{p} |\beta_j - \hat{\eta}_j|, \tag{5}$$

where $\lambda \geq 0$ is the tuning parameter that controls the amount of regularization. For $\lambda = 0$, $\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}}$ is equivalent to the LR estimate. On the contrary, for a sufficiently large value of λ, $\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}}$ will provide predicted probabilities that are the same as NB. Note that the NB probabilities depend on the value of its intercept. To regularize the LR model with an NB estimate, we penalize the intercept term, which is uncommon in traditional regularization techniques, such as lasso and ridge regression.

In practice, when p is relatively small compared to n, the sparsity assumption is more likely to be violated. However, NB's assumption of conditional independence is different. The smaller number of predictors raises the chance of satisfying the conditional independence assumption among features, which makes the NB estimates more reliable. In these settings, shrinking coefficients toward the NB estimates instead of zero will produce less bias while reducing the variance compared to LR. Although, the proposed NBRLR estimator will have a larger variance than lasso because no coefficient will be set to zero, and $\hat{\eta}_j$ is an estimate for all $j \in \{0, \ldots, p\}$. This issue will be more problematic for larger p's. However, the proposed method can outperform lasso when the predictors are informative and the number of predictors is small relative to the sample size.

2.4. Asymptotic Results

This subsection will provide a consistent result for the NBRLR estimator. The result is a general result that will apply to shrinking towards any values and does depend on some conditions. We allow categorical and continuous predictors as in Section 2.3. Let $\hat{\eta}_j$ represent the, potentially estimated, value the jth coefficient is being shrunk towards and $\hat{\boldsymbol{\eta}} = (\hat{\eta}_0, \ldots, \hat{\eta}_p)^T \in \mathbb{R}^{p+1}$. Suppose,

$$Z_n(\beta, \hat{\boldsymbol{\eta}}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i x_i'\beta - \log\left[1 + \exp(x_i'\beta)\right]) + \frac{\lambda}{n}\sum_{j=0}^{p}|\beta_j - \hat{\eta}_j| \tag{6}$$

and

$$\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} Z_n(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}). \tag{7}$$

Define $\psi_{jkm}(\boldsymbol{\beta}) = \frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_m} l(\boldsymbol{\beta})$. To prove $\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}}$ is a consistent estimator of $\boldsymbol{\beta}^*$, as defined in Eq. (1), we require the following conditions.

Condition 1. *The number of predictors must satisfy* $p = o\left(n^{1/4}\right)$. *There exist positive constant* b_1,b_2,b_3 *and* b_4 *such that for any vector* $\mathbf{a} \in \mathbb{R}^{p+1}$,

$$b_1\|\mathbf{a}\|_2^2 \leq \mathbf{a}^T\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{a} \leq b_2\|\mathbf{a}\|_2^2$$

and

$$b_3\|\mathbf{a}\|_2^2 \leq \mathbf{a}^T\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}^*)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}^*)\right]^2}\mathbf{a} \leq b_4\|\mathbf{a}\|_2^2.$$

In addition $\max_{1\leq i\leq n}\frac{1}{n}||\mathbf{x}_i||_2^2 \to 0$.

Condition 2. *For some positive constant* C_1, *suppose* $B_n = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 \leq C_1\sqrt{p/n}\}$. *There exist positive constants* C_2 *and* C_3 *such that for all* $\boldsymbol{\beta} \in B_n$, $j \in \{0, \ldots, p\}$, $k \in \{0, \ldots, p\}$ *and* $m \in \{0, \ldots, p\}$ *that*

$$C_2 < |\psi_{jkm}(\boldsymbol{\beta})| < C_3.$$

Condition 1 is a generalization of assumptions made in Knight and Fu (2000) for fixed p least squares regression with lasso. Condition 2 ensures that when $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}^*$, $l(\boldsymbol{\beta})$ can be well approximated by a second-order Taylor expansion. Similar conditions have been made on the third partial derivative of a likelihood when analyzing the asymptotic behavior of a penalized likelihood method (Fan and Peng, 2004; Kwon and Kim, 2012).

Theorem 1. *Assume that Eq.* (1) *and Conditions 1 and 2 hold and that* $\lambda = O(\sqrt{n})$ *then* $||\hat{\boldsymbol{\beta}}_{\lambda,\hat{\boldsymbol{\eta}}} - \boldsymbol{\beta}^*|| = O_P(\sqrt{p/n})$.

Proof. By the properties of convex functions, for more details, see the proof of Theorem 2.1 in He and Shi (1994) and Corollary 25, p.47, of Eggleston (1958), it is sufficient to show that there exists L such that

$$P\left[\inf_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2=L\sqrt{p/n}} Z_n(\boldsymbol{\beta},\hat{\boldsymbol{\eta}}) - Z_n(\boldsymbol{\beta}^*,\hat{\boldsymbol{\eta}}) > 0\right] \to 1. \tag{8}$$

By Taylor's approximation, for any $\boldsymbol{\beta}$ there is $\tilde{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ such that

$$\begin{aligned} l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}^*) &= (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\left[\frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{\beta}^*)\right] \\ &+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\left[\frac{\partial^2}{\partial\boldsymbol{\beta}^2} l(\boldsymbol{\beta}^*)\right](\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\ &+\sum_{j=0}^{p}\sum_{k=0}^{p}\sum_{m=0}^{p}(\beta_j-\beta_j^*)(\beta_k-\beta_k^*)(\beta_m-\beta_m^*)\psi_{jkm}(\tilde{\boldsymbol{\beta}}). \end{aligned}$$

Notice,

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{\beta}) &= -\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left[y_i - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right], \\ \frac{\partial^2}{\partial\boldsymbol{\beta}^2} l(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{[1+\exp(\mathbf{x}_i'\boldsymbol{\beta})]^2}. \end{aligned}$$

For any $j \in \{0,\ldots,p\}$, by Condition 1 and the Lindeberg-Feller CLT, $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\frac{1}{n}\sum_{i=1}^{n}x_{ij}\left[y_i - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right] = O_P(||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 n^{-1/2})$ and thus $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left[y_i - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right] = O_P(||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2\sqrt{p/n})$. Therefore

$$\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2=L\sqrt{p/n}}\left|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\left[\frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{\beta}^*)\right]\right| = O_P(pn^{-1}L).$$

In addition by Condition 1

$$\begin{aligned} \inf_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2=L\sqrt{p/n}}&\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\left[\frac{\partial^2}{\partial\boldsymbol{\beta}^2} l(\boldsymbol{\beta}^*)\right](\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\ &\geq b_3||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2 = b_3L^2pn^{-1}. \end{aligned}$$

By Condition 2

$$\begin{aligned}
&\sum_{j=0}^{p}\sum_{k=0}^{p}\sum_{m=0}^{p}(\beta_j-\beta_j^*)(\beta_k-\beta_k^*)(\beta_m-\beta_m^*)\psi_{jkm}(\tilde{\boldsymbol{\beta}})\\
&\leq C_3\sum_{j=0}^{p}|\beta_j-\beta_j^*|\sum_{k=0}^{p}|\beta_k-\beta_k^*|\sum_{m=0}^{p}|\beta_m-\beta_m^*|\\
&\leq C_3 p^{3/2}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^3,
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{j=0}^{p}\sum_{k=0}^{p}\sum_{m=0}^{p}(\beta_j-\beta_j^*)(\beta_k-\beta_k^*)(\beta_m-\beta_m^*)\psi_{jkm}(\tilde{\boldsymbol{\beta}})\\
&\geq -C_3\sum_{j=0}^{p}|\beta_j-\beta_j^*|\sum_{k=0}^{p}|\beta_k-\beta_k^*|\sum_{m=0}^{p}|\beta_m-\beta_m^*|\\
&\geq -C_3 p^{3/2}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^3.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2=L\sqrt{p/n}}\left|\sum_{j=0}^{p}\sum_{k=0}^{p}\sum_{m=0}^{p}(\beta_j-\beta_j^*)(\beta_k-\beta_k^*)(\beta_m-\beta_m^*)\psi_{jkm}(\tilde{\boldsymbol{\beta}})\right|\\
&\quad = O_P(L^3p^3n^{-3/2}).
\end{aligned}$$

Note, under assumption that $p = o\left(n^{1/4}\right)$, from Condition 1, that $p^3n^{-3/2} = o(pn^{-1})$. Finally, under the assumption that $\lambda = O(\sqrt{n})$,

$$\begin{aligned}
\frac{\lambda}{n}\sum_{j=0}^{p}|\beta_j-\hat{\eta}_j|-|\beta_j^*-\hat{\eta}_j|&\leq \frac{\lambda}{n}\sum_{j=0}^{p}|\beta_j-\beta_j^*|\\
&\leq \frac{\lambda}{n}\sqrt{p}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2= O(pLn^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
\frac{\lambda}{n}\sum_{j=0}^{p}|\beta_j-\hat{\eta}_j|-|\beta_j^*-\hat{\eta}_j|&\geq -\frac{\lambda}{n}\sum_{j=0}^{p}|\beta_j-\beta_j^*|\\
&\geq -\frac{\lambda}{n}\sqrt{p}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2= O(pLn^{-1}).
\end{aligned}$$

Therefore, for sufficiently large L, the lower bound of the quadratic term will dominate the other terms, and (8) holds.

□

Theorem 1 proves that the convergence rate of the proposed estimator is asymptotically equivalent to logistic regression, assuming that λ gets smaller as n increases. It allows p to increase with n, suggesting that the proposed method can work well even with many predictors.

3. Algorithm

We consider a coordinate descent step for solving Eq. (7), following Friedman et al. (2010). The unpenalized log-likelihood $l(\boldsymbol{\beta})$ is maximized by implementing Newton's method with the iteratively reweighted least square algorithm. Specifically, given the current estimates of the parameters $\hat{\boldsymbol{\beta}}^{old} = (\hat{\beta}_0^{old}, \hat{\beta}_1^{old}, \ldots, \hat{\beta}_p^{old})^T$ with corresponding probability $\hat{p}^{old}(\mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}}^{old})$ for observation i, we obtain a quadratic approximation to the $l(\boldsymbol{\beta})$ as:

$$l_Q(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n} w_i \left\{z_i - \mathbf{x}_i^T\boldsymbol{\beta}\right\}^2 + C, \tag{9}$$

where

$$\begin{aligned} z_i &= \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{old} + \frac{y_i - \hat{p}^{old}(\mathbf{x}_i)}{\hat{p}^{old}(\mathbf{x}_i)\left[1 - \hat{p}^{old}(\mathbf{x}_i)\right]}, \\ w_i &= \hat{p}^{old}(\mathbf{x}_i)\left[1 - \hat{p}^{old}(\mathbf{x}_i)\right], \end{aligned}$$

and C is a constant term. Then, our task becomes minimizing the following penalized weighted least-squares problem

$$-\frac{1}{n}l_Q(\boldsymbol{\beta}) + \frac{\lambda}{n}\sum_{j=0}^{p}|\beta_j - \hat{\eta}_j|. \tag{10}$$

Define soft-thresholding operator $S(a, b) = sign(a)(|a|-b)_+$, the update of coordinate descent is performed by

$$\beta_j^{new} \leftarrow \hat{\eta}_j + \frac{S(A - \hat{\eta}_j B, \lambda)}{B}, \tag{11}$$

where

$$
\begin{aligned}
A &= \sum_{i=1}^{n} w_i x_{ij} z_i, \\
B &= \sum_{i=1}^{n} w_i x_{ij}^2.
\end{aligned}
$$

Thus, Eq. (10) is minimized by iterating through $j \in \{0, 1, \ldots, p\}$ until its difference between two iterations is less than 10^{-7}.

We propose the following algorithm with a fixed value of λ.

1. Begin with initial estimates of $\hat{\boldsymbol{\beta}}^0 = \{\hat{\beta}_0^0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\}$.
2. For the t^{th} step, where $t \geq 1$, repeat the steps below until the difference of the penalized log-likelihood Eq. (5) between $(t-1)^{th}$ and t^{th} step is less than 10^{-7}.
   (a) Update the quadratic approximation l_Q with the current parameters $\hat{\beta}^{t-1}$.
   (b) Given current l_Q, the t^{th} iterative estimate of $\boldsymbol{\beta}$ is:

$$
\hat{\boldsymbol{\beta}}^t = \arg\min_{\beta} -\frac{1}{n} l_Q(\beta) + \frac{\lambda}{n} \sum_{j=0}^{p} |\beta_j - \hat{\eta}_j|,
$$

   where it can be solved following the coordinate descent solution from Eq. (11) using $\hat{\beta}^{t-1}$ as the current estimate $\hat{\beta}^{old}$.

An R package implementing the described algorithm will be made publicly available upon acceptance of the publication of this work.

4. Simulations

This section compares the NBRLR with pure LR, pure NB, regularized LR (lasso, ridge, and elastic net), and hybrid LR-NB (Tan and Shenoy, 2020). Lasso, ridge, and elastic net are fit using the glmnet package (Friedman et al., 2010) in R. Tuning parameters for NBRLR, lasso, ridge, elastic net, and hybrid LR-NB are determined using 10-fold cross-validation to minimize the out-of-sample prediction error. When the predictors are categorical, the parameters of the pure NB model are estimated using the Laplace correction (Niblett, 1987) to prevent the high influence of zero probabilities. Specifically, we add one of each class to the data. The Laplace corrected values are used for the NBRLR estimator and the hybrid LR-NB model.

The prediction performance of the models is evaluated on 1000 testing instances using the average prediction 0-1 loss ($L_{0\text{-}1}$) and root squared prediction error ($RSPE$).

Let N_{te} (= 1000) be the testing sample size, y_i be the observed class for the ith testing observation, $\hat{P}(y_i = 1)$ be the predicted probability that testing observation i belongs to class 1, and $\hat{y}_i$ be the predicted class of that observation. Specifically, if the predicted probability of an observation is below .5, we predict that the sample belongs to class 0. Otherwise, we predict it belongs to class 1. Let $\mathbf{1}()$ be an indicator function for the condition in the parenthesis. $L_{0\text{-}1}$ and $RSPE$ are defined as

$$L_{0\text{-}1} = 1 - \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \mathbf{1}(y_i = \hat{y}_i),$$

and

$$RSPE = \sqrt{\frac{1}{2 \cdot N_{te}} \sum_{i=1}^{N_{te}} \{ \sum_{\tilde{y}=0}^{1} (\hat{P}(y_i = \tilde{y}) - \mathbf{1}(y_i = \tilde{y}))^2 \}}.$$

For NBRLR, pure LR, pure NB, lasso, ridge, and elastic net, we also report the mean squared error of the estimator $\hat{\boldsymbol{\beta}}$, $MSE(\hat{\boldsymbol{\beta}})$, which for the truth $\boldsymbol{\beta}^*$ is defined as

$$MSE(\hat{\boldsymbol{\beta}}) = \frac{1}{p+1} ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2.$$

This metric is not applicable for hybrid LR-NB because it cannot be expressed in the regression form as Eq. (1).

Four simulation settings are considered in this study. First, we consider generating data with categorical predictors from a discriminative LR model. Second, we consider generating data with categorical predictors from a generative NB model. Third, we generate data with continuous predictors from a discriminative LR model. Fourth, we generate data using categorical and continuous predictors from a discriminative LR model. We put the results of simulations 3 & 4 in the supplemental material because they lead to similar conclusions compared to simulations 1 & 2. In addition, we vary the number of training instances (in simulations 1 & 2), the number of predictors, and the dependence among predictors (conditional dependence for the second simulation setting) to see how these factors affect the models' performance. We repeat the entire procedure 100 times. Boxplots are presented to compare the $L_{0\text{-}1}$, $RSPE$ and $MSE(\hat{\boldsymbol{\beta}})$ of the models. Further, we compare the results of NBRLR to the other benchmark methods by reporting the averages of the three metrics, performing two-tailed, paired t-tests, and reporting the corresponding p-values. Note that the hybrid LR-NB model is only included in the comparison under the first two simulation settings as it was proposed for categorical predictors.

4.1. Simulation Setting 1: Discriminative LR

In the first simulation, p categorical variables $\mathbf{x}_i = \{x_{i1}, \dots x_{ip}\}$ are simulated by first generating $\tilde{\mathbf{x}}_i \sim N(0_p, \Sigma_{p\times p})$, where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = r$ for $j \neq k$ and then x_{ij} is dichotomized as 0 if $\tilde{x}_{ij}$ is smaller than 0, and 1 otherwise. The class variable y_i is then simulated from

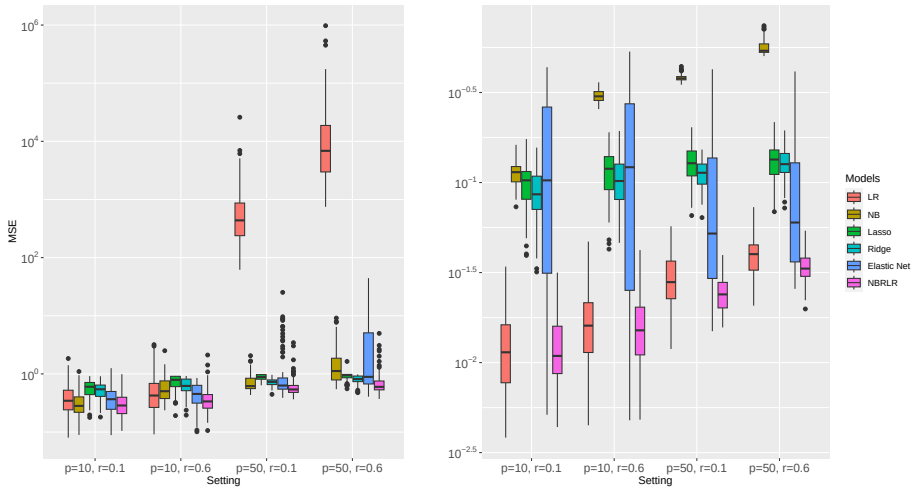
$$P(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$

where $\boldsymbol{\beta} = (0, -1_{p/2}^T, 1_{p/2}^T)^T$. We generate N_{tr} training samples to fit the models and $N_{te} = 1000$ test samples to assess their prediction performance. Eight situations in which $N_{tr} \in \{100, 2000\}$, $p \in \{10, 50\}$ and $r \in \{0.1, 0.6\}$ are considered.

Comparisons of the estimators in terms of $MSE(\hat{\boldsymbol{\beta}})$, $L_{0\text{-}1}$ and $RSPE$ across all the combinations of N_{tr}, p and r for this setting are reported in Figures 1 - 3, respectively. Table 1 provides the averages of the three metrics across the eight different combinations of N_{tr}, p, and r, and includes the p-values from two-tailed, paired t-tests comparing the performance of NBRLR to the other benchmark methods. Notice that when $N_{tr} = 100$ and $p = 50$, $MSE(\hat{\boldsymbol{\beta}})$ for LR is excessively large, which may be due to the estimation convergence failures. As a result in this section, all the boxplots of MSE use a log-10 scale for the Y axis. The results show that our proposed NBRLR estimator always performs the best with respect to MSE. However, the ridge regression is competitive with or outperforms NBRLR in both $L_{0\text{-}1}$ and $RSPE$, especially when $N_{tr} = 100$. This may be due to ridge regression's extraordinary ability to eliminate multicollinearity under the discriminative setting. Besides, LR does better than or as well as NBRLR in $L_{0\text{-}1}$ on the either large size ($N_{tr} = 2000$) or low dimensional ($p = 10$) datasets. This is because LR is optimal for large values of N_{tr} compared to p, as the MLE of it is asymptotically unbiased and has the smallest variance. Also, notice that the difference in the prediction performance for NBRLR compared to LR and regularized LR is very marginal when the training set size is large ($N_{tr} = 2000$). This supports our theorem of convergence that the proposed NBRLR and standard LR behave asymptotically equivalently.

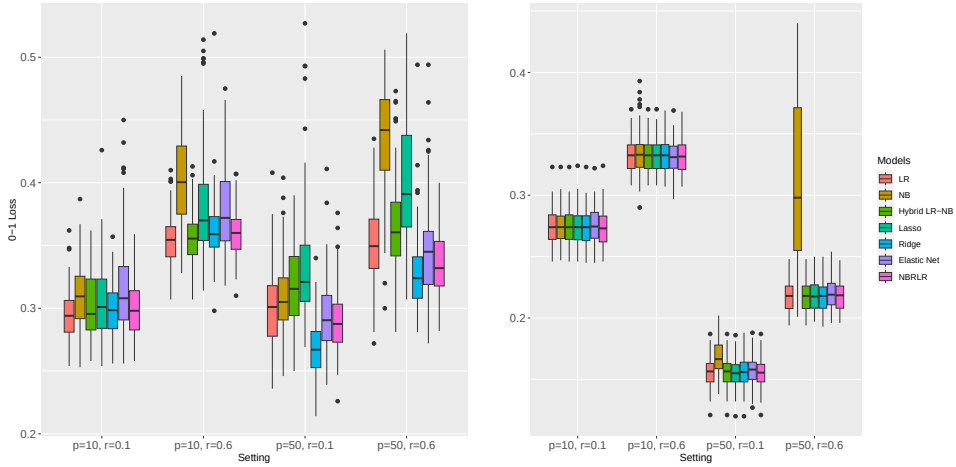
4.2. Simulation Setting 2: Generative NB

In the second simulation, we start by generating the class variable y_i from $y_i \sim Bern(0.5)$. Then, p categorical variables $\mathbf{x}_i = \{x_{i1}, \dots x_{ip}\}$ are simulated in a two step process. First, $\tilde{\mathbf{x}}_i \mid y_i = 1 \sim N(\mathbf{u}_1, \Sigma_{p\times p})$ where $\mathbf{u}_1 = \{0.2_{p/2}, -0.2_{p/2}\}$, and $\tilde{\mathbf{x}}_i \mid y_i = 0 \sim N(\mathbf{u}_0, \Sigma_{p\times p})$ where $\mathbf{u}_0 = \{-0.2_{p/2}, 0.2_{p/2}\}$. For both distributions $\Sigma_{jj} = 1$ and $\Sigma_{jk} = r$ if $j \neq k$. Define $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$ as a vector of ones and zeros, where x_{ij} is zero if $\tilde{x}_{ij}$ is smaller than 0 and x_{ij} is one otherwise. To get



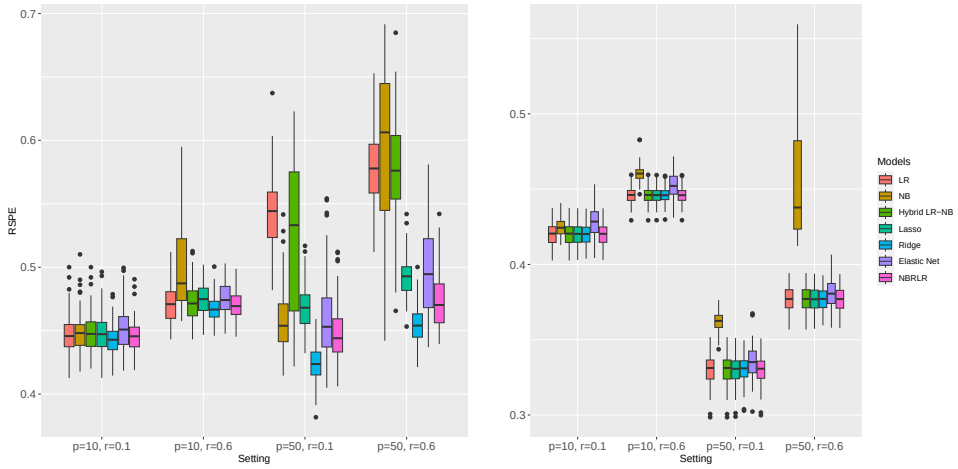
(a) $N_{tr} = 100$ (b) $N_{tr} = 2000$

Figure 1: MSE results for simulation setting 1. The x-axis includes the four different combinations of p and r. The plot does not include hybrid LR-NB because it cannot be expressed in the regression form as Eq.(1).



(a) $N_{tr} = 100$ (b) $N_{tr} = 2000$

Figure 2: $L_{0\text{-}1}$ results for simulation setting 1. The x-axis includes the four different combinations of p and r.



(a) $N_{tr} = 100$ (b) $N_{tr} = 2000$

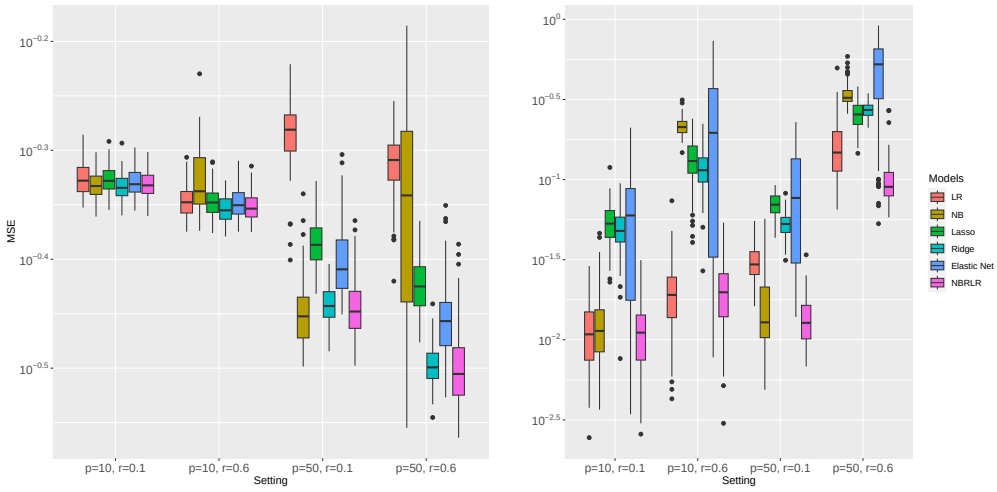
Figure 3: $RSPE$ results for simulation setting 1. The x-axis includes the four different combinations of p and r.

the value of $\boldsymbol{\beta}^*$, we generate 500,000 training samples, fit an LR model and treat the corresponding coefficients as $\boldsymbol{\beta}^*$.

Comparisons of the estimators in terms of $MSE(\hat{\boldsymbol{\beta}})$, $L_{0\text{-}1}$ and $RSPE$ for different values of N_{tr}, p and r are reported in Figures 4 - 6, respectively. Table 2 is the equivalent of Table 1 but for simulation setting 2. It is worth noting that when r = 0.1, NB is highly competitive with NBRLR, especially when $N_{tr} = 100$, because the conditional independence assumption is only weakly violated. Besides, NBRLR does worse than ridge in $RSPE$ when $N_{tr} = 100$ and $p = 10$; however, NBRLR performs the best or among the best in the rest of the settings. Similar to simulation 1, the difference between NBRLR versus (regularized) LR regarding the prediction performance is less significant when $N_{tr} = 2000$.

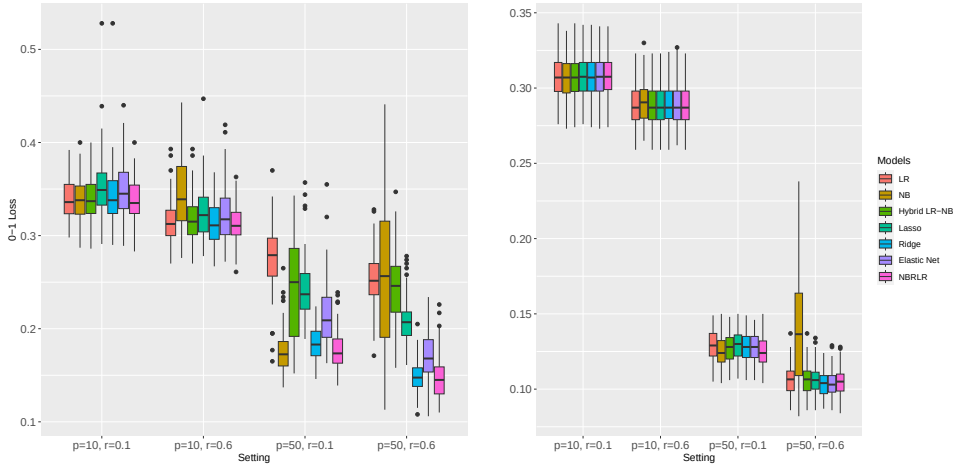
4.3. Bias and Variance Analysis

Our proposed method follows the traditional bias-variance tradeoff strategy. To provide valuable insight into the components of the error of the classifiers, we discuss the squared bias and variance of LR, NB, lasso, ridge, elastic net, and NBRLR. Similarly, as for $MSE(\hat{\boldsymbol{\beta}})$, the hybrid LR-NB is not included in the discussion because it cannot be expressed in the regression form as Eq. (1). Let $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{0,n}, \ldots, \hat{\beta}_{p,n})^T \in \mathbb{R}^{p+1}$ represent an estimator from the nth simulation and $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{100}) \in \mathbb{R}^{(p+1)\times 100}$ represent the 100 estimators for a given method. The squared bias and variance of an estimator for a given simulation setting, with a true coefficient vector



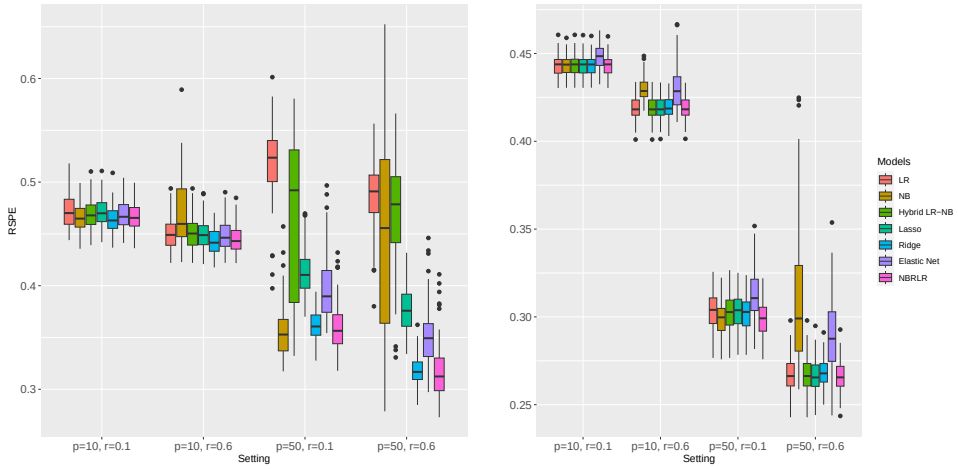
(a) $N_{tr} = 100$ (b) $N_{tr} = 2000$

Figure 4: MSE results for simulation setting 2. The x-axis includes the four different combinations of p and r. The plot does not include hybrid LR-NB because it cannot be expressed in the regression form as Eq.(1).



(a) $N_{tr} = 100$ (b) $N_{tr} = 2000$

Figure 5: $L_{0\text{-}1}$ results for simulation setting 2. The x-axis includes the four different combinations of p and r.



(a) $N_{tr} = 100$ (b) $N_{tr} = 2000$

Figure 6: $RSPE$ results for simulation setting 2. The x-axis includes the four different combinations of p and r.

of $\boldsymbol{\beta}^*$, is

$$Bias^2(\hat{\mathbf{B}}) = \frac{1}{p+1}\sum_{j=0}^{p}\left(\beta_j^* - \frac{1}{100}\sum_{n=1}^{100}\hat{\beta}_{j,n}\right)^2,$$

and

$$Var(\hat{\mathbf{B}}) = \frac{1}{p+1}\cdot\frac{1}{99}\sum_{j=0}^{p}\sum_{n=1}^{100}\left(\hat{\beta}_{j,n} - \frac{1}{100}\sum_{n=1}^{100}\hat{\beta}_{j,n}\right)^2.$$

Tables 3 and 4 present the results $Bias^2(\hat{\mathbf{B}})$ and the $Var(\hat{\mathbf{B}})$, respectively, of the six estimators given simulation 1 and 2. The results are mostly consistent with our intuition. The proposed NBRLR estimator generally has higher variance but lower bias than NB, lasso, and ridge. On the contrary, NBRLR is expected to have a higher bias but lower variance than LR. This holds when $N_{tr} = 2000$. However, when $N_{tr} = 100$, NBRLR does better than LR in bias and variance, which may be due to the convergence failures in LR, especially when p is large. Regarding elastic net, it achieves a lower bias but higher variance than lasso and ridge by applying both L_1 and L_2 regularization. As a result, NBRLR generally outperforms elastic net in bias and variance. Exceptions are that NBRLR has a higher bias in simulation 1 when $N_{tr} = 100$ and $p = 50$, and higher variance in simulation 2 when $N_{tr} = 100$ and $p = 10$. Similar conclusions can be derived from simulations 3 and 4 in the

supplemental material.

5. Empirical Results

In this section, we evaluate the performance of our proposed NBRLR estimator on 30 different machine-learning datasets from either the UCI Machine Learning Repository or Kaggle. Table 5 summarizes these datasets, including the number of predictors, instances, and the predictor type. The datasets are selected such that we have ten datasets with categorical predictors, ten with continuous predictors, and ten with a mix of categorical and continuous predictors. For datasets with missing values, the missing values of categorical predictors are imputed with the conditional probability given the response variable, i.e., $P(x_{ij} \mid y_i)$. For continuous predictors with missing values, we assume they are conditionally normally distributed given the response variable, $x_{ij} \mid y_i = a \sim N(u_{ja}, \sigma^2_{ja})$. The missing values of continuous predictors are imputed with the corresponding conditional distribution given the response variable, i.e., $f(x_{ij} \mid y_i)$. The imputation procedure is conducted before we analyze the data.

The estimators used in Section 4 are tested on each dataset using 20 rounds of 5-fold cross-validation. We compare $L_{0\text{-}1}$ and $RSPE$ of NBRLR to the benchmark methods by reporting Win/Draw/Loss (W/D/L) results as the counted number of datasets for which NBRLR performs numerically better, equally well, or worse on the given metrics. We also conduct two-tailed, paired samples Wilcoxon signed-ranks test (Wilcoxon, 1992) for the comparison of classifiers over multiple datasets and consider the results to be significant if the corresponding p-value is less than 0.05. The Wilcoxon test is a non-parametric test which ranks the differences in performances of two classifiers for each dataset while ignoring the signs, and then compares the ranks for the positive and the negative differences. The detailed results in the averages of $L_{0\text{-}1}$ and $RSPE$ for each estimator on each dataset are presented in the supplemental material. We do not compare the MSE for the empirical datasets as we do not assume to know the true β^* of the predictors. Also, notice that hybrid LR-NB is only included in the comparison for categorical datasets.

5.1. NBRLR versus LR and NB

The Win/Draw/Loss (W/D/L) results of NBRLR against pure LR and pure NB are in Table 6. It can be seen that NBRLR does only slightly better than pure LR in both $L_{0\text{-}1}$ and $RSPE$. However, the improvement of NBRLR over LR becomes more notable on small datasets, especially for $RSPE$. Wilcoxon test suggests that NBRLR significantly outperforms LR at 0.05 level in $RSPE$ on datasets with less

than 300 instances. On the other hand, NBRLR has significantly better $L_{0\text{-}1}$ and $RSPE$ than pure NB over the entire 30 datasets. The corresponding p-value from two-tailed, paired samples Wilcoxon test is smaller than 0.001 for both metrics. Unlike the comparison against LR, the improvement of NBRLR against NB is more significant on large datasets with at least 10,000 instances. LR is a low-bias high-variance estimator, while NB is a low-variance high-bias estimator. These results are consistent with what we expect: our proposed NBRLR estimator will be more like LR, thus outperforming NB for large datasets, and more like NB, thus outperforming LR for smaller datasets.

5.2. NBRLR versus hybrid LR-NB

Comparisons of NBRLR with hybrid LR-NB in terms of $L_{0\text{-}1}$ and $RSPE$ for categorical datasets are reported in Table 7. NBRLR does slightly worse than hybrid LR-NB in $L_{0\text{-}1}$ and $RSPE$. Two-tailed, paired samples Wilcoxon test suggests that the difference is insignificant for either of the two metrics. Notice these results are only based on ten categorical datasets.

5.3. NBRLR versus regularized LR

The Win/Draw/Loss (W/D/L) results of NBRLR against lasso, ridge, and elastic net are in Table 8. It can be seen that NBRLR does better than all three regularized LR methods in both $L_{0\text{-}1}$ and $RSPE$ over the entire 30 datasets, although the improvement over lasso is not significant for both $L_{0\text{-}1}$ and $RSPE$, and the improvement over elastic net is not significant for $L_{0\text{-}1}$. However, when the number of predictors is small, as we have discussed, the sparsity assumption of lasso is more likely to be violated, making the regression estimates suboptimal. Also, predictors tend to have less multicollinearity, limiting ridge regression's benefits. As a result, the improvement of NBRLR is more pronounced on low-dimensional datasets with less than 10 predictors.

6. Conclusion

This paper presents a naïve Bayes regularized logistic regression model for classification problems. As LR is a low–bias, high–variance classifier, many regularized methods have been proposed to overcome LR's overfitting issue, which may lead to poor prediction performance when the training sample is limited, or there is a large number of parameters to be estimated. Most of these methods assume that the true coefficients of LR are sparse. However, this sparsity assumption is often violated when p is relatively small compared to n, which makes the regression estimates suboptimal. Meanwhile, there also tends to be less multicollinearity among predictors.

This limits the benefits of ridge regression, which is not motivated by sparsity. We argue that when p is relatively small compared to n, shrinking the coefficients towards a low–variance data-driven estimate could be a better regularization strategy.

Our approach is primarily motivated by the fact that NB has an equivalent functional form compared to LR, given NB's conditional independence assumption holds. The resulting classifier tends to have higher variance but lower bias than lasso when p is relatively small compared to n. Simulation and empirical, experimental results suggest that NBRLR can generally outperform pure LR and pure NB. Also, it is highly competitive with regularized LR, especially on low and moderate-dimension datasets.

References

Domingos, P., Pazzani, M., 1996. Beyond independence: Conditions for the optimality of the simple Bayesian classifier, in: Proc. 13th Intl. Conf. Machine Learning, pp. 105–112.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. The Annals of Statistics 32, 407–499.

Eggleston, H.G., 1958. Convexity, Cambridge Tracts in Mathematics and Mathematical Physics, No.47. Cambridge University Press.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics 32, 928–961.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1.

Fujino, A., Ueda, N., Saito, K., 2007. A hybrid generative/discriminative approach to text classification with additional information. Information Processing and Management 43, 379–392.

Hand, D.J., Yu, K., 2001. Idiot's Bayes—not so stupid after all? International Statistical Review 69, 385–398.

Hastie, T., Friedman, J., Tibshirani, R., 2001. The Elements of Statistical Learning: Data mining, Inference and Prediction. Springer, New York, NY.

He, X., Shi, P., 1994. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. Journal of Nonparametric Statistics 3, 299–308.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Kang, C., Tian, J., 2006. A hybrid generative/discriminative Bayesian classifier, in: Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference, pp. 562–567.

Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28, 1356–1378.

Kwon, S., Kim, Y., 2012. Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. Statistica Sinica 22, 629–653.

Mitchell, T.M., 1997. Machine Learning. WCB/McGraw-Hill, Boston, MA.

Ng, A.Y., Jordan, M.I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes, in: Advances in Neural Information Processing Systems, pp. 841–848.

Niblett, T., 1987. Constructing decision trees in noisy domains, in: Proceedings of the Second European Working Session on Learning, Sigma, Bled, Yugoslavia. pp. 67–78.

Raina, R., Shen, Y., Ng, A.Y., McCallum, A., 2003. Classification with hybrid generative/discriminative models, in: Thrun, S., Saul, L.K., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems, pp. 545–552.

Tan, Y., Shenoy, P.P., 2020. A bias-variance based heuristic for constructing a hybrid logistic regression-naïve Bayes model for classification. International Journal of Approximate Reasoning 117, 15–28.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.

Wilcoxon, F., 1992. Individual comparisons by ranking methods, in: Breakthroughs in Statistics: Methodology and Distribution. Springer, pp. 196–202.

Zaidi, N.A., Carman, M.J., Cerquides, J., Webb, G.I., 2014. Naive-Bayes inspired effective pre-conditioner for speeding-up logistic regression, in: IEEE 2014 International Conference on Data Mining (ICDM), IEEE. pp. 1097–1102.

Zaidi, N.A., Cerquides, J., Carman, M.J., Webb, G.I., 2013. Alleviating naive Bayes attribute independence assumption by attribute weighting. Journal of Machine Learning Research 14, 1947–1988.

Zaidi, N.A., Webb, G.I., Carman, M.J., Petitjean, F., Cerquides, J., 2016. ALR^n: accelerated higher-order logistic regression. Machine Learning 104, 151–194.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Table 1: Summary of results from simulation setting 1 comparing NBRLR with pure LR, pure NB, hybrid LR-NB, lasso, ridge, and elastic net at different values of N_{tr}, p, and r. The p-values (in parenthesis) are from two-tailed, paired t-tests comparing the performance of NBRLR with the corresponding estimator. Best performances are in boldface.

Simu. 1	N_{tr}	p	r	NBRLR	LR	NB	Hybrid LR-NB	Lasso	Ridge	Elastic Net
MSE	100	10	0.1	**0.332**	0.429 (<0.001)	0.335 (0.906)	NA	0.574 (<0.001)	0.541 (<0.001)	0.403 (0.001)
			0.6	**0.402**	0.626 (<0.001)	0.607 (<0.001)	NA	0.726 (<0.001)	0.643 (<0.001)	0.473 (0.017)
		50	0.1	**0.643**	1072.686 (<0.001)	0.734 (0.051)	NA	0.872 (<0.001)	0.725 (0.068)	1.616 (0.002)
			0.6	**0.777**	36180.090 (0.003)	1.795 (<0.001)	NA	0.909 (0.043)	0.812 (0.595)	4.721 (<0.001)
	2000	10	0.1	**0.013**	0.013 (0.665)	0.113 (<0.001)	NA	0.101 (<0.001)	0.088 (<0.001)	0.153 (<0.001)
			0.6	**0.017**	0.017 (0.404)	0.303 (<0.001)	NA	0.116 (<0.001)	0.107 (<0.001)	0.166 (<0.001)
		50	0.1	**0.024**	0.030 (<0.001)	0.381 (<0.001)	NA	0.129 (<0.001)	0.113 (<0.001)	0.098 (<0.001)
			0.6	**0.034**	0.040 (<0.001)	0.566 (<0.001)	NA	0.133 (<0.001)	0.129 (<0.001)	0.098 (<0.001)
$L_{0\text{-}1}$	100	10	0.1	0.300	**0.296** (<0.001)	0.311 (<0.001)	0.301 (0.388)	0.306 (<0.001)	0.299 (0.699)	0.316 (<0.001)
			0.6	0.360	**0.355** (<0.001)	0.401 (<0.001)	0.357 (0.003)	0.381 (<0.001)	0.362 (0.331)	0.381 (<0.001)
		50	0.1	0.290	0.300 (0.007)	0.310 (<0.001)	0.320 (<0.001)	0.334 (<0.001)	**0.269** (<0.001)	0.296 (0.068)
			0.6	0.335	0.350 (<0.001)	0.434 (<0.001)	0.365 (<0.001)	0.401 (<0.001)	**0.328** (0.006)	0.345 (0.004)
	2000	10	0.1	**0.274**	0.274 (0.211)	0.274 (0.640)	0.274 (0.211)	0.274 (0.207)	0.274 (0.419)	0.275 (0.122)
			0.6	0.331	0.331 (0.741)	0.334 (0.125)	0.331 (0.741)	0.331 (0.623)	0.331 (0.807)	**0.331** (0.710)
		50	0.1	**0.156**	0.156 (0.915)	0.168 (<0.001)	0.156 (0.915)	0.156 (0.473)	0.156 (0.315)	0.157 (<0.001)
			0.6	0.218	**0.218** (0.283)	0.310 (<0.001)	**0.218** (0.283)	0.218 (0.893)	0.218 (0.871)	0.219 (0.004)
RSPE	100	10	0.1	0.445	0.447 (0.017)	0.448 (0.001)	0.448 (<0.001)	0.448 (<0.001)	**0.443** (<0.001)	0.452 (<0.001)
			0.6	0.470	0.471 (0.099)	0.501 (<0.001)	0.473 (0.003)	0.475 (<0.001)	**0.467** (<0.001)	0.475 (<0.001)
		50	0.1	0.448	0.542 (<0.001)	0.459 (<0.001)	0.522 (<0.001)	0.468 (<0.001)	**0.424** (<0.001)	0.460 (0.001)
			0.6	0.475	0.577 (<0.001)	0.592 (<0.001)	0.577 (<0.001)	0.492 (<0.001)	**0.455** (<0.001)	0.498 (<0.001)
	2000	10	0.1	0.419	**0.419** (0.320)	0.425 (<0.001)	**0.419** (0.320)	0.419 (0.068)	0.420 (<0.001)	0.428 (<0.001)
			0.6	**0.446**	0.446 (0.592)	0.460 (<0.001)	0.446 (0.592)	0.446 (0.910)	0.446 (0.006)	0.453 (<0.001)
		50	0.1	**0.330**	0.330 (<0.001)	0.362 (<0.001)	0.330 (<0.001)	0.330 (0.172)	0.330 (<0.001)	0.335 (<0.001)
			0.6	**0.377**	0.377 (0.081)	0.456 (<0.001)	0.377 (0.081)	0.377 (0.263)	0.377 (0.081)	0.381 (<0.001)

Table 2: Summary of results from simulation setting 2 comparing NBRLR with pure LR, pure NB, hybrid LR-NB, lasso, ridge, and elastic net at different values of N_{tr}, p, and r. The p-values (in parenthesis) are from two-tailed, paired t-tests comparing the performance of NBRLR with the corresponding estimator. Best performances are in boldface.

Simu. 2	N_{tr}	p	r	NBRLR	LR	NB	Hybrid LR-NB	Lasso	Ridge	Elastic Net
MSE	100	10	0.1	0.261	0.385 (<0.001)	0.241 (0.009)	NA	0.341 (<0.001)	0.286 (0.162)	**0.239** (0.066)
			0.6	**0.458**	1.213 (0.023)	0.524 (0.166)	NA	0.734 (<0.001)	0.617 (0.002)	0.495 (0.335)
		50	0.1	0.405	1351.846 (0.013)	0.375 (0.358)	NA	0.358 (0.183)	**0.254** (<0.001)	0.788 (0.002)
			0.6	**0.581**	27558.56 (0.048)	1.337 (<0.001)	NA	1.002 (<0.001)	0.753 (<0.001)	1.530 (<0.001)
	2000	10	0.1	**0.011**	0.012 (0.275)	0.013 (<0.001)	NA	0.055 (<0.001)	0.050 (<0.001)	0.060 (<0.001)
			0.6	**0.020**	0.021 (0.554)	0.216 (<0.001)	NA	0.135 (<0.001)	0.116 (<0.001)	0.234 (<0.001)
		50	0.1	**0.014**	0.031 (<0.001)	0.017 (<0.001)	NA	0.070 (<0.001)	0.053 (<0.001)	0.081 (<0.001)
			0.6	**0.099**	0.162 (<0.001)	0.343 (<0.001)	NA	0.258 (<0.001)	0.272 (<0.001)	0.477 (<0.001)
$L_{0\text{-}1}$	100	10	0.1	**0.338**	0.339 (0.593)	0.339 (0.172)	0.339 (0.420)	0.352 (<0.001)	0.342 (0.029)	0.350 (<0.001)
			0.6	**0.312**	0.315 (0.023)	0.343 (<0.001)	0.317 (0.002)	0.325 (<0.001)	0.314 (0.229)	0.323 (<0.001)
		50	0.1	0.178	0.276 (<0.001)	**0.176** (0.197)	0.242 (<0.001)	0.242 (<0.001)	0.184 (<0.001)	0.215 (<0.001)
			0.6	**0.147**	0.250 (<0.001)	0.256 (<0.001)	0.243 (<0.001)	0.207 (<0.001)	0.149 (0.171)	0.170 (<0.001)
	2000	10	0.1	0.307	0.307 (0.687)	**0.307** (0.363)	0.307 (0.781)	0.307 (0.724)	0.307 (0.489)	0.308 (0.443)
			0.6	**0.288**	0.289 (0.589)	0.290 (0.069)	0.289 (0.589)	0.288 (0.610)	0.289 (0.091)	0.289 (0.259)
		50	0.1	**0.125**	0.129 (<0.001)	0.125 (0.212)	0.128 (<0.001)	0.129 (<0.001)	0.128 (<0.001)	0.128 (<0.001)
			0.6	0.105	0.106 (0.015)	0.142 (<0.001)	0.106 (0.015)	0.106 (0.013)	**0.104** (<0.001)	0.104 (0.015)
RSPE	100	10	0.1	0.467	0.472 (<0.001)	0.466 (0.008)	0.469 (<0.001)	0.471 (<0.001)	**0.464** (0.008)	0.469 (0.037)
			0.6	0.445	0.450 (<0.001)	0.472 (<0.001)	0.451 (<0.001)	0.449 (<0.001)	**0.443** (0.001)	0.449 (<0.001)
		50	0.1	0.360	0.519 (<0.001)	**0.356** (0.014)	0.461 (<0.001)	0.413 (<0.001)	0.362 (0.265)	0.397 (<0.001)
			0.6	**0.317**	0.486 (<0.001)	0.450 (<0.001)	0.467 (<0.001)	0.376 (<0.001)	0.318 (0.540)	0.351 (<0.001)
	2000	10	0.1	**0.443**	0.443 (0.066)	0.443 (0.310)	0.443 (0.005)	0.443 (0.028)	0.443 (0.348)	0.448 (<0.001)
			0.6	0.419	**0.419** (0.096)	0.430 (<0.001)	**0.419** (0.096)	0.419 (0.401)	0.419 (<0.001)	0.430 (<0.001)
		50	0.1	**0.299**	0.303 (<0.001)	0.300 (0.009)	0.303 (<0.001)	0.303 (<0.001)	0.302 (<0.001)	0.313 (<0.001)
			0.6	**0.266**	0.267 (<0.001)	0.311 (<0.001)	0.267 (<0.001)	0.266 (0.036)	0.268 (<0.001)	0.290 (<0.001)

Table 3: Squared bias of the six compared estimators at different values of N_{tr}, p, and r for the first two simulation settings.

Simu. Setting	N_{tr}	p	r	NBRLR	LR	NB	Lasso	Ridge	Elastic Net
1	100	10	0.1	0.009	0.022	0.103	0.453	0.498	0.070
			0.6	0.025	0.038	0.291	0.638	0.598	0.139
		50	0.1	0.219	268.963	0.364	0.828	0.692	0.096
			0.6	0.175	5569.745	0.536	0.824	0.780	0.066
	2000	10	0.1	0.0005	0.0002	0.101	0.090	0.080	0.102
			0.6	0.0006	0.0002	0.289	0.102	0.097	0.106
		50	0.1	0.0029	0.0027	0.365	0.114	0.102	0.043
			0.6	0.0036	0.0030	0.518	0.112	0.114	0.043
2	100	10	0.1	0.0019	0.016	0.0021	0.259	0.249	0.037
			0.6	0.025	0.059	0.208	0.580	0.564	0.156
		50	0.1	0.003	173.347	0.006	0.280	0.223	0.032
			0.6	0.122	1682.375	0.293	0.809	0.700	0.241
	2000	10	0.1	0.001	0.000	0.002	0.045	0.044	0.038
			0.6	0.0010	0.0003	0.1989	0.1192	0.1063	0.1543
		50	0.1	0.0016	0.0019	0.0021	0.057	0.044	0.050
			0.6	0.030	0.021	0.305	0.233	0.268	0.395

Table 4: Variance of the six compared estimators at different values of N_{tr}, p, and r for the first two simulation settings.

Simu. Setting	N_{tr}	p	r	NBRLR	LR	NB	Lasso	Ridge	Elastic Net
1	100	10	0.1	0.326	0.412	0.234	0.122	0.044	0.337
			0.6	0.382	0.594	0.319	0.088	0.046	0.337
		50	0.1	0.428	811.841	0.374	0.045	0.033	1.535
			0.6	0.608	30919.540	1.272	0.086	0.032	4.702
	2000	10	0.1	0.0124	0.0128	0.0118	0.0111	0.0081	0.0514
			0.6	0.0168	0.0175	0.0142	0.0137	0.0099	0.0606
		50	0.1	0.022	0.026	0.016	0.015	0.011	0.056
			0.6	0.031	0.037	0.048	0.021	0.015	0.055
2	100	10	0.1	0.262	0.374	0.240	0.075	0.029	0.201
			0.6	0.438	1.164	0.322	0.160	0.059	0.345
		50	0.1	0.405	1190.320	0.373	0.083	0.035	0.765
			0.6	0.463	26137.610	1.053	0.194	0.053	1.301
	2000	10	0.1	0.0111	0.0118	0.0110	0.0098	0.0063	0.0216
			0.6	0.020	0.021	0.018	0.016	0.010	0.080
		50	0.1	0.012	0.029	0.015	0.011	0.009	0.031
			0.6	0.073	0.139	0.050	0.036	0.016	0.096

Table 5: A Summary of the 30 datasets used in the empirical results. The Type column indicates if the predictors are categorical or continuous. Instances are the number of observations in the data set.

Dataset	*# Predictors*	*# Instances*	*# Type*
Balloons	4	20	Categorical
Qualitative Bankruptcy	6	250	Categorical
Blogger	5	100	Categorical
SPECT Heart	22	267	Categorical
Tic-Tac-Toe Endgame	9	958	Categorical
Congressional Voting Records	16	435	Categorical
Chess1	36	3196	Categorical
Chess2	6	28056	Categorical
Phishing Website Detector	30	11054	Categorical
Monkey-Pox	9	25000	Categorical
Blood Transfusion Service Center	4	748	Continuous
Connectionist Bench	60	208	Continuous
Haberman's Survival	3	306	Continuous
Liver Disorders	6	345	Continuous
Pima Indians Diabetes	8	768	Continuous
Vertebral Column	6	310	Continuous
Occupancy Detection	5	8143	Continuous
HTRU2	8	17898	Continuous
Pulsar Classication	8	17898	Continuous
Microcalcication Classication	6	11183	Continuous
ILPD (Indian Liver Patient Dataset)	10	583	Mixed
Cryotherapy	6	90	Mixed
Heart Failure Clinical Records	12	299	Mixed
Statlog (Heart)	13	270	Mixed
Australian Credit Approval	14	690	Mixed
Hepatitis	19	155	Mixed
Hypothyroid	17	3163	Mixed
Bank Customer Churn Prediction	10	10000	Mixed
Gender Classication	7	5001	Mixed
Default of Credit Card Clients	23	30000	Mixed

Table 6: Win/Draw/Loss: NBRLR versus LR and NB

	NBRLR vs. LR		*NBRLR vs. NB*	
	W/D/L	p-value	W/D/L	p-value
$L_{0\text{-}1}$	15/2/13	0.715	21/2/7	<0.001
$RSPE$	17/0/13	0.472	26/0/4	<0.001
	Small Size Dataset (9)		*Large Size Dataset (8)*	
$L_{0\text{-}1}$	5/1/3	0.294	6/1/1	<0.001
$RSPE$	7/0/2	0.019	8/0/0	<0.001

Table 7: Win/Draw/Loss: NBRLR versus hybrid LR-NB

	NBRLR vs. hybrid LR-NB	
	W/D/L	p-value
$L_{0\text{-}1}$	3/2/5	0.363
$RSPE$	4/0/6	0.813

Table 8: Win/Draw/Loss: NBRLR versus regularized LR

	NBRLR vs. Lasso		*NBRLR vs. Ridge*		*NBRLR vs. Elastic Net*	
	W/D/L	p-value	W/D/L	p-value	W/D/L	p-value
$L_{0\text{-}1}$	15/3/12	0.532	18/3/9	0.041	15/2/13	0.855
$RSPE$	16/0/14	0.943	26/0/4	<0.001	23/0/7	0.003
	Low-Dim Dataset (17)		*Low-Dim Dataset (17)*		*Low-Dim Dataset (17)*	
$L_{0\text{-}1}$	13/3/1	0.008	12/3/2	0.008	12/2/3	0.012
$RSPE$	12/0/5	0.041	17/0/0	<0.001	17/0/0	<0.001