# Feature Augmentation via Nonparametrics and Selection (FANS) in High Dimensional Classification*

Jianqing Fan, Yang Feng, Jiancheng Jiang and Xin Tong

**Abstract**

We propose a high dimensional classification method that involves nonparametric feature augmentation. Knowing that marginal density ratios are building blocks of the Bayes rule for each feature, we use the ratio estimates to transform the original feature measurements. Subsequently, we invoke penalized logistic regression, taking as input the newly transformed or augmented features. This procedure trains models with high local complexity and a simple global structure, thereby avoiding the curse of dimensionality while creating a flexible nonlinear decision boundary. The resulting method is called Feature Augmentation via Nonparametrics and Selection (FANS). We motivate FANS by generalizing the Naive Bayes model, writing the log ratios of joint densities as a linear combination of those of marginal densities. It is related to generalized additive models, but has better interpretability and computability. Risk bounds are developed for FANS. In numerical analysis, FANS is compared with competing methods, so as to provide a guideline on its best application domain. Real data analysis demonstrates that FANS performs very competitively on benchmark email spam and gene expression data sets. Moreover, FANS is implemented by an extremely fast algorithm through parallel computing.

**Keywords:** density estimation, classification, high dimensional space, nonlinear decision boundary, feature augmentation, feature selection, parallel computing.

*Jianqing Fan is Frederick L. Moore Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton NJ 08544 (Email: jqfan@princeton.edu). Yang Feng is Assistant Professor, Department of Statistics, Columbia University, New York, NY 10027 (Email: yangfeng@stat.columbia.edu). Jiancheng Jiang is Associate Professor, Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, 28223 (Email: jjiang1@uncc.edu). Xin Tong is Assistant Professor, Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, (Email: xint@marshall.usc.edu). The financial support from National Institutes of Health grants R01-GM072611 and R01GM100474-01 and National Science Foundation grants DMS-1206464 and DMS-1308566 is greatly acknowledged.

# 1 Introduction

Classification aims to identify to which category a new observation belongs based on feature measurements. Common applications include disease classification using high-throughput data such as microarrays, SNPs, spam detection and image recognition. Well known classification methods include Fisher's linear discriminant analysis (LDA), logistic regression, Naive Bayes, $k$-nearest neighbor, neural networks, and many others. All these methods can perform well in the classical low dimensional settings, in which the number of features is much smaller than the sample size. However, in many contemporary applications, the number of features $p$ is large compared to the sample size $n$. For instance, the dimensionality $p$ in microarray data is frequently in thousands or beyond, while the sample size $n$ is typically in the order of tens. Besides computational issues, the central conflict in high dimensional setup is that the model complexity is not supported by limited access to data. In other words, the "variance" of conventional models is high in such new settings, and even simple models such as LDA need to be regularized. We refer to Hastie et al. (2009) and Bühlmann and van de Geer (2011) for overviews of statistical challenges associated with high dimensionality.

In this paper, we propose a classification procedure FANS (Feature Augmentation via Nonparametrics and Selection). Before introducing the algorithm, we first detail its motivation. Suppose feature measurements and responses are coded by a pair of random variables $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ denotes the features and $Y \in \{0, 1\}$ is the binary response. Recall that a classifier $h$ is a data-dependent mapping from the feature space to the labels. Classifiers are usually constructed to minimize the risk $P(h(\boldsymbol{X}) \neq Y)$.

Denote by $g$ and $f$ the class conditional densities respectively for class 0 and class 1, i.e., $(\boldsymbol{X}|Y = 0) \sim g$ and $(\boldsymbol{X}|Y = 1) \sim f$. It can be shown that the Bayes rule is $\mathbb{1}(r(\boldsymbol{x}) \geq 1/2)$, where $r(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$. Let $\pi = P(Y = 1)$, then

$$r(\boldsymbol{x}) = \frac{\pi f(\boldsymbol{x})}{\pi f(\boldsymbol{x}) + (1 - \pi) g(\boldsymbol{x})} .$$

Assume for simplicity that $\pi = 1/2$, then the oracle decision boundary is

$$\{\boldsymbol{x} : f(\boldsymbol{x})/g(\boldsymbol{x}) = 1\} = \{\boldsymbol{x} : \log f(\boldsymbol{x}) - \log g(\boldsymbol{x}) = 0\} ,$$

Denote by $g_1, \cdots, g_p$ the marginals of $g$, and by $f_1, \cdots, f_p$ those of $f$. Nonparametric Naive Bayes model assumes that the conditional distributions of each feature given the class labels are independent, i.e.,

$$\log \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = \sum_{j=1}^{p} \log \frac{f_j(x_j)}{g_j(x_j)} . \tag{1.1}$$

Naive Bayes is a simple approach, but it is useful in high-dimensional settings. Taking a two class Gaussian model with common covariance matrix, Bickel and Levina (2004) showed that naively carrying out Fisher's discriminant rule performs poorly due to diverging spectra. In addition, the authors argued that independence rule which ignores the covariance structure performs better than Fisher's rule in high-dimensional settings. However, correlation among features is usually an essential characteristic of data, and it can help classification under suitable models and sample size to feature dimension ratios. Examples in bioinformatics study can be found in Ackermann and Strimmer (2009). Recently, Fan et al. (2012) showed that the independence assumption can lead to huge loss in classification power when correlation prevails, and proposed a Regularized Optimal Affine Discriminant (ROAD). ROAD is a linear plug-in rule targeting directly on the classification error, and it takes advantages of the unregularized pooled sample covariance matrix.

Relaxing the two-class Gaussian assumption in parametric Naive Bayes gives us a general Naive Bayes formulation such as (1.1). However, this model also fails to capture the correlation, or dependence among features in general. This consideration motivates us to ask the following question: what if we add weights in front of each log marginal density ratio and optimize them under certain criterions (all coefficients are 1 in Naive Bayes model, so there is no need for optimization). More precisely, we would like to learn a decision boundary from the following set

$$\mathcal{D} = \left\{ \boldsymbol{x} : \beta_0 + \beta_1 \log \frac{f_1(x_1)}{g_1(x_1)} + \cdots + \beta_p \log \frac{f_p(x_p)}{g_p(x_p)} = 0, \beta_0, \cdots, \beta_p \in \mathbb{R} \right\}. \tag{1.2}$$

For univariate problems, the marginal density ratio delivers the best classifier based on only one feature. Therefore, the marginal density ratios can be regarded as the best transformations of the future measurements, and (1.2) is an effort towards combining those most powerful univariate transforms to build more powerful classifiers.

This is in a similar spirit as the sure independence screening in Fan and Lv (2008) where the best marginal predictors are used as probes for their utilities in the joint model. By combining these marginal density ratios and optimizing over their coefficients $\beta_j$'s, we wish to build a good classifier that takes into account feature dependence. Note that although our target boundary $\mathcal{D}$ is not linear in the original features, but it is linear in the parameters $\beta_j$'s. Therefore, after renaming the transformed variables, any linear classifiers can be applied. For example, we can use logistic regression, one of the most popular linear classification rules. Other choices, such as SVM (linear kernel), are good alternatives, but we fix logistic regression for the rest of discussion.

Recall that logistic regression models the log odds ratio by

$$\log \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \,,$$

where the $\beta_j$'s are estimated by the maximum likelihood approach. We should note that without explicitly modeling correlations, logistic regression takes into account features' joint effects and levels a good linear combination of features as the decision boundary. Its performance is similar to LDA, but both models can only capture decision boundaries linear in original features.

On the other hand, logistic regression might serve as a building block for the more flexible FANS. Concretely, if we know the marginal densities $f_j$ and $g_j$, and run logistic regression on the transformed features $\{\log(f_j(x_j)/g_j(x_j))\}$, we create a decision boundary nonlinear in original features. This use of the transformed features is easily interpretable: one naturally combines the "most powerful" univariate transforms (building blocks of univariate Bayes rules) $\{\log(f_j(x_j)/g_j(x_j))\}$ rather than the original measurements. In special cases such as the two-class Gaussian model with common covariance matrix, the transformed features are not different from the original ones. Some caution should be taken: if $f_j = g_j$ for some $j$, i.e., the marginal densities for feature $j$ are exactly the same, this feature will not have any contribution in classification. Deletion like this might lose power, because features having no marginal contribution on their own might boost classification performance when they are used jointly with other features. In view of this defect, an alternative method is to augment the original features with the transformed ones.

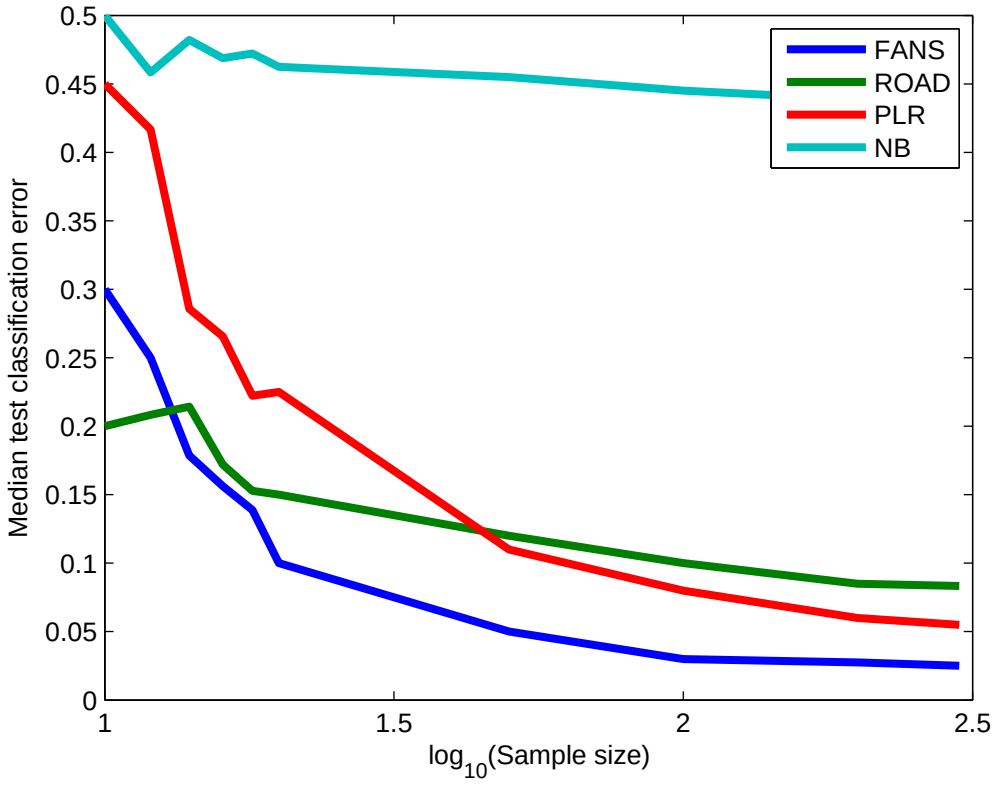
Since marginal densities $f_j$ and $g_j$ are unknown, we need to first estimate them, and then run a penalized logistic regression (PLR) on the estimated transforms. Such regularization (penalization) is necessary to reduce model complexity in the high dimensional paradigm. This two-step classification rule of feature augmentation via nonparametrics and selection will be called "FANS" for short. Precise algorithmic implementation of FANS is described in the next section. Numerical results show that our new method excels in many scenarios, in particular when no linear decision boundary can separate the data well.

To understand where FANS stands compared to Naive Bayes (NB), penalized logistic regression (PLR), and the regularized optimal affine discriminant (ROAD), we showcase a simple simulation example. In this example, the choice is between a multivariate Gaussian distribution and the componentwise mixture of two multivariate Gaussian distributions:

Class 0: $N((5 \times \mathbf{1}_{10}^T, \mathbf{0}_{p-10}^T)^T, \boldsymbol{\Sigma})$,

Class 1: $\boldsymbol{w} \circ N(\mathbf{0}_p, \boldsymbol{I}_p) + (1-\boldsymbol{w}) \circ N((6 \times \mathbf{1}_{10}^T, \mathbf{0}_{p-10}^T)^T, \boldsymbol{\Sigma})$, where $p = 1000$, $\circ$ is the elementwise product between matrices, $\Sigma_{ii} = 1$ for all $i = 1, \cdots, p$, $\Sigma_{ij} = 0.5$ for all $i, j = 1, \cdots, p$ and $i \neq j$. $\boldsymbol{w} = (w_1, \cdots, w_p)^T$, where $w_j \sim^{iid}$ Bernoulli(0.5).

Figure 1: The median test errors for Gaussian vs. mixture of Gaussian when the training data size varies.




The median classification error for 100 repetitions as a function of training sample size $n$ is rendered in Figure 1. This figure suggests that increasing sample sizes does not help NB boost performance, because the NB model is severely biased in view of significant correlation presence. It is interesting to compare PLR with ROAD. ROAD is a more efficient approach when the sample size is small; however, PLR eventually does better when the sample size becomes large enough. This is not surprising because the underlying true model is not two class Gaussian with common covariance matrix. So the less "biased" PLR beats ROAD on large samples. Nevertheless, even if ROAD uses a misspecified model, it still benefits from a specific model assumption on small samples. Finally, since the oracle decision boundary in this example is nonlinear, the newly proposed FANS approach performs significantly better than others when the sample size is reasonably large. The above analysis seems to suggest that FANS does well as long as we have enough data to construct accurate marginal density estimates. Note also that ROAD is better than FANS when the training sample size is extremely small. The best method in practice largely depends on the available data.

A popular extension of logistic regression and close relative to FANS is the additive logistic regression, which belongs to generalized additive models (Hastie and Tibshirani, 1990).

Additive logistic regression allows (smooth) nonparametric feature transformations to appear in the decision boundary by modeling

$$\log \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \sum_{j=1}^{p} h_j(x_j)\,, \tag{1.3}$$

where $h_j$'s are smooth functions. This additive decision boundary is very general, in which FANS and logistic regression are special cases. It works well for small-$p$-large-$n$ scenarios, while its penalized versions adapt to high dimensional settings. We will compare FANS with penalized additive logistic regression in numerical studies. The major drawback of additive logistic regression (generalized additive model) is the heavy computational complexity (e.g., the backfitting algorithm) involved in searching the transformation functions $h_j(\cdot)$. Moreover, the available algorithms, e.g., the algorithm for penGAM (Meier et al., 2009), fail to give an estimate when the sample size is very small. Compared to FANS, the generalized additive model uses a factor of $K_n$ more parameters, where $K_n$ is the number of knots in the approximation of every additive components $\{h_j(\cdot)\}_{j=1}^{p}$. While this reduces possible biases in comparison with FANS, it increases variances in the estimation and results in more computation cost (see Table 2). Moreover, FANS admits a nice interpretation of optimal combination of optimal univariate classifiers.

Besides the aforementioned references, there is a huge literature on high dimensional classification. Examples include principal component analysis in Bair et al. (2006) and Zou et al. (2006), partial least squares in Nguyen and Rocke (2002), Huang (2003) and Boulesteix (2004), and sliced inverse regression in Li (1991) and Antoniadis et al. (2003). Recently, there has been a surge of interest for extending the linear discriminant analysis to high-dimensional settings including Guo et al. (2007), Wu et al. (2009), Clemmensen et al. (2011), Shao et al. (2011), Cai and Liu (2011), Fan et al. (2012), Mai et al. (2012) and Witten and Tibshirani (2012).

The rest of the paper is organized as follows. Section 2 introduces an efficient algorithm for FANS. Section 3 is dedicated to simulation studies and real data analysis. Theoretical results are presented in Section 4. We conclude with a short discussion in Section 5. Longer proofs and technical results are relegated to the Appendix.

# 2 Algorithm

In this section, we detail an efficient algorithm ($S1 - S5$) for FANS. A variant of FANS (FANS2), which uses the transformed features to augment the original ones, is also described.

## 2.1 FANS and its Running Time Bound

*S*1. Given $n$ pairs of observations $D = \{(\boldsymbol{X}_i, Y_i), i = 1, \cdots, n\}$. Randomly split the data into two parts for $L$ times: $D_l = (D_{l1}, D_{l2})$, $l = 1, \cdots, L$.

*S*2. On each $D_{l1}$, $l \in \{1, \cdots, L\}$, apply kernel density estimation and denote the estimates by $\hat{\boldsymbol{f}} = (\hat{f}_1, \cdots, \hat{f}_p)^T$ and $\hat{\boldsymbol{g}} = (\hat{g}_1, \cdots, \hat{g}_p)^T$.

*S*3. Calculate the transformed observations $\hat{\boldsymbol{Z}}_i = \boldsymbol{Z}_{\hat{\boldsymbol{f}},\hat{\boldsymbol{g}}}(\boldsymbol{X}_i)$, where $\hat{Z}_{ij} = \log \hat{f}_j(X_{ij}) - \log \hat{g}_j(X_{ij})$, for each $i \in D_{l2}$ and $j \in \{1, \ldots, p\}$.

*S*4. Fit an $L_1$-penalized logistic regression to the transformed data $\{(\hat{\boldsymbol{Z}}_i, Y_i), i \in D_{l2}\}$, using cross validation to get the best penalty parameter. For a new observation $\boldsymbol{x}$, we estimate transformed features by $\log \hat{f}_j(x_j) - \log \hat{g}_j(x_j)$, $j = 1, \ldots, p$, and plug them into the fitted logistic function to get the predicted probability $p_l = P(Y = 1 | \boldsymbol{X} = \boldsymbol{x})$.

*S*5. Repeat (*S*2)-(*S*4) for $l = 1, \cdots, L$, use the average predicted probability $prob = L^{-1} \sum_{l=1}^{L} p_l$ as the final prediction, and assign the observation $\boldsymbol{x}$ to class 1 if $prob \geq 1/2$, and 0 otherwise.

A few comments on the technical implementation are made as follows.

**Remark 1**

*i). In S*2*, if an estimated marginal density is less than some threshold* $\varepsilon$ *(say* $10^{-2}$*), we set it to be* $\varepsilon$*. This Winsorization increases the stability of the transformations, because the estimated transformations* $\log \hat{f}_j$ *and* $\log \hat{g}_j$ *are unstable in regions where true densities are small.*

*ii). In S*4*, we take penalized logistic regression, but any linear classifier can be used. For example, support vector machine (SVM) with linear kernel is also a good choice.*

*iii). In S*4*, the* $L_1$ *penalty (Tibshirani, 1996) was adopted since our primary interest is the classification error. We can also apply other penalty functions, such as SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006) and MCP (Zhang, 2010).*

*iv). In S*5*, the average predicted probability is taken as the final prediction. An alternative approach is to make a decision on each random split, and listen to majority votes.*

In *S*1, we split the data multiple times. The rationale behind multiple splitting lies in the two-step prototype nature of FANS, which uses the first part of the data for marginal

nonparametric density estimates (in $S2$) and (transformation of) the second part for penalized logistic regression (in $S4$). Multiple splitting and prediction averaging not only make our procedure more robust against arbitrary assignments of data usage, but also make more efficient use of limited data. We take $L = 20$ in our numerical studies. This choice reflects our cluster's node number. Interested readers can as well leverage their better computing resources for a larger $L$, but our experience is that more splits do not help universally in our numerical examples. Also, we recommend a balanced assignment by switching the role of data used for feature transformation and for feature selection, i.e., $D_{2l} = (D_{(2l-1),2}, D_{(2l-1),1})$ when $D_{2l-1} = (D_{(2l-1),1}, D_{(2l-1),2})$.

It is straightforward to derive a running time bound for our algorithm. Suppose splitting has been done. In $S2$, we need to perform kernel density estimation for each variable, which costs $O(n^2p)$[1]. The transformations in $S3$ cost $O(np)$. In $S4$, we call the R package **glmnet** to implement penalized logistic regression, which employs the coordinate decent algorithm for each penalty level. This step has a computational cost at most $O(npT)$, where $T$ is the number of penalty levels, i.e., the number of times the coordinate descent algorithm is run (see Friedman et al. (2007) for a detailed analysis). The default setting is $T = 100$, though we can set it to other constants. Therefore, a running time bound for the whole algorithm is $O(L(n^2p + np + npT)) = O(Lnp(n + T))$.

The above bound does not look particularly interesting. However, smart implementation of the FANS procedure can fully unleash the potential of our algorithm. Indeed, not only the $L$ repetitions, but also the marginal density estimates in $S2$ can be done via parallel computing. Suppose $L$ is the number of available nodes, and the cpu core number in each node is $N \geq n/T$. This assumption is reasonable because $T = 100$ by default, $N = 8$ for our implementation, and sample sizes $n$ for many applications are less than a multiple of $TN$. Under this assumption, the $L$ predicted probabilities calculations can be carried out simultaneously and the results are combined later in $S5$. Moreover in $S2$, the running time bound becomes $O(n^2p/N)$. Henceforth, a bound for the whole algorithm will be $O(npT)$, which is the same as for penalized logistic regression. The exciting message here is that, by leveraging modern computer architecture, we are able to implement our nonparametric classification rule FANS within running time at the order of a parametric method. The computation times for various simulation setups are reported in Table 2, where the first column reports results when only $L$ repetitions are paralleled, and the second column reports the improvement when marginal density estimates in $S2$ are paralleled within each node.

[1]Approximate kernel density estimates can be computed faster, see e.g., Raykar et al. (2010).

### 2.2 Augmenting Linear Features

As we argued in the introduction, features with no marginal discrimination power do not make contribution in FANS. One remedy is to run (in $S4$) the penalized logistic regression using both the transformed features and the original features, which amounts to modeling the log odds ratio by

$$\beta_0 + \beta_1 \log \frac{f_1(x_1)}{g_1(x_1)} + \ldots + \beta_p \log \frac{f_p(x_p)}{g_p(x_p)} + \beta_{p+1} x_1 + \ldots + \beta_{2p} x_p \,.$$

This variant of FANS is named FANS2, and it allows features with no marginal power to enter the model in a linear fashion. FANS2 augments the original features with the log-likelihood ratios and helps when a linear decision boundary separates data reasonably well.

## 3 Numerical Studies

### 3.1 Simulation

In simulation studies, FANS and FANS2 are compared with competing methods: penalized logistic regression (PLR, Friedman et al. (2010)), penalized additive logistic regression models (penGAM, Meier et al. (2009)), support vector machine (SVM), regularized optimal affine discriminant (ROAD, Fan et al. (2012)), linear discriminant analysis (LDA), Naive Bayes (NB) and feature annealed independence rule (FAIR, Fan and Fan (2008)).

In all simulation settings, we set $p = 1000$ and training and testing data sample sizes of each class to be 300. Five-fold cross-validation is conducted when needed, and we repeat 50 times for each setting (The relative small number of replications is due to the intensive computation time of penGAM, c.f. Table 2). Table 1 summarizes median test errors for each method along with the corresponding standard errors. This table omits Fisher's classifier (using pseudo inverse for sample covariance matrix), because it gives a test error around 50%, equivalent to random guessing.

**Example 1** *We consider the two class Gaussian settings where* $\Sigma_{ii} = 1$ *for all* $i = 1, \cdots, p$ *and* $\Sigma_{ij} = \rho^{|i-j|}$, $\boldsymbol{\mu}_1 = \mathbf{0}_{1000}$ *and* $\boldsymbol{\mu}_2 = (\mathbf{1}_{10}^T, \mathbf{0}_{990}^T)^T$, *in which* $\mathbf{1}_d$ *is a length* $d$ *vector with all entries 1, and* $\mathbf{0}_d$ *is a length* $d$ *vector with all entries 0. Two different correlations* $\rho = 0$ *and* $\rho = 0.5$ *are investigated.*

This is the classical LDA setting. In view of the linear optimal decision boundary, the nonparametric transformations in FANS is not necessary. Table 1 indicates some efficiency (not much) loss due to the more complex model FANS. However, by including the original features, FANS2 is comparable to the methods (e.g., PLR and ROAD) which learn linear

boundaries of original features. In other words, the price to pay for using the unnecessarily more complex method FANS (FANS2) is small in terms of the classification error.

An interesting observation is that penGAM, which is based on a more general model class than FANS and FANS2, performs worse than our new methods. This is also expected as the complex parameter space considered by penGAM is unnecessary in view of linear optimal decision boundary. Surprisingly, SVM performs poorly (even worse than NB), especially when all features are independent.

**Example 2** *The same settings as Example 1 except the common covariance matrix is an equal correlation matrix, with a common correlation* $\rho = 0.9$.

Same as in Example 1, FANS and FANS2 have performance comparable to PLR and ROAD. Although FAIR works very well in Example 1, where the features are independent (or nearly independent), it fails badly when there is significant global pairwise correlation. Similar observations also hold for NB. This example shows, ignoring correlation among features could lead to significant loss of information and deterioration in the classification error.

**Example 3** *One class follows a multivariate Gaussian distribution, and the other a mixture of two multivariate Gaussian distributions. Precisely,*
*Class 0:* $N((3 \times \mathbf{1}_{10}^T, \mathbf{0}_{p-10}^T)^T, \mathbf{\Sigma}_p)$,
*Class 1:* $0.5 \times N(\mathbf{0}_p, \boldsymbol{I}_p) + 0.5 \times N((6 \times \mathbf{1}_{10}^T, \mathbf{0}_{p-10}^T)^T, \mathbf{\Sigma}_p)$,
*where* $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ *for* $i \neq j$. *Correlations* $\rho = 0$ *and* $\rho = 0.5$ *are considered.*

In this example, Class 0 and Class 1 have the same mean, so the differences are in higher order moments. Table 1 shows that all methods based on linear boundary perform like random guessing, because the optimal decision boundary is highly nonlinear. penGAM is comparable to FANS and FANS2, but SVM cannot capture the oracle decision boundary well even if a nonlinear kernel is applied.

**Example 4** *Two classes follow uniform distributions,*
*Class 0: Unif* $(A)$,
*Class 1: Unif* $(B \backslash A)$,
*where* $A = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 \leq 1\}$ *and* $B = [-1, 1]^p$.

Clearly, the oracle decision boundary is $\{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 = 1\}$. Again, FANS and FANS2 capture this simple boundary well while the linear-boundary based methods fail to do so.

Computation times (in seconds) for various classification algorithms are reported in Table 2. FANS is extremely fast thanks to parallel computing. While penGAM performs similarly to FANS in the simulation examples, its computation cost is much higher. The similarity

in performance is due to the abundance in training examples. We will demonstrate with an email spam classification example that penGAM fails to deliver satisfactory results on small samples.

Table 1: Median test error (in percentage) for the simulation examples. Standard errors are in the parentheses.

| Ex($\rho$) | FANS | FANS2 | ROAD | PLR | penGAM | NB | FAIR | SVM |
|---|---|---|---|---|---|---|---|---|
| 1(0) | 6.8(1.1) | 6.2(1.2) | 6.0(1.3) | 6.5(1.2) | 6.6(1.1) | 11.2(1.4) | 5.7(1.0) | 13.2(1.5) |
| 1(0.5) | 16.5(1.7) | 16.2(1.8) | 16.5(5.3) | 15.9(1.7) | 16.9(1.6) | 20.6(1.7) | 17.2(1.6) | 22.5(1.8) |
| 2(0.5) | 4.2(0.9) | 2.0(0.6) | 2.0(0.6) | 2.5(0.6) | 3.7(0.9) | 43.5(11.1) | 25.3(1.6) | 5.3(1.1) |
| 2(0.9) | 3.1(1.1) | 0.0(0.0) | 0.0(0.0) | 0.0(0.0) | 0.2(1.4) | 46.8(8.8) | 30.2(1.9) | 0.0(0.1) |
| 3(0) | 0.0(0.0) | 0.0(0.0) | 49.6(2.4) | 50.0(1.3) | 0.0(0.1) | 50.4(2.2) | 50.2(2.1) | 31.8(2.4) |
| 3(0.5) | 3.4(0.7) | 3.4(0.7) | 49.3(2.4) | 50.0(1.3) | 3.7(0.8) | 50.0(2.1) | 50.2(2.0) | 19.8(2.4) |
| 4 | 0.0(0.0) | 0.0(0.0) | 28.2(1.8) | 50.0(10.7) | 0.0(0.0) | 41.0(1.1) | 34.6(1.4) | 0.0(0.0) |

Table 2: Computation time (in seconds) comparison for FANS, SVM, ROAD and penGAM. The parallel computing technique is applied. Standard errors are in the parentheses.

| Ex($\rho$) | FANS | FANS(para) | SVM | ROAD | penGAM |
|---|---|---|---|---|---|
| 1(0) | 12.0(2.6) | 3.8(0.2) | 59.4(12.8) | 99.1(98.2) | 243.7(151.8) |
| 1(0.5) | 12.7(2.1) | 3.5(0.2) | 81.3(19.2) | 100.7(89.3) | 325.8(194.3) |
| 2(0.5) | 16.0(3.1) | 4.0(0.2) | 77.6(18.1) | 106.8(90.7) | 978.0(685.7) |
| 2(0.9) | 22.0(4.6) | 4.5(0.3) | 75.7(17.8) | 98.3(83.9) | 3451.1(3040.2) |
| 3(0) | 12.1(2.1) | 3.4(0.2) | 152.1(27.4) | 96.3(68.8) | 254.6(130.0) |
| 3(0.5) | 11.9(2.0) | 3.4(0.2) | 342.1(58.0) | 95.9(74.8) | 298.7(167.4) |
| 4 | 22.4(3.9) | 6.6(0.4) | 264.3(45.0) | 75.1(54.0) | 4811.9(3991.7) |

## 3.2 Real Data Analysis

We study two real examples, and compare FANS(FANS2) with competing methods.

### 3.2.1 Email Spam Classification

First, we investigate a benchmark email spam data set. This data set has been studied by Hastie et al. (2009) among others to demonstrate the power of additive logistic regression models. There are a total of $n = 4,601$ observations with $p = 57$ numeric attributes. The attributes are, for instance, the percentage of specific words or characters in an email, the average and maximum run lengths of upper case letters, and the total number of such letters. To show suitable application domains of FANS and FANS2, we vary the training proportion, from 5%, 10%, 20%, $\cdots$, to 80% of the data while assigning the rest as test set. Splits are repeated for 100 times and we report the median classification errors.

Figure 2 and Table 3 summarize the results. First, we notice that FANS and FANS2 are very competitive when training sample sizes are small. As the training sample size increases, SVM becomes comparable to FANS2 and slightly better than FANS. In general, these three methods dominate throughout different training proportions. The more complex model penGAM failed to yield classifiers when training data proportion is less than 30% due to the difficulty of matrix inversion with the splines basis functions. For larger training samples, penGAM performs better than linear decision rules; however, it is not as competitive as either FANS or FANS2. Also interestingly, Naive Bayes (NB) is the favored method given smallest training sample (5%), and but its test error remains almost unchanged when the sample size increases. In other words, NB's independence assumption allows good training given very few data points, but it cannot benefit from larger samples due to severe model bias.

Table 3: Median classification error (in percentage) on e-mail spam data when the size of the training data varies. Standard errors are in the parentheses.

| % | FANS | FANS2 | ROAD | PLR | penGAM | LDA | NB | FAIR | SVM |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 11.1(2.6) | 10.5(1.1) | 13.6(0.9) | 13.5(1.7) | - | 13.6(1.1) | 10.5(5.0) | 15.6(1.7) | 11.2(0.8) |
| 10 | 8.7(2.4) | 8.5(0.9) | 11.3(0.8) | 10.5(1.1) | - | 11.3(0.9) | 10.7(4.2) | 13.5(0.9) | 9.4(0.7) |
| 20 | 8.0(2.1) | 7.7(0.7) | 10.6(0.6) | 9.0(0.8) | - | 10.3(0.6) | 10.7(5.3) | 12.4(0.7) | 8.1(0.7) |
| 30 | 7.8(1.7) | 7.4(0.5) | 10.3(0.4) | 8.9(0.6) | 9.2(0.6) | 10.1(0.5) | 10.7(4.0) | 11.7(0.4) | 7.4(0.6) |
| 40 | 7.2(2.2) | 6.9(0.5) | 10.1(0.5) | 9.0(0.6) | 8.6(0.5) | 10.0(0.4) | 10.5(5.1) | 11.5(0.6) | 7.0(0.5) |
| 50 | 7.4(2.2) | 7.0(0.5) | 9.9(0.5) | 8.5(0.6) | 8.3(0.5) | 9.9(0.4) | 10.7(4.1) | 11.8(0.6) | 6.9(0.5) |
| 60 | 7.4(2.2) | 6.8(0.5) | 9.8(0.6) | 9.3(0.6) | 7.8(0.6) | 9.5(0.5) | 10.6(4.8) | 11.8(0.7) | 6.5(0.6) |
| 70 | 7.2(1.6) | 6.4(0.6) | 9.5(0.7) | 9.2(0.7) | 7.4(0.7) | 9.4(0.6) | 10.5(4.6) | 11.4(0.7) | 6.4(0.7) |
| 80 | 6.9(1.6) | 6.3(0.7) | 9.4(0.6) | 9.3(0.9) | 7.4(0.8) | 9.2(0.6) | 10.4(4.7) | 11.4(0.8) | 6.3(0.9) |

Figure 2: The median test classification error for the spam data set using various proportions of the data as training sets for different classification methods.

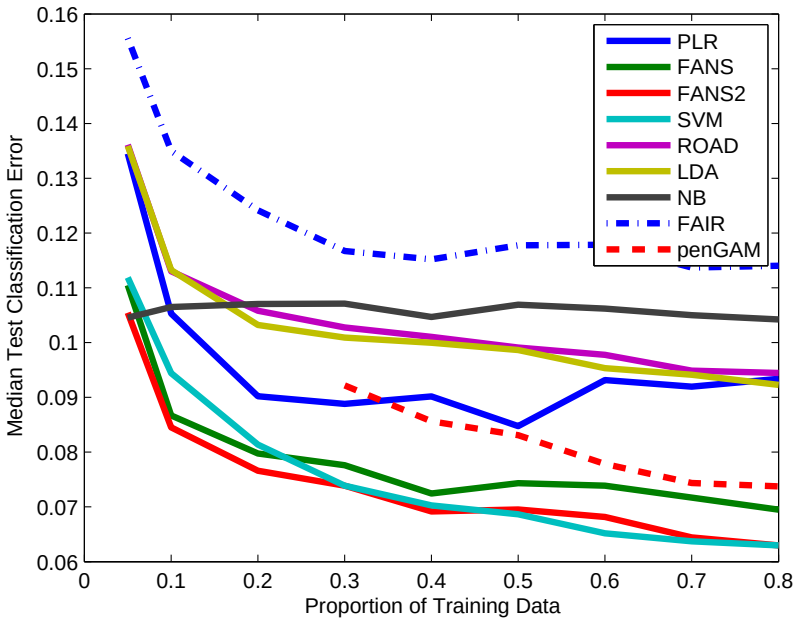



Table 4: Classification error and number of selected genes on lung cancer data.

| | FANS | FANS2 | ROAD | PLR | penGAM | FAIR | NB |
|---|---|---|---|---|---|---|---|
| Training Error | 0 | 0 | 1 | 0 | 0 | 0 | 6 |
| Testing Error | 0 | 0 | 1 | 6 | 2 | 7 | 36 |
| No. of selected genes | 52 | 52 | 52 | 15 | 16 | 54 | 12533 |

### 3.2.2 Lung Cancer Classification

We now evaluate the newly proposed classifiers on a popular gene expression data set "Lung Cancer" (Gordon et al., 2002), which comes with predetermined, separate training and test sets. It contains $p = 12,533$ genes for $n_0 = 16$ adenocarcinoma (ADCA) and $n_1 = 16$ mesothelioma training vectors, along with 134 ADCA and 15 mesothelioma test vectors.

Following Dudoit et al. (2002), Fan and Fan (2008), and Fan et al. (2012), we standardized each sample to zero mean and unit variance. The classification results for FANS, FANS2, ROAD, penGAM, FAIR and NB are summerized in Table 4. FANS and FANS2 achieve 0 test classification error, while the other methods fail to do so.

# 4 Theoretical Results

In this section, an oracle inequality regarding the excess risk is derived for FANS. Denote by $\boldsymbol{f} = (f_1, \cdots, f_p)^T$ and $\boldsymbol{g} = (g_1, \cdots, g_p)^T$ vectors of marginal densities of each class with $\boldsymbol{f}_0 = (f_{0,1}, \cdots, f_{0,p})^T$ and $\boldsymbol{g}_0 = (g_{0,1}, \cdots, g_{0,p})^T$ being the true densities. Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of $(\boldsymbol{X}, Y)$, and the regression function be modeled by

$$P(Y_1 = 1|\boldsymbol{X}_1) = \frac{1}{1 + \exp(-m(\boldsymbol{Z}_1))},$$

where $\boldsymbol{Z}_1 = (Z_{11}, \cdots, Z_{1p})^T$, each $Z_{1j} = Z_{1j}(\boldsymbol{X}_1) = \log f_j(X_{1j}) - \log g_j(X_{1j})$, and $m(\cdot)$ is a generic function in some function class $\mathcal{M}$ including the linear functions. Now, let $\mathcal{Q} = \{q = (m, \boldsymbol{f}, \boldsymbol{g})\}$ be the parameter space of interest with constraints on $m$, $\boldsymbol{f}$ and $\boldsymbol{g}$ be specified later. The loss function we consider is

$$\rho(q) = \rho(m, \boldsymbol{f}, \boldsymbol{g}) = \rho_q(\boldsymbol{X}_1, Y_1) = -Y_1 m(\boldsymbol{Z}_1) + \log(1 + \exp[m(\boldsymbol{Z}_1)]).$$

Let $m_0 = \arg\min_{m \in \mathcal{M}} P\rho(m, \boldsymbol{f}_0, \boldsymbol{g}_0)$. Then the target parameter is $q^* = (m_0, \boldsymbol{f}_0, \boldsymbol{g}_0)$. We use a working model with $m_{\boldsymbol{\beta}}(\boldsymbol{Z}_1) = \boldsymbol{\beta}^T \boldsymbol{Z}_1$ to approximate $m_0$. Under this working model, for a given parameter $q = (m_{\boldsymbol{\beta}}, \boldsymbol{f}, \boldsymbol{g})$, let

$$\pi_q(\boldsymbol{X}_1) = P(Y_1 = 1|\boldsymbol{X}_1) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \boldsymbol{Z}_1)}.$$

With this linear approximation, the loss function is the logistic loss

$$\rho(q) = \rho_q(\boldsymbol{X}_1, Y_1) = -Y_1 \boldsymbol{\beta}^T \boldsymbol{Z}_1 + \log(1 + \exp[\boldsymbol{\beta}^T \boldsymbol{Z}_1]).$$

Denote the empirical loss by $P_n\rho(q) = \sum_{i=1}^n \rho_q(\boldsymbol{X}_i, Y_i)/n$, and the expected loss by $P\rho(q) = E\rho_q(\boldsymbol{X}, Y)$. In the following, we take $\mathcal{M}$ as linear combinations of the transformed features so that $m_0 = m_{\boldsymbol{\beta}_0}$, where

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} P\rho(m_{\boldsymbol{\beta}}, \boldsymbol{f}_0, \boldsymbol{g}_0).$$

In other words, $q_0 = (m_{\boldsymbol{\beta}_0}, \boldsymbol{f}_0, \boldsymbol{g}_0) = q^*$. Hence, the excess risk for a parameter $q$ is

$$\mathcal{E}(q) = P[\rho(q) - \rho(q^*)] = P[\rho(q) - \rho(q_0)]. \tag{4.4}$$

As described in Section 2, densities $\boldsymbol{f}_0$ and $\boldsymbol{g}_0$ are unavailable and must be estimated. Different from the numerical implementation, we now proceed with a modified sampling scheme and procedure that are more friendly for theoretical derivation. Suppose we have labeled samples $\{\boldsymbol{X}_1^+, \cdots, \boldsymbol{X}_{n_1}^+\}$ (used to learn $\boldsymbol{f}_0$) and $\{\boldsymbol{X}_1^-, \cdots, \boldsymbol{X}_{n_1}^-\}$ (used to learn $\boldsymbol{g}_0$; theory carries over for different sample sizes), in addition to an i.i.d. sample $\{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\}$

(used to select features). Moreover, suppose $\{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\}$ is independent of $\{\boldsymbol{X}_1^+, \cdots, \boldsymbol{X}_{n_1}^+\}$ and $\{\boldsymbol{X}_1^-, \cdots, \boldsymbol{X}_{n_1}^-\}$. A simple way to comprehend the above theoretical set up is that the sample size of $2n_1 + n$ has been split into three groups. The notations $P$ and $E$ are regarding the random couple $(\boldsymbol{X}, Y)$. We use the notation $P^n$ to denote the probability measure induced by the sample $\{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\}$, and notation $P^+$ and $P^-$ for the probability measure induced by the labeled samples $\{\boldsymbol{X}_1^+, \cdots, \boldsymbol{X}_{n_1}^+\}$ and $\{\boldsymbol{X}_1^-, \cdots, \boldsymbol{X}_{n_1}^-\}$.

The density estimates $\hat{\boldsymbol{f}} = (\hat{f}_1, \cdots, \hat{f}_p)^T$ and $\hat{\boldsymbol{g}} = (\hat{g}_1, \cdots, \hat{g}_p)^T$ are based on samples $\{\boldsymbol{X}_1^+, \cdots, \boldsymbol{X}_{n_1}^+\}$ and $\{\boldsymbol{X}_1^-, \cdots, \boldsymbol{X}_{n_1}^-\}$:

$$\hat{f}_j(x) = \frac{1}{n_1 h} \sum_{i=1}^{n_1} K\left(\frac{X_{ij}^+ - x}{h}\right) \text{ and } \hat{g}_j(x) = \frac{1}{n_1 h} \sum_{i=1}^{n_1} K\left(\frac{X_{ij}^- - x}{h}\right) \text{ for } j = 1, \cdots, p\,,$$

in which $K(\cdot)$ is a kernel function and $h$ is the bandwidth. Then with these estimated marginal densities, we have an "oracle estimate" $q_1 = (\boldsymbol{\beta}_1, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}})$, where

$$\boldsymbol{\beta}_1 = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} P\rho(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}})\,.$$

It is the oracle given marginal density estimates $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$, and is estimated in FANS by

$$\hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} P_n\rho(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \lambda\|\boldsymbol{\beta}\|_1\,.$$

Let $\hat{q}_1 = (m_{\hat{\boldsymbol{\beta}}_1}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}})$. Our goal is to control the excess risk $\mathcal{E}(\hat{q}_1)$, where $\mathcal{E}$ is defined by (4.4). In the following, we introduce technical conditions for this task.

Let $\boldsymbol{Z}^0$ be the $n \times p$ design matrix consisting of transformed covariates based on the true densities $\boldsymbol{f}_0$ and $\boldsymbol{g}_0$. That is $Z_{ij}^0 = \log f_{0,j}(X_{ij}) - \log g_{0,j}(X_{ij})$, for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. In addition, let $\boldsymbol{Z}^0 = (\boldsymbol{Z}_1^0, \boldsymbol{Z}_2^0, \ldots, \boldsymbol{Z}_n^0)^T$. Also, denote by $|S|$ the cardinality of the set $S$. Denote by $\|\boldsymbol{D}\|_{\max} = \max_{ij}|D_{ij}|$ for any matrix $\boldsymbol{D}$ with elements $D_{ij}$.

**Assumption 1 (Compatibility Condition)** *The matrix $\boldsymbol{Z}^0$ satisfies compatibility condition with a compatibility constant $\phi(\cdot)$, if for every subset $S \subset \{1, \cdots, p\}$, there exists a constant $\phi(S)$, such that for all $\boldsymbol{\beta} \in \mathbb{R}^p$ that satisfy $\|\boldsymbol{\beta}_{S^c}\|_1 \le 3\|\boldsymbol{\beta}_S\|_1$, it holds that*

$$\|\boldsymbol{\beta}_S\|_1^2 \le \frac{1}{n\phi^2(S)}\|\boldsymbol{Z}^0\boldsymbol{\beta}\|^2|S|\,.$$

A direct application of Corollary 6.8 in Bühlmann and van de Geer (2011) leads to a compatibility condition on the estimated transform matrix $\hat{\boldsymbol{Z}}$, in which $\hat{Z}_{ij} = \log \hat{f}_j(X_{ij}) - \log \hat{g}_j(X_{ij})$.

**Lemma 1** *Denote by $\boldsymbol{E} = \hat{\boldsymbol{Z}} - \boldsymbol{Z}^0$ the estimation error matrix of $\boldsymbol{Z}^0$. If the compatibility condition is satisfied for $\boldsymbol{Z}^0$ with a compatibility constant $\phi(\cdot)$, and the following inequalities hold*

$$\frac{32\|\boldsymbol{E}\|_{\max}|S|}{\phi(S)^2} \le 1\,, \text{ for every } S \subset \{1, \cdots, p\}, \tag{4.5}$$

*the compatibility condition holds for $\hat{\boldsymbol{Z}}$ with a new compatibility constant $\phi_1(\cdot) \ge \phi(\cdot)/\sqrt{2}$.*

The Compatibility Condition can be interpreted as a condition that bounds the restricted eigenvalues. The irrepresentable condition (Zhao and Yu, 2006) and the Sparse Riesz Condition (SRC) (Zhang and Huang, 2008) are in similar spirits. Essentially, these conditions avoid high correlation among subsets where signals are concentrated; such high correlation may cause difficulty in parameter estimation and risk prediction.

To help theoretical derivation, we introduce two intermediate $L_0$-penalized estimates. Given the true densities $\boldsymbol{f}_0$ and $\boldsymbol{g}_0$, consider a penalized theoretical solution $q_0^* = (\boldsymbol{\beta}_0^*, \boldsymbol{f}_0, \boldsymbol{g}_0)$, where

$$\boldsymbol{\beta}_0^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} 3P\rho(m_{\boldsymbol{\beta}}, \boldsymbol{f}_0, \boldsymbol{g}_0) + 2H\left(\frac{4\lambda\sqrt{s_{\boldsymbol{\beta}}}}{\phi(S_{\boldsymbol{\beta}})}\right), \tag{4.6}$$

in which $H(\cdot)$ is a strictly convex function on $[0,\infty)$ with $H(0)=0$, $s_{\boldsymbol{\beta}} = |S_{\boldsymbol{\beta}}|$ is the cardinality of $S_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$, and $\phi(\cdot)$ is the compatibility constant for $\boldsymbol{Z}^0$. Throughout the paper, we consider a specific quadratic function[2] $H(v) = v^2/(4c)$ whose convex conjugate is $G(u) = \sup_v\{uv - H(v)\} = cu^2$. Then, equation (4.6) defines an $L_0$-penalized oracle:

$$\boldsymbol{\beta}_0^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} 3P\rho(m_{\boldsymbol{\beta}}, \boldsymbol{f}_0, \boldsymbol{g}_0) + \frac{8\lambda^2 s_{\boldsymbol{\beta}}}{c\phi^2(S_{\boldsymbol{\beta}})}. \tag{4.7}$$

Similarly, with density estimate vectors $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$, we define an $L_0$-penalized oracle estimate $q_1^* = (m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}})$, where

$$\boldsymbol{\beta}_1^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} 3P\rho(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \frac{8\lambda^2 s_{\boldsymbol{\beta}}}{c\phi_1^2(S_{\boldsymbol{\beta}})}. \tag{4.8}$$

To study the excess risk $\mathcal{E}(\hat{q}_1)$, we consider its relationship with $\mathcal{E}(q_1^*)$ and $\mathcal{E}(q_0^*)$.

**Assumption 2 (Uniform Margin Condition)** *There exists $\eta > 0$ such that for all $(m_{\boldsymbol{\beta}}, \boldsymbol{f}, \boldsymbol{g})$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty + \max_{1\le j\le p}\|f_j - f_{0,j}\|_\infty + \max_{1\le j\le p}\|g_j - g_{0,j}\|_\infty \le 2\eta$, we have*

$$\mathcal{E}(m_{\boldsymbol{\beta}}, \boldsymbol{f}, \boldsymbol{g}) \ge c\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2, \tag{4.9}$$

*where $c$ is the positive constant in* (4.7).

The uniform margin condition is related to the one defined in Tsybakov (2004) and van de Geer (2008). It is a type of "identifiability" condition. Basically, near the target parameter $q_0 = (m_{\boldsymbol{\beta}_0}, \boldsymbol{f}_0, \boldsymbol{g}_0)$, the functional value needs to be sufficiently different from the value on $q_0$ to enable enough separability of parameters. Note that we impose the uniform margin condition in both the neighborhood of the parametric component $\boldsymbol{\beta}_0$ and the nonparametric components $\boldsymbol{f}_0$ and $\boldsymbol{g}_0$, because we need to estimate the densities, in addition to the

[2]The following theoretical results can be derived for a generic strictly convex function $H(\cdot)$ along the same lines.

parametric part. A related concept in binary classification is called "Margin Assumption", which was first introduced in Polonik (1995) for densities.

To study the relationship between $\mathcal{E}(\hat{q}_1)$ and $\mathcal{E}(q_1^*)$, we define

$$v_n(\boldsymbol{\beta}) = (P_n - P)\rho(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) \text{ and } W_M = \sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\|\leq M} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_1^*)| .$$

Denote by

$$2\epsilon^* = 3\mathcal{E}(m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \frac{8\lambda^2 s_{\boldsymbol{\beta}_1^*}}{c\phi_1^2(S_{\boldsymbol{\beta}_1^*})} .$$

Set $M^* = \epsilon^*/\lambda_0$ ($\lambda_0$ to specified in Theorem 1) and

$$\mathcal{J}_1 = \{W_{M^*} \leq \lambda_0 M^*\} = \{W_{M^*} \leq \epsilon^*\} .$$

The idea here is to choose $\lambda_0$ such that the event $\mathcal{J}_1$ has high probability.

A few more notations are introduced to facilitate the discussion. Let $\tau > 0$. Denote by $\lfloor\tau\rfloor$ the largest integer strictly less than $\tau$. For any $x, x' \in \mathbb{R}$ and any $\lfloor\tau\rfloor$ times continuously differentiable real valued function $u$ on $\mathbb{R}$, we denote by $u_x$ its Taylor polynomial of degree $\lfloor\tau\rfloor$ at point $x$:

$$u_x(x') = \sum_{|s|\leq\lfloor\tau\rfloor} \frac{(x'-x)^s}{s!} D^s u(x) .$$

For $L > 0$, the $(\tau, L, [-1,1])$-Hölder class of functions, denoted by $\Sigma(\tau, L, [-1,1])$, is the set of functions $u : \mathbb{R} \to \mathbb{R}$ that are $\lfloor\tau\rfloor$ times continuously differentiable and satisfy, for any $x, x' \in [-1,1]$, the inequality:

$$|u(x') - u_x(x')| \leq L|x - x'|^\tau .$$

The $(\tau, L, [-1,1])$-Hölder class of density is defined as

$$\mathcal{P}_\Sigma(\tau, L, [-1,1]) = \left\{ p \,:\, p \geq 0, \int p = 1, p \in \Sigma(\tau, L, [-1,1]) \right\} .$$

**Assumption 3** *Assume that $\boldsymbol{\beta}_1$ is in the interior of some compact set $\mathcal{C}_p$. There exists an $\epsilon_0 \in (0,1)$ such that for all $\boldsymbol{\beta} \in \mathcal{C}_p$, $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{P}_\Sigma(2, L, [-1,1])$, $\epsilon_0 < \pi_{(m_{\boldsymbol{\beta}}, \boldsymbol{f}, \boldsymbol{g})}(\cdot) < 1 - \epsilon_0$.*

**Assumption 4** *$\|\boldsymbol{Z}^0\|_{\max} \leq K$ for some absolute constant $K > 0$, and $\|\boldsymbol{\beta}_0\|_\infty \leq C_1$ for some absolute constant $C_1 > 0$.*

**Assumption 5** *The penalty level $\lambda$ is in the range of $(8\lambda_0, L\lambda_0)$ for some $L > 8$. Moreover, the following holds*

$$\frac{8KL^2(e^\eta/\epsilon_0 + 1)^2}{\eta} \frac{\lambda_0 s_{\boldsymbol{\beta}_1^*}}{\phi_1^2(S_{\boldsymbol{\beta}_1^*})} \leq 1,$$

*where $\eta$ is as in the uniform margin condition.*

Assumption 3 is a regularity condition on the probability of the event that the observation belongs to class 1. Since the FANS estimator is based on the estimated densities, we impose the constraints in a neighborhood of the oracle estimate $\boldsymbol{\beta}_1$ (when using $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$). Assumption 4 bounds the maximum absolute entry of the design matrix as well as the maximum absolute true regression coefficient. Assumption 5 posits a proper range of the penalty parameter $\lambda$ to guarantee that the penalized estimator mimics the un-penalized oracle.

**Assumption 6** *Suppose the feature measurement $\boldsymbol{X}$ has a compact support $[-1,1]^p$, and $f_{0,j}, g_{0,j} \in \mathcal{P}_\Sigma(2, L, [-1,1])$ for all $j = 1, \cdots, p$, where $\mathcal{P}_\Sigma$ denotes a Hölder class of densities.*

**Assumption 7** *Suppose there exists $\epsilon_l > 0$ such that for all $j = 1, \cdots, p$, $\epsilon_l \le f_{0,j}, g_{0,j} \le \epsilon_l^{-1}$. Also we truncate estimates $\hat{f}_j$ and $\hat{g}_j$ at $\epsilon_l$ and $\epsilon_l^{-1}$.*

**Assumption 8**

$$n_1^{\frac{7}{20}-\frac{3}{4}\alpha}(\log(3p))^{\frac{3}{4}}(\log n_1)^{\frac{1}{10}} = o(1)\,,$$

*and,*

$$n_1^{\frac{1}{10}-\alpha}(\log(3p))^{\frac{1}{2}}(\log n_1)^{\frac{2}{5}} = o(1)\,,$$

*for some constant $\alpha > 7/15$.*

Assumption 6 imposes constraints on the support of $\boldsymbol{X}$ and smoothness condition on the true densities $\boldsymbol{f}_0$ and $\boldsymbol{g}_0$, which help control the estimation error incurred by the nonparametric density estimates. Assumption 7 assumes that the marginal densities and the kernel are strictly positive on $[-1,1]^p$. Assumption 8 puts a restriction on the growth of the dimensionality $p$ in terms of sample size $n_1$.

We now provide a lemma to bound the uniform deviation between $\hat{f}_j$ and $f_{0,j}$ for $j = 1, \cdots, p$.

**Lemma 2** *Under Assumptions 6-8, taking the bandwidth $h = \left(\frac{\log n_1}{n_1}\right)^{1/5}$, for any $\delta_1 > 0$, there exists $N_1^*$ such that if $n_1 \ge N_1^*$,*

$$P^{+-}\left(\max_{1\le j\le p}\|\hat{f}_j - f_{0,j}\|_\infty \ge m\right) \le \delta_1\,,\ \text{and}\ P^{+-}\left(\max_{1\le j\le p}\|\hat{g}_j - g_{0,j}\|_\infty \ge m\right) \le \delta_1\,,$$

*for $m = C_2\sqrt{\frac{2\log(3p/\delta_1)}{n_1^{1-\alpha}}}$, and $C_2$ is an absolute constant.*

Denote by

$$\mathcal{J}_2 = \left\{\max_{1\le j\le p}\|\hat{f}_j - f_{0,j}\|_\infty \le \eta/2, \max_{1\le j\le p}\|\hat{g}_j - g_{0,j}\|_\infty \le \eta/2\right\},$$

where $\eta$ is the constant in the uniform margin condition. It is straightforward from Lemma 2 that

$$P^{+-}(\mathcal{J}_2) \ge 1 - \frac{6p}{\exp(\eta^2 n_1^{1-\alpha}/4C_2^2)}\,.$$

The next lemma can be similarly derived as Lemma 2, so its proof is omitted.

**Lemma 3** *Under Assumptions 6-8, taking the bandwidth* $h = \left(\frac{\log n_1}{n_1}\right)^{1/5}$*, for any* $\delta > 0$*, there exists* $N_2^*$ *such that if* $n_1 \geq N_2^*$*,*

$$P^{+-}\left(\|\boldsymbol{E}\|_{\max} \geq m\right) \leq \delta\,,$$

*where* $\boldsymbol{E}$ *is the estimation error matrix as defined in Lemma 1 and* $m = C_3\sqrt{\frac{2\log(3p/\delta)}{n_1^{1-\alpha}}}$*.*

**Corollary 1** *Under Assumptions 6-8, take the bandwidth* $h = \left(\frac{\log n_1}{n_1}\right)^{1/5}$*. On the event* $\mathcal{J}_3 = \left\{\|\boldsymbol{E}\|_{\max} \leq C_3\sqrt{\frac{2\log(3p/\delta)}{n_1^{1-\alpha}}}\right\}$ *(regarding labeled samples) with* $P^{+-}(\mathcal{J}_3) > 1-\delta$*, there exists* $N_2^* \in \mathbb{N}$ *and* $C_4 > 0$ *such that if* $n_1 \geq N_2^*$*,* $|F_{kl}| = |\hat{Z}_{1k} - Z_{1k}^0| \cdot |\hat{Z}_{1l} - Z_{1l}^0| \leq C_4 b_{n_1}$ *uniformly for* $k, l = 1, \cdots, p$*, where* $b_{n_1} = 2\log(3p/\delta)/n_1^{1-\alpha}$*.*

Denote by

$$\mathcal{J}_4 = \left\{32\|\boldsymbol{E}\|_{\max} \max_{S\subset\{1,\ldots,p\}} \frac{|S|}{\phi(S)^2} \leq 1\right\}.$$

On the event $\mathcal{J}_4$, the inequality (4.5) holds, and the compatibility condition is satisfied for $\hat{\boldsymbol{Z}}$ (by Lemma 1). Moreover, it can be derived from Lemma 3 by taking a specific $\delta$,

$$P^{+-}(\mathcal{J}_4) \geq 1 - 3p\exp\{-n_1^{1-\alpha}/(2048C_3^2A_p^2)\}\,,$$

where $A_p = \max_{S\subset\{1,\ldots,p\}} |S|/\phi(S)^2$. Combining Lemma 2 and the uniform margin condition, we see that for given estimators $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$, the margin condition holds for the estimated transformed matrix $\hat{\boldsymbol{Z}}$ involved in the FANS estimator $\hat{\boldsymbol{\beta}}_1$. Following similar lines as in van de Geer (2008) delivers the following theorem, so a formal proof is omitted.

**Theorem 1 (Oracle Inequality)** *In addition to Assumptions 1-8, assume* $\|m_{\boldsymbol{\beta}_1^*} - m_{\boldsymbol{\beta}_0}\|_\infty \leq \eta/2$ *and* $\mathcal{E}(m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}})/\lambda_0 \leq \eta/4$*. Then on the event* $\mathcal{J}_1 \cap \mathcal{J}_2 \cap \mathcal{J}_3 \cap \mathcal{J}_4$*, we have*

$$\mathcal{E}(m_{\hat{\boldsymbol{\beta}}_1}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \lambda\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq 6\mathcal{E}(m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \frac{16\lambda^2 s_{\beta_1^*}(e^\eta/\epsilon_0 + 1)^2}{c\phi_1^2(S_{\beta_1^*})}\,.$$

*In addition, when* $n_1 \geq \max(N_1^*, N_2^*)$ *and under the normalization condition that* $\|Z_{1j}\|_\infty \leq 1$ *for all* $j$*, it holds that*

$$\mathbb{P}(\mathcal{J}_1 \cap \mathcal{J}_2 \cap \mathcal{J}_3 \cap \mathcal{J}_4) \geq 1 - \exp(-t) - 6p\exp\{-\eta^2 n_1^{1-\alpha}/(4C_2^2)\} - \delta - 3p\exp\{-n_1^{1-\alpha}/(2048C_3^2A_p^2)\}\,,$$

*for*

$$\lambda_0 := 4\lambda^* + \frac{tK}{3n} + \sqrt{\frac{2t}{n}(1 + 8\lambda^*)}\,,$$

*where* $\mathbb{P}$ *is the probability with regards to all the samples and*

$$\lambda^* = \sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{3n}\,.$$

Theorem 1 shows that with high probability, the excess risk of the FANS estimator can be controlled in terms of the excess risk of $q_1^*$ when using the estimated density functions $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$ plus an explicit order term. Next, we will study the excess risk of $q_1^*$.

**Assumption 9** *Let $\boldsymbol{Z}_1^0(\boldsymbol{\beta}_1)$ be the subvector of $\boldsymbol{Z}_1^0$ corresponding to the nonzero components of $\boldsymbol{\beta}_1$, and $b_{n_1} = \log(3p/\delta_1)/n_1^{1-\alpha}$. Assume $s_{\boldsymbol{\beta}_1} \leq a_{n_1}$ for some deterministic sequence $\{a_{n_1}\}$, and $a_{n_1} \cdot b_{n_1} = o(1)$. In addition, $0 < C_5 \leq \lambda_{\min}(P\{\boldsymbol{Z}_1^0(\boldsymbol{\beta}_1)\boldsymbol{Z}_1^0(\boldsymbol{\beta}_1)^T\})$, for some absolute constant $C_5$.*

Assumption 9 allows the number of nonzero elements of $\boldsymbol{\beta}_1$ to diverge at a slow rate with $n_1$. Also, it demands a lower bound of the restricted eigenvalue of the sub-matrix of $\boldsymbol{Z}^0$ corresponding to the nonzero components of $\boldsymbol{\beta}_1$.

**Lemma 4** *Let $Q(\boldsymbol{\beta}) = P\rho(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \lambda\|\boldsymbol{\beta}\|_0$, and $\bar{\boldsymbol{\beta}}_1 = \min\{|\boldsymbol{\beta}_{1,j}| : j \in S_{\boldsymbol{\beta}_1}\}$. Under Assumptions 3, 6, 7, 8 and 9, on the event $\mathcal{J}_3$, there exists a $N_3^*$ such that, if $n_1 \geq N_3^*$ and the penalty parameter $\lambda < 0.5C_5\epsilon_0(1-\epsilon_0)\bar{\boldsymbol{\beta}}_1^2$, the $L_0$ penalized solution coincides with the unpenalized version; that is $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$.*

**Theorem 2 (Oracle Inequality)** *In addition to Assumptions 1-9, suppose $4C_1C_4s_{\boldsymbol{\beta}_0}^2 b_{n_1} \leq \lambda_0\eta$, the penalty parameter $\lambda \in (8\lambda_0, \min(L\lambda_0, 0.5C_5\epsilon_0(1-\epsilon_0) \cdot \min_{j:\beta_{1,j}\neq 0}(|\beta_{1,j}|)))$, where $C_5$ is defined in Assumption 9, and $\|m_{\boldsymbol{\beta}_1^*} - m_{\boldsymbol{\beta}_0}\|_\infty \leq \eta/2$. Taking the bandwidth $h = \left(\frac{\log n_1}{n_1}\right)^{1/5}$, on the event $\mathcal{J}_1 \cap \mathcal{J}_2 \cap \mathcal{J}_3 \cap \mathcal{J}_4$ as in Theorem 1, we have*

$$\mathcal{E}(m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) \leq C_1C_4s_{\boldsymbol{\beta}_0}^2 b_{n_1}\,.$$

*Then in view of Theorem 1, we have*

$$\mathcal{E}(m_{\hat{\boldsymbol{\beta}}_1}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) \leq \frac{16\lambda^2 s_{\beta_1}^*(e^\eta/\epsilon_0 + 1)^2}{c\phi_1^2(S_{\beta_1^*})} + 6C_1C_4s_{\boldsymbol{\beta}_0}^2 b_{n_1}\,.$$

From Theorem 2, it is clear that the excess risk of the FANS estimator is naturally decomposed into two parts. One part is due to the nonparametric density estimations while the other part is due to the regularized logistic regression on the estimated transformed covariates. When both the penalty parameter $\lambda$ and the bandwidth $h$ of the nonparametric density estimates $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$ are chosen appropriately, the FANS estimator will have a diminishing excess risk with high probability. Note that one can make explicit $\lambda$ to obtain a bound on the excess risk in terms of the sample sizes $n$ and $n_1$, and the dimensionality $p$.

# 5 Discussion

We propose a new two-step nonlinear rule FANS (and its variant FANS2) to tackle binary classification problems in high-dimensional settings. FANS first augments the original feature space by leveraging flexibility of nonparametric estimators, and then achieves feature selection through regularization (penalization). It combines linearly the best univariate transforms that essentially augment the original features for classification. Since nonparametric techniques are only performed on each dimension, we enjoy a flexible decision boundary without suffering from the curse of dimensionality. An array of simulation and real data examples, supported by an efficient parallelized algorithm, demonstrate the competitive performance of the new procedures.

A few extensions are worth further investigation. Beyond a specific procedure, FANS establishes a general two-step classification framework. For the first step, one can use other types of marginal density estimators, e.g., local polynomial density estimates. For the second step, one might rely on other classification algorithms, e.g., the support vector machine, $k$-nearest neighbors, etc. Searching for the best two-step combination is an important but difficult task, and we believe that the answer mainly depends on the data structures.

We can further augment the features by adding pairwise bivariate density ratios. These bivariate densities can be estimated by the bivariate kernel density estimators. Alternatively, we can restrict our attention to bivariate ratios of features selected by FANS. The latter has significantly fewer features.

Dimensions of data sets (e.g., SNPs) in many contemporary applications could be in millions. In such ultra-high dimensional scenarios, directly applying the FANS (FANS2) approach could cause problems due to high computational complexity and instability of the estimation. It will be beneficial to have a prior step to reduce the dimensionality in the original data. Notable works towards this effort on the theoretical front include Fan and Lv (2008), which introduced the sure independence screening (SIS) property to screen out the marginally unimportant variables. Subsequently, Fan et al. (2011) proposed nonparametric independence screening (NIS), which is an extension of SIS to the additive models.

# 6 Appendix

The appendix contains technical proofs and Lemma 5.

**Proof of Lemma 2** For any $r, m > 0$,

$$\begin{aligned} & P^{+-}\left(\max_{1\le j\le p}\|\hat{f}_j - f_{0,j}\|_\infty \ge m\right) \\ \le\ & e^{-rm}E^{+-}\exp\left(\max_{1\le j\le p} r\|\hat{f}_j - f_{0,j}\|_\infty\right) \\ =\ & e^{-rm}E^{+-}\left(\max_{1\le j\le p}\exp r\|\hat{f}_j - f_{0,j}\|_\infty\right) \\ \le\ & e^{-rm}\sum_{j=1}^{p} E^{+-}\left(\exp r\|\hat{f}_j - f_{0,j}\|_\infty\right). \end{aligned}$$

Since we assumed that all $\hat{f}_j$ and $f_{0,j}$ are uniformly bounded by $\epsilon_l^{-1}$, $\|\hat{f}_j - f_{0,j}\|_\infty$ is bounded by $\epsilon_l^{-1}$ for all $j \in \{1, \cdots, p\}$. This coupled with Lemma 1 in Tong (2013), provides a high probability bound for $\|\hat{f}_j - f_{0,j}\|_\infty$, gives rise to the following inequality,

$$E^{+-}\exp\left(r\|\hat{f}_j - f_{0,j}\|_\infty\right) \le \exp\left(r\sqrt{\frac{\log(n_1/\delta_2)}{n_1 h}}\right) + \exp(r\epsilon_l^{-1})\cdot\delta_2\,,$$

where $\delta_2$ playes the role of $\epsilon$ in Lemma 1 of Tong (2013)(taking constant $C = 1$ for simplicity).

Finding the optimal order for $r$ does not seem to be feasible. So we plug in $r = n_1^{1-\alpha}m$ and $\delta_2 = \exp(-r\epsilon_l^{-1})$, then

$$\begin{aligned} & P^{+-}\left(\max_{1\le j\le p}\|\hat{f}_j - f_{0,j}\|_\infty \ge m\right) \\ \le\ & p\exp(-n_1^{1-\alpha}m^2)\left\{1+\exp\left(n_1^{1-\alpha}m\sqrt{\frac{\log n_1}{n_1 h} + \frac{n_1^{1-\alpha}m\epsilon_l^{-1}}{n_1 h}}\right)\right\} \\ \le\ & p\exp(-n_1^{1-\alpha}m^2)\left\{1+\exp\left[\sqrt{2}n_1^{1-\alpha}m\left(\sqrt{\frac{\log n_1}{n_1 h}} + \sqrt{\frac{m\epsilon_l^{-1}}{n_1^{\alpha}h}}\right)\right]\right\} \\ \le\ & p\exp(-n_1^{1-\alpha}m^2)\left\{1+\exp\left[\sqrt{2}n_1^{1-\alpha}m\left(\frac{\log n_1}{n_1}\right)^{\frac{2}{5}} + \sqrt{2}m^{\frac{3}{2}}\epsilon_l^{-\frac{1}{2}}n_1^{\frac{11}{10}-\frac{3}{2}\alpha}(\log n_1)^{\frac{1}{10}}\right]\right\}, \end{aligned}$$

where in the last inequality we have used the bandwidth $h = \left(\frac{\log n_1}{n_1}\right)^{1/5}$.

The results are derived by taking $m = \sqrt{\frac{2\log(3p/\delta_1)}{n_1^{1-\alpha}}}$ ( so $\delta_1 = 3p\exp(-n_1^{1-\alpha}m^2)$), and by taking Assumption 8. Note that we need to introduce $\alpha > 0$ because the consistency conditions do not hold for $\alpha = 0$. In fact, we need at least $\alpha > 7/15$. Under this assumption, there exists a positive integer $N_1^*$ such that if $n_1 \ge N_1^*$,

$$1+\exp\left[2^{\frac{5}{4}}\epsilon_l^{-\frac{1}{2}}n_1^{\frac{7}{20}-\frac{3}{4}\alpha}(\log(3p/\delta_1))^{\frac{3}{4}}(\log n_1)^{\frac{1}{10}} + 2n_1^{\frac{1}{10}-\alpha}(\log(3p/\delta_1))^{\frac{1}{2}}(\log n_1)^{\frac{2}{5}}\right] \le 3\,.$$

Therefore, for $n_1 \geq N_1^*$,

$$P^{+-}\left(\max_{1\leq j\leq p}\|\hat{f}_j - f_{0,j}\|_\infty \geq m\right) \leq \delta_1, \text{ for } m = \sqrt{\frac{2\log(3p/\delta_1)}{n_1^{1-\alpha}}}.$$

**Lemma 5** *For any vector $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,p})^T$, let $S_{\boldsymbol{\theta}_0} = \{j : \theta_{0,j} \neq 0\}$, and let the minimum signal level be $\bar{\boldsymbol{\theta}}_0 = \min\{|\theta_{0,j}| : j \in S_{\boldsymbol{\theta}_0}\}$. Let $g(\theta_j) = c_j(\theta_j - \theta_{0,j})^2 + \lambda\|\theta_j\|_0$, where $c_j > 0$. If $\lambda \leq c_j\bar{\boldsymbol{\theta}}_0^2$, $g(\theta_j)$ achieves the unique minimum at $\theta_j = \theta_{0,j}$.*

**Proof of Lemma 5** For $\theta_{0,j} = 0$, the result is obvious. For $\theta_{0,j} \neq 0$, we have $j \in S_{\boldsymbol{\theta}_0}$ and

$$\begin{aligned} g(\theta_j) &\geq \lambda\|\theta_j\|_0 I(\theta_j \neq 0) + c_j(\theta_j - \theta_{0,j})^2 I(\theta_j = 0) \\ &= \lambda\|\theta_j\|_0 I(\theta_j \neq 0) + c_j\theta_{0,j}^2 I(\theta_j = 0). \end{aligned}$$

If $\lambda\|\theta_{0,j}\|_0 \leq c_j\bar{\boldsymbol{\theta}}_0^2$,

$$g(\theta_j) \geq \lambda\|\theta_j\|_0 I(\theta_j \neq 0) + \lambda\|\theta_{0,j}\|_0 I(\theta_j = 0) = \lambda\|\theta_{0,j}\|_0.$$

Since $g(\theta_{0,j}) = \lambda\|\theta_{0,j}\|_0$, the lemma follows.

**Proof of Lemma 4** Denote $Q_0(\boldsymbol{\beta}) = P\rho(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}})$. Then we have $\boldsymbol{\beta}_1 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} Q_0(\boldsymbol{\beta})$. Since $\nabla Q_0(\boldsymbol{\beta}_1) = 0$ and

$$\nabla^2 Q_0(\boldsymbol{\beta}) = P\{\hat{\boldsymbol{Z}}_1\hat{\boldsymbol{Z}}_1^T \exp(\hat{\boldsymbol{Z}}_1^T\boldsymbol{\beta})(1+\exp(\hat{\boldsymbol{Z}}_1^T\boldsymbol{\beta}))^{-2}\} \geq \epsilon_0(1-\epsilon_0)P\{\hat{\boldsymbol{Z}}_1\hat{\boldsymbol{Z}}_1^T\} \succeq \mathbf{0}.$$

By Taylor's expansion of $Q_0(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_1$,

$$Q(\boldsymbol{\beta}) = Q_0(\boldsymbol{\beta}_1) + 0.5(\boldsymbol{\beta} - \boldsymbol{\beta}_1)^T\nabla^2 Q_0(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}_1) + \lambda\|\boldsymbol{\beta}\|_0\,, \tag{6.10}$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_1$. Let $\widehat{\boldsymbol{M}} = P\{\hat{\boldsymbol{Z}}_1(\boldsymbol{\beta}_1)\hat{\boldsymbol{Z}}_1(\boldsymbol{\beta}_1)^T\}$, where $\hat{\boldsymbol{Z}}_1(\boldsymbol{\beta}_1)$ is the subvector of $\hat{\boldsymbol{Z}}_1$ corresponding to the nonzero components of $\boldsymbol{\beta}_1$, and $\boldsymbol{M} = P\{\boldsymbol{Z}_1^0(\boldsymbol{\beta}_1)\boldsymbol{Z}_1^0(\boldsymbol{\beta}_1)^T\}$, where $\boldsymbol{Z}_1^0(\boldsymbol{\beta}_1)$ is the subvector of $\boldsymbol{Z}_1^0$ corresponding to the nonzero components of $\boldsymbol{\beta}_1$. Let $\mathbf{F} = \widehat{\boldsymbol{M}} - \boldsymbol{M}$ (a symmetric matrix). From the uniform deviance result of Lemma 3, with probability $1-\delta$ regarding the labeled samples, there exists a constant $C_4 > 0$ such that $|F_{kl}| \leq C_4 b_{n_1}$ uniformly for $k, l = 1, \cdots, s_{\boldsymbol{\beta}_1}$, where $b_{n_1} = 2\log(3p/\delta)/n_1^{1-\alpha}$.

Hence, $\|\mathbf{F}\|_2 \leq \|\mathbf{F}\|_F \leq C_4 s_{\boldsymbol{\beta}_1} b_{n_1} \leq C_4 a_{n_1} b_{n_1}$. For any eigenvalue $\lambda(\widehat{\boldsymbol{M}})$, by the Bauer-Fike inequality (Bhatia, 1997), we have $\min_{1\leq k\leq s_{\boldsymbol{\beta}_1}} |\lambda(\widehat{\boldsymbol{M}}) - \lambda_k(\boldsymbol{M})| \leq \|\mathbf{F}\|_2 \leq C_4 a_{n_1} b_{n_1}$, where $\lambda_k(\boldsymbol{A})$ denotes the $k$-th largest eigenvalue of $\boldsymbol{A}$. In addition, in view of Assumption 9, there exists $k \in S_{\boldsymbol{\beta}_1}$ such that

$$\lambda_{\min}(\widehat{\boldsymbol{M}}) \geq \lambda_k(\boldsymbol{M}) - C_4 a_{n_1} b_{n_1} \geq \lambda_{\min}(\boldsymbol{M}) - C_4 a_{n_1} b_{n_1} \geq C_5 - C_4 a_{n_1} b_{n_1}.$$

Since $a_{n_1}b_{n_1} = o(1)$, there exists $N_3^*(\delta)$ such that when $n_1 > N_3^*(\delta)$, we have $\lambda_{\min}(\widehat{\boldsymbol{M}}) > 0$.

Let $\boldsymbol{\beta}_1^{(1)}$ be the subvector of $\boldsymbol{\beta}_1$ consisting of the nonzero components. Then by (6.10) and Lemma 5 for each $j \in S_{\boldsymbol{\beta}_1}$ with $\lambda < 0.5C_5\epsilon_0(1-\epsilon_0)\bar{\boldsymbol{\beta}_1}^2$, we have

$$\begin{aligned} Q(\boldsymbol{\beta}) &\geq Q_0(\boldsymbol{\beta}_1) + 0.5(C_5 - C_4 a_{n_1} b_{n_1})\epsilon_0(1-\epsilon_0)\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_1^{(1)}\|^2 + \lambda\|\boldsymbol{\beta}\|_0 \\ &\geq Q_0(\boldsymbol{\beta}_1) + \sum_{j \in S_{\boldsymbol{\beta}_1}} \big\{0.5(C_5 - C_4 a_{n_1} b_{n_1})\epsilon_0(1-\epsilon_0)(\beta_j - \beta_{1,j})^2 + \lambda\|\beta_j\|_0\big\}, \quad (6.11) \end{aligned}$$

where $\beta_j$ and $\beta_{1,j}$ are the $j$-th components of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_1$, respectively. For $n_1 \geq N_3^*(\delta)$,

$$Q(\boldsymbol{\beta}) \geq Q_0(\boldsymbol{\beta}_1) + \lambda \sum_{j \in S_{\boldsymbol{\beta}_1}} \|\beta_{1,j}\|_0 = Q_0(\boldsymbol{\beta}_1) + \lambda\|\boldsymbol{\beta}_1\|_0 \,.$$

By (6.10), we have

$$Q(\boldsymbol{\beta}_1) = Q_0(\boldsymbol{\beta}_1) + \lambda\|\boldsymbol{\beta}_1\|_0 \,.$$

Therefore, $\boldsymbol{\beta}_1$ is a local minimizer of $Q(\boldsymbol{\beta})$. It then follows from the convexity of $Q(\boldsymbol{\beta})$ that $\boldsymbol{\beta}_1$ is the global minimizer $\boldsymbol{\beta}_1^*$ of $Q(\boldsymbol{\beta})$.

**Proof of Theorem 2** For simplicity, denote by $\rho(m(\boldsymbol{Z}_1), Y_1)$ the loss function $\rho_q(\boldsymbol{X}_1, Y_1) = -Y_1 m(\boldsymbol{Z}_1) + \log(1 + \exp(m(\boldsymbol{Z}_1))$. Note that

$$\frac{\partial \rho(m(\boldsymbol{Z}_1), Y_1)}{\partial m(\boldsymbol{Z}_1)} = -Y_1 + \frac{\exp(m(\boldsymbol{Z}_1))}{1 + \exp(m(\boldsymbol{Z}_1))} = -Y_1 + \pi_{m, \boldsymbol{f}_0, \boldsymbol{g}_0}(\boldsymbol{X}_1),$$

and

$$\frac{\partial^2 \rho(m(\boldsymbol{Z}_1), Y_1)}{[\partial m(\boldsymbol{Z}_1)]^2} = \frac{\exp(m(\boldsymbol{Z}_1))}{[1 + \exp(m(\boldsymbol{Z}_1))]^2}.$$

By the second order Taylor expansion, we obtain that

$$\begin{aligned} \rho(m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1), Y_1) &= \rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1) + [\partial \rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1)/\partial m_{\boldsymbol{\beta}}(\boldsymbol{Z}_1)](m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1) - m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0)) \\ &\quad + \frac{1}{2}\frac{\partial^2 \rho(m^*, Y_1)}{[\partial m_{\boldsymbol{\beta}}(\boldsymbol{Z}_1)]^2}(m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1) - m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1))^2, \quad (6.12) \end{aligned}$$

where $m^*$ lies between $m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1)$ and $m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0)$. Since

$$P\left[\frac{\partial \rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1)}{\partial m_{\boldsymbol{\beta}}(\boldsymbol{Z}_1)}\right] = 0 \quad (6.13)$$

and $0 < \partial^2 \rho(m^*, Y_1)/[\partial m_{\boldsymbol{\beta}}(\boldsymbol{Z}_1)]^2 < 1$, taking the expectation we obtain that

$$\begin{aligned} |P\rho(m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1), Y_1) - P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1)| &< 0.5P[(m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1) - m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0))^2] \\ &= 0.5P[(\hat{\boldsymbol{Z}}_1^T\boldsymbol{\beta} - (\boldsymbol{Z}_1^0)^T\boldsymbol{\beta}_0)^2]. \end{aligned}$$

Hence, from Corollary 1, on the event $\mathcal{J}_3$,

$$\begin{aligned} |P\rho(m_{\boldsymbol{\beta}_0}(\hat{\boldsymbol{Z}}_1), Y_1) - P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1)| &\leq 0.5\boldsymbol{\beta}_0^T P[(\hat{\boldsymbol{Z}}_1 - \boldsymbol{Z}_1^0)(\hat{\boldsymbol{Z}}_1 - \boldsymbol{Z}_1^0)^T]\boldsymbol{\beta}_0 \\ &\leq C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}, \end{aligned}$$

where $s_{\boldsymbol{\beta}} = |S_{\boldsymbol{\beta}}|$ is the cardinality of $S_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$. Naturally, $P\rho(m_{\boldsymbol{\beta}_0}(\hat{\boldsymbol{Z}}_1), Y_1) \leq P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1) + C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}$.

In addition, by definition of $\boldsymbol{\beta}_1$, $P\rho(m_{\boldsymbol{\beta}_1}(\hat{\boldsymbol{Z}}_1), Y_1) = \min_{\boldsymbol{\beta}} P\rho(m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1), Y_1)$. As a result, $P\rho(m_{\boldsymbol{\beta}_1}(\hat{\boldsymbol{Z}}_1), Y_1) \leq P\rho(m_{\boldsymbol{\beta}_0}(\hat{\boldsymbol{Z}}_1), Y_1)$. Thus, we have

$$P\rho(m_{\boldsymbol{\beta}_1}(\hat{\boldsymbol{Z}}_1), Y_1) \leq P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1) + C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}. \tag{6.14}$$

In addition, by (6.12) and (6.13), for any $\boldsymbol{\beta}$ we have $P\rho(m_{\boldsymbol{\beta}}(\hat{\boldsymbol{Z}}_1), Y_1) \geq P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1)$. Then, setting $\boldsymbol{\beta} = \boldsymbol{\beta}_1$ on the left side leads to

$$P\rho(m_{\boldsymbol{\beta}_1}(\hat{\boldsymbol{Z}}_1), Y_1) \geq P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1). \tag{6.15}$$

Combining (6.14) and (6.15) leads to

$$|P\rho(m_{\boldsymbol{\beta}_1}(\hat{\boldsymbol{Z}}_1), Y_1) - P\rho(m_{\boldsymbol{\beta}_0}(\boldsymbol{Z}_1^0), Y_1)| \leq C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}. \tag{6.16}$$

As a result, we have

$$\mathcal{E}(m_{\boldsymbol{\beta}_1}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) \leq C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}. \tag{6.17}$$

(6.17) combined with Lemma 4 ($\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$) leads to

$$\mathcal{E}(m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) \leq C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}. \tag{6.18}$$

Recall the oracle estimator

$$\boldsymbol{\beta}_1^* = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \Big\{ \mathcal{E}(m_{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \frac{8\lambda s_\beta}{c\phi_1^2(S_\beta)} \Big\}.$$

Then by Theorem 1,

$$\mathcal{E}(\hat{q}_1) = \mathcal{E}(m_{\hat{\boldsymbol{\beta}}_1}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) \leq 6\mathcal{E}(m_{\boldsymbol{\beta}_1^*}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}) + \frac{16\lambda^2 s_{\beta_1^*}(e^\eta/\epsilon_0 + 1)^2}{c\phi^2(S_{\beta_1^*})}. \tag{6.19}$$

Therefore, by (6.18) and (6.19),

$$\mathcal{E}(\hat{q}_1) \leq \frac{16\lambda^2 s_{\beta_1^*}(e^\eta/\epsilon_0 + 1)^2}{c\phi^2(S_{\beta_1^*})} + C_1 C_4 s_{\boldsymbol{\beta}_0}^2 b_{n_1}.$$

# References


Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10** 1471–2105.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19** 563–570.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.*, **101** 119–137.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10** 989–1010.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol.*, **3** Art. 33, 32 pp. (electronic).

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, **106** 1566–1577.

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

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97** 77–87.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36** 2605–2637.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.*, **106** 544–557.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **74** 745–771.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc., Ser. B: Statistical Methodology*, **70** 849–911.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1** 302–332.

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

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62** 4963–4967.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8** 86–100.

Hastie, T. and Tibshirani, R. (1990). *Generalized Addtive Models*. Chapman & Hall/CRC.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer-Verlag Inc.

Huang, P. W., X. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19** 2072–2978.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86** 316–342. With discussion and a rejoinder by the author.

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99** 29–42.

Meier, L., Geer, V. and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist*, **37** 3779–3821.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data . *Bioinformatics*, **18** 39–50.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Annals of Statistics*, **23** 855–881.

Raykar, V., Duraiswami, R. and Zhao, L. (2010). Fast computation of kernel estimators. *Journal of Computational and Graphical Statistics*, **19** 205–220.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, **39** 1241–1265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, **58** 267–288.

Tong, X. (2013). A plug-in approach to anomaly detection. *Journal of Machine Learning Research*, **14** 3011–3040.

Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32** 135–166.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.*, **36** 614–645.

Witten, D. and Tibshirani, R. (2012). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society Series B*, **73** 753–772.

Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C. and Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, **25** 1145–1151.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38** 894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, **36** 1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7** 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101** 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, **15** 265–286.