Regularized Optimal Affine Discriminant (ROAD)

A statistical report on Regularized Optimal Affine Discriminant
Department of Mathematical and Computational Sciences
University of Toronto Mississauga

Names: Kungania Peter.
Student Number: 1002534171
Names: Kintu Brian
Student Number: 1001802034
March 2019 - April 2019

Contents

1 Acknowlegdements

We would like to first thank our statistics professor, Dr. Dehan Kong for being a great and a very passionate educator and for taking all the time to challenge us on a statistical intellectual level. Professor Dehan, you've been very patient with our slow understanding of the class material, and you've always answered every question that we've asked you. We also appreciate the fact that you've always taken the time to show us the mathematics behind different classification approaches, and this has enabled us to continue to appreciate the beauty of mathematics in statistics. On a personal note, your deep and grounded insights about statistics have not only inspired us, but they've also propelled us to appreciate the beauty behind advanced statistical learning especially in this digital age in which data is most certainly the new oil. From the words of Ian Hacking [7], " The quiet statisticians have changed our world; not by discovering new facts or technical developments , but by changing the ways that we reason, experiment and form our opinions", we too believe that Professor Dehan has heighten our understanding of the statistical methods that are used by institutions and corporations to construct a world in which many of us rarely understand.

We also want to thank Jianqing Fan, Yang Feng, and Xin Tong, the three researchers behind a paper, ROAD; a paper that triggered our attention. From the words of Rutherford D. Roger [7], " We are drowning in information and starving for knowlegde", with all the enormous technological advancements that have occurred and how such advancements have had and still do have an inevitable impact on society, we find an interest in the thereotical work of Jianqing, Yang, and Xin's, and this is simply because their work inspires many of us to call into question the effectiveness and efficiency of some of the classical classification approaches developed in such a young science field such as statistics.

Lastly, we want to thank our STA314H5 and STA315H5 TA, Peng Liu, for being so resourceful in explaining all the statistical notions that we still found very difficult to understand. Not to forget, we sincerely give great thanks to the Department of Mathematical and Computational sciences at the University of Toronto Mississauga.

2 Introduction

In this report, we briefly summarize the **regularized optimal affine discriminant (ROAD)**, a new high dimensional classification procedure proposed by Jianquing Fan, Yang Feng and Xin Tong. Jianqing, Yang and Xin argue that in biological applications and in particular microarray studies, in which a group of correlated genes are responsible for clinical outcomes, the ROAD is a very fruitful high dimensional classification procedure than Fisher discriminant rule and the independence rule.

In section 3, we briefly revisit the main idea behind the Fisher discriminant analysis (FDA) and we also briefly discuss the diverging spectrum and noise accumulation as some of the inevitable statistical challenges associated with using Fisher discriminant analysis (FDA) in high dimensional settings. As an alternative of FDA, in this section, we also briefly revisit the main idea behind the features annealed independence rule (FAIR) a special kind of independence rule. As with the issues associated with FDA, we also revisit some of the non-negligible statistical issues associated with using FAIR in high dimensional settings.

In section 4, we thoroughly discuss the regularized optimal affine discriminant (ROAD) and also discuss its variants namely, **screening-based ROAD version one (S-ROAD1)**, and **screening-based ROAD version two (S-ROAD2)** in a two-class setting. In addition, we also break down the mathematical ground-work of ROAD in a multi-class settings.

In section 5, we discuss how we can solve the minimization problem that is presented in section 4 using the constrained coordinate descent. We also break down to the objective function to show how its convexity is used to show that a co-ordinate-wise minimum is also local minimum.

And finally in section 6, we present simulations done with the ROAD method and compare it against its variants S-ROAD1, S-ROAD2 and DROAD. In this section we also try to explain as to how certain a settings affect the performance of ROAD methods and also why these performance of the ROAD is affected in the presented manner.

3 Statistical challenges associated with Fisher discriminant rule and Independence rule in high dimensional settings.

3.1 Main idea behind Fisher Discriminant Analysis (FDA)

With Fisher discriminant Analysis, one wants to find a projection of data points onto a line (or generally one wants to find a projection of data points onto a $P-1$ dimension surface) such that samples from different classes are well separated.

In order to project the data points onto a line or a plane or another dimensional surface, one usually has to solve the eigenvalue problem that is always associated with the implementation of FDA. Solving the eigenvalue problem enables one to find the eigenvectors that are necessary in determining a good data projection.

In a two dimensional setting, FDA is optimal for instance, it uses all the data points to estimate the covariance matrix and hence, it provides more information for classification decision purposes. However, not only does FDA ends up not being robust against outliers in a two-class problem, but also as we discuss in 3.1.1 in high dimensional settings, FDA performs worse than its other competitors such as, logistic regression, independence rule and many other new classification approaches namely, regularized optimal affine discriminant.

3.1.1 Limitations of Fisher discriminant rule

Based on the statistical findings of Bickel and Levina [1], Jianqing, Yang and Xin [3] argue that when the dimension p of the predictor space is very high compared with the sample size, n the Fisher discriminant rule performs very poorly due to diverging spectra. In other words, in high dimensional settings such as in microarray studies, the eigenvalues of the covariance matrix, Σ in the Fisher discriminant function do not converge. Note that when the associated eigenvalue problem cannot be solved, this raises difficulties with maximising the distance between projected class means and also raises difficulties with minimizing the projected class variances. In addition, even with some structural assumptions imposed on the covariance matrix, Σ, when dimension p is greater than sample size n, the pooled sample covariance, $\hat{\Sigma}$ is not necessarily anymore a good estimate of Σ or simply say that the covariance matrix, Σ is not invertible in a $p > n$ setting.

Due to the accumulation of noise in high dimensional settings such as in the microarray settings, in which the dimensionality is frequently in thousands or more, where as the sample size is typically of the order of tens, Fisher discriminant rule leads to poor classification results. In high dimensional settings, FDA

performs worse simply because FDA is very vulnerable to estimation errors that result from having too much noise in the setting. This accumulated noise raises difficulties with estimating population centroids such as, the mean, median and many other population centroids.

3.2 Main idea behind Features Annealed Independence rule (FAIR)

As a special kind of independence rule proposed by Fan and Fan [4], the features annealed independence rule (FAIR) is a high dimensional classification approach that makes classification decisions of samples from different classes by exploiting information from a selected subset of important features for classification.

As with Fisher discriminant rule that has some advantages in low dimensional settings, based on the statistical findings of Dudoit et al [2] features annealed independence rule leads to better classification results when correlations among variable are ignored. This makes FAIR to out-perform Fisher discriminant analysis which as we discussed in 3.1.1 performs worse in high dimensional settings. However, in some studies such as, microarray studies in which correlations are not necessarily negligible, Jianqing, Yang and Xin argue that features annealed independence rule is very incompetent since it dismisses any existing correlation information not realizing that such information may be very critical for classification purposes. As we shall discuss more in details of limitations of FAIR, limitations that end up provoking Jianqing, Yang and Xin to dismiss FAIR in proposition of the regularized optimal affine discriminant (ROAD).

3.2.1 Limitations of the Features Annealed Discriminant Independence rule

Even though in general independence rules circumvent diverging spectra and also to mitigate some accumulated noise in the setting, the fact they use all the features makes to still suffer from some other accumulated noise. This leads to problems with estimating population centroids in high dimensional, which Jianqing, Yang and Xin [3] argue on the basis of the statistical findings of Fan and Fan [4] can be very poor as random guessing.

In high dimensional settings such as, in microarray studies in which correlation between different genes can have a significant impact on the outcome, the independence assumption which acts as if covariance matrix, Σ is diagonal among variables say, genes etc., is very unrealistic in some settings in which correlations are already naturally incorporated. This over reliance on the independence assumption leads to a large misclassification rate simply because some already existing critical information especially on a group of correlated variables has been not taken into any serious account in the setting.

As briefly discussed in 3.1.1 and 3.2.1, due to such inevitable limitations of both Fisher discriminant rule and Independence rule in high dimensional settings such as, microarray settings, Jianqing, Yang and Xin [3] propose the **regularized optimal affine discriminant, (ROAD)**, a new high dimensional classification procedure as we will discuss in section 4, incorporates the covariance structure in the analysis with protection against diverging spectra and significant noise accumulation.

4 ROAD in high dimensional settings.

4.1 Definition of ROAD

Jianqing, Yang and Xin [3] define the **regularized optimal affine discriminant (ROAD)** as a type of classification tool that selects an increasing number of features as the regularization relaxes.

Recall that in LASSO, as we increased the tuning parameter, we ended up left with fewer model parameters since LASSO sets most of the model parameters to zero, and when we decreased the tuning parameter, we retained more model parameters than when the tuning parameter was large. Thus to "relax a regularization", means that if the tuning parameter is reduced or increased, the penalization effect is also reduced or increased.

Since in high dimensional settings such as in microarray studies, correlated genes have an inevitable impact on the outcome, and thus, to fully tackle the classification error that results from not incorporating correlation information into great consideration, Jianqing, Yang and Xin suggest the addition of an L_1- constraint of the form, $\parallel w \parallel_1 \leq c$, for regularization purposes.

In addition to an affine constraint or L_1- constraint, Jianqing, Yang and Xin suggest the addition of a complexity penalty, which as we will see is of the form,$w^T\mu_d = 1$. Therefore, with the addition of an affine constraint and complexity penalty constraint , Jianqing, Yang and Xin propose the following regularized optimal affine discriminant problem,

$$w_c = arg\ min(w^T \Sigma w) \tag{1}$$

where w runs through all $\parallel w \parallel \leq c$, $w^T\mu_d = 1$ and $\sigma_{w_c}(.)$ is the regularized optimal affine discriminant and w is the data projection direction.

To see the mathematics behind Road, lets revisit the following example from [3]: suppose that random variables representing two classes, C_1 and C_2 that follow p-variate normal distributions $X|Y = 1 \sim N_p(\mu_1, \Sigma)$ and $X|Y = 2 \sim N_p(\mu_2, \Sigma)$, and also assume that $P(Y = 1) = \frac{1}{2}$. Since Jianqing, Yang and Xin's [3] mission is to find a good data projection direction, lets call it w, the three researchers suggest that for any linear discriminant rule,

$$\delta_w(X) = I\{w^T(X - \mu_a) > 0\} \tag{2}$$

where $\mu_a = \frac{(\mu_2+\mu_1)}{2}$ and I denotes the indicator function of X with value 1 corresponds to assigning X to class C_1 and 0 corresponds to assigning X to C_2. then, the misclassification rate of the (pseudo) classifier δ_w in Equation (2)

$$W(\delta_w) = \frac{1}{2}\ P_2\{\delta_w(X) = 0\} + \frac{1}{2}\ P_1\{\delta_w(X) = 1\} \tag{3}$$

where $\mu_d = \frac{(\mu_2-\mu_1)}{2}$ and P_i is the conditional distribution of X given its class label i.

Note that the misclassification rate of the linear classifier, $\delta_w(X)$ in Equation (3) can also be rewritten in the following form,

$$W(\delta_w) = 1 - \Phi\{w^T\mu_d/(w^T\Sigma w)^{\frac{1}{2}}\} \tag{4}$$

Note that the classifier in Equation (2) is a linear classifier, meaning that the classification decision between the two classes, C_1 and C_2 is entirely based on using a linear function of inputs.

In contrast with Fisher discriminant rule, the Fisher discriminant classifier in a two classification setting is of the following form,

$$\delta_F(X) = I\{(\Sigma^{-1}\mu_d)^T(X - \mu_a) > 0\} \tag{5}$$

and Fisher's misclassification rate in a two classification setting is of the following form,

$$W(\delta_F) = 1 - \Phi\{(\mu_d^T\Sigma^{-1}\mu_d)^{\frac{1}{2}}\} \tag{6}$$

Unlike the linear classifier in Equation (2) which as we will discuss later is still optimal in high dimensional settings, it is worth stating beforehand that the Fisher classifier in Equation (5) is not optimal in high dimensional settings simply because the covariance matrix, Σ is not invertible and this is raises difficulties in computing the eigenvalues. The unexpected inefficiencies of Fisher discriminant rule in high dimensional settings suggests to us why Jianqing, Yang and Xin propose the regularized optimal affine discriminant.

Jianqing, Yang and Xin [3] argue that minimizing the classification error $W(\delta_W)$ in Equation (4) is the same as maximizing $w^T\mu_d/(w^T\Sigma w)^{\frac{1}{2}}$ in Equation (7), which is equivalent to minimizing $w^T\Sigma w$ in Equation (8) subject to $w^T\mu_d$ = 1. Mathematically, the three researchers argue in the following

$$W(\delta_w) = arg\ max\{w^T\mu_d/(w^T\Sigma w)^{\frac{1}{2}}\} \tag{7}$$

$$W(\delta_w) = arg\ min\{w^T\Sigma w\} \tag{8}$$

subject to $w^T \mu_d = 1$

Theorem 4.1. *Let μ_d be an element of R^p and Σ be a positive definite matrix of dimension $p * p$. Let $w_c = arg\ min\{w^T \Sigma w\}$ where $||w||_1 \leq c$, and $w^T \mu_d = 1$. Then w_c is a continuous piecewise linear function in c [3]*

According to this theorem, the solution path to the optimization problem in Equation (8) is a continuous piecewise linear function in c.

4.2 Two of variants of regularized optimal affine discriminant (ROAD)

According to Jianqing, Yang and Xin [3], the regularized optimal affine discriminant has two variants namely, screening-based version one (S-ROAD1) and the screening-based version two (S-ROAD2). As we will later discuss in section (6), it is worth stating beforehand that the two variants of ROAD serve entirely different purposes in the ROAD method.

4.2.1 Screening-based ROAD version one (S-ROAD1)

This type of the regularized optimal affine discriminant involves first applying the two-sample $t-$test to select any features with the corresponding $t-$test statistic with absolute value larger than the maximum absolute $t-$test statistic value calculated on the permuted data.

In other words, given a set of m features in high dimensional settings, one can select a given number of k out of the m features by using S-ROAD1 implemented in the pre-screening step of ROAD settings. For instance, in Fig.1 in [3], Jianqing, Yang and Xin [3] strongly suggest the implementation of S-ROAD1 in order to be able to keep track of the performance of oracle procedures of sub-Fisher(10 features).

4.2.2 Screening-based ROAD version two (S-ROAD2)

In addition to what S-ROAD1 does namely, selecting a given number of features from a given set of features, as another variant of regularized optimal affine discriminant, S-ROAD2 incorporates an additional variable that is most correlated with an already selected variable.

In contrast with independence rule in which one assumes that there is no correlation among features even in high dimensional settings, with S-ROAD2 implemented, one can incorporate correlations which are most cases are not very negligible in high dimensional settings, such as in microarray settings and proteomics and metabolomics settings. For instance, in Fig.1 in [3], S-ROAD2

is implemented in order to keep track of the performance of the Sub-Fisher, for instance in sub-Fisher of twenty (20) features.

As discussed above, the two variants end up having different feature spaces; for instance, before implementing ROAD, one uses S-ROAD1 to select a given number of k features from a given feature space, and then one uses S-ROAD2 to create another feature space by incorporating additional variables that are most correlated with each of the variables in the already selected feature space. Also, S-ROAD1 keeps track of the performance of the oracle estimator on a smaller feature space while as S-ROAD2 keeps track of the performance of the larger feature space.

Thus, the two variants do serve different purposes of ROAD and as we will also discuss later that in some cases one version will tend to outperform its contrast in a high dimensional setting. It is also worth stating that the difference in the feature space between ROAD and Independence rule results from the implementation of S-ROAD2.

4.3 Extension of regularized optimal affine discriminant to multi class settings.

As discussed in the above sections, the regularized optimal affine discriminant as a new high dimensional classification tool proposed by Jianqing, Yang and Xin, answers problems that Fisher discriminant analysis and independence rule fail to provide answers to in a two-class setting.

It is worth emphasizing that with the exception of the oracle, ROAD outperforms other classification approaches not only by guarding against diverging spectra and significant noise accumulation, but simply by also incorporating the correlation structure and taking advantage of the correlation information. Making the best use of the correlation information in high dimensional studies such as, in microarray studies yet at the same time protecting against diverging spectra and accumulation is what in the end enables ROAD to produce amazing results as compared to FDA and features annealed independence rule in a two-class problem.

To fully pin down the benefits of the regularized optimal affine discriminant in high dimensional settings, Jianqing Yang and Xin, argue that the regularized optimal affine discriminant can also be extended to multi-class Gaussian settings. Unlike in the two-class Gaussian setting in which we have discussed the theoretical and the simulated findings of Jianqing, Yang and Xin about ROAD, in this section of the report, we only break-down the mathematical ground work of ROAD.

Suppose that there are $K-$ classes, and for $j = 1, 2, 3, .., K-1, K$, the jth class has mean μ_j and common covariance, Σ. Denote $\mu_a = K^{-1}\Sigma_{j=1}^{K}\mu_j$ be the overall mean of the features. Now lets find $s \leq K-1$ discriminant co-ordinates $(w_1^*, w_2^*, w_3^*, w_4^*, ..., w_s^*)$ that separate the population centroids $\{\mu_j\}_{j=1}^{K}$ the most in the projected space $S = \text{span}(w_1^*, w_2^*, w_3^*, w_4^*, ..., w_s^*)$. Now suppose that X is the new observation that we want to assign to some class. Then Jianqing, Yang and Xin argue that the population centroids μ_j and the new observation,X are projected onto S and the observation X is assigned to the class whose projected centroid is closest to X onto S. As we accounted for diverging spectrum and noise accumulation in a two-class setting, the same is done in a multi-class setting by regularizing the data projection direction, w in the following way

$$arg\ min\{w^T\Sigma w\} \tag{9}$$

where w runs through all $\parallel w \parallel \leq c$ and subject to $w^T B w = 1$, where $B = \Psi^T\Psi$ and jth column of Ψ^T is $\mu_j - \mu_a$. Jianqing, Yang and Xin argue that the solution of the optimization problem in Equation (9) is the first regularized discriminant co-ordinate $\bar{w_1^*}$. With an additional constraint $\bar{w_1^*}^T\Sigma w = 0$, Jianqing, Yang and Xin argue we get the second regularized discriminant co-ordinate $\bar{w_2^*}$, and other regularized discriminant co-ordinates are obtained in the same way by simply using the previous regularized discriminant co-ordinates and covariance matrix, Σ as parts of a constraint imposed to produce a new regularized discriminant co-ordinate. Now with all $s(\leq K-1)$ regularized discriminant co-ordinates obtained, Jianqing, Yang and Xin argue that the ROAD classifier will be based on the minimum distance to the projected centroids in the $s-$dimensional space that is being by $\{\bar{w_j^*}\}_{j=1}^{s}$

5 Solution of the regularized optimal affine discriminant optimization problem

5.1 Constrained co-ordinate descent method

In order to solve the minimization problem in the regularized affine discriminant classification procedures, Jianqing, Yang and Xin [3] propose a constrained co-ordinate descent algorithm.

Jianqing, Yang and Xin [3] define the constrained co-ordinate descent algorithm to be an algorithm in which the p search directions are just unit vectors $e_1, e_2, ..., e_p$, where e_i denotes the i^{th} element in the standard basis of R^p. To avoid cases in which convergence might not be met, Jianqing, Yang and Xin [1] use unit vectors as search directions in each search cycle.

Unlike other optimization algorithms that don't give explicit formulas for solving optimization problems, Jianqing, Yang and Xin [1] propose the constrained co-ordinate descent algorithm simply as it has an explicit formula for

each co-ordinate update.

From the researchers, Jianqing, Yang and Xin [1], we have the following objective function:

$$g(w_1) = \frac{1}{2}(w_1^T \tilde{\mathbf{w}}_2^T) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ \tilde{\mathbf{w}}_2 \end{pmatrix} + \lambda|w_1| + \lambda|\tilde{\mathbf{w}}_2|_1 + \frac{1}{2}\gamma(\mathbf{w}^T\mu_d - 1)^2$$

which will be a strictly convex function for each of the coordinates on $\mathbb{R}$. By decomposing the objective function into two parts such that we have [1]:

$$g(w_1) = g_1(w_1) + g_2(w_1)$$

, where $g_2(w_1) = \lambda|w_1|$ and

$$g_1(w_1) = \frac{1}{2}(w_1^T \tilde{\mathbf{w}}_2^T) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ \tilde{\mathbf{w}}_2 \end{pmatrix} + \lambda|\tilde{\mathbf{w}}_2|_1 + \frac{1}{2}\gamma(\mathbf{w}^T\mu_d - 1)^2$$

.

From [1] it can be see that $g_1(w_1)$ is strictly convex on $\mathbb{R}$ as it is a quadratic function and in addition, the researchers argue that $g_2(w_1)$ is also convex on $\mathbb{R}$.

The researchers now argue that the co-ordinate descent algorithms will converge to the minimum since the objective function is strictly convex and that its non-diferentiable part, $g_2(w_1)$, is separable [1]. And with this every co-ordinate wise minima will also a local minima since all directional derivatives exist. [1]

6 Simulations of the ROAD

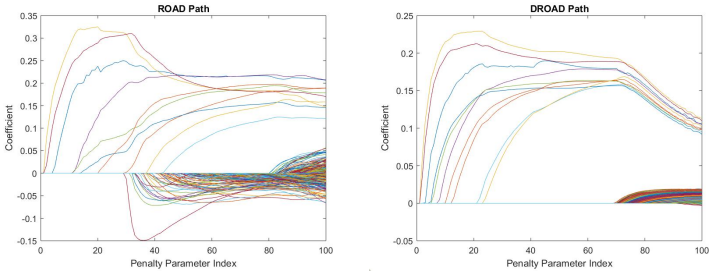
6.1 Results from Simulated Results

In this section we will begin by simulating one of the studies conducted by Jianqing, Yang and Xin [1]. It should be noted that the following simulations are under the same conditions as the simulations done in the paper:

- Number of variables will be 1000 ($p = 1000$)
- Sample size will be held at 300 ($n = 300$)
- Mean vector μ_1 set to 0 for the first class
- Equal correlation setting ($\rho = 0.5$)

The simulations in Figure 1 were done using the MATLAB code that was provided by the researchers along with some changes in order to display the results [5].

The simulations in Figure 1 show how the ROAD and its variants DROAD select more features as we reduce the penalization term (penalization term decreases as penalty parameter index increases). From Figure 1 we can see that the DROAD begins to reduce the number of features earlier than the ROAD method. This may be due to the fact that the DROAD does not contain information about covariance between different features.

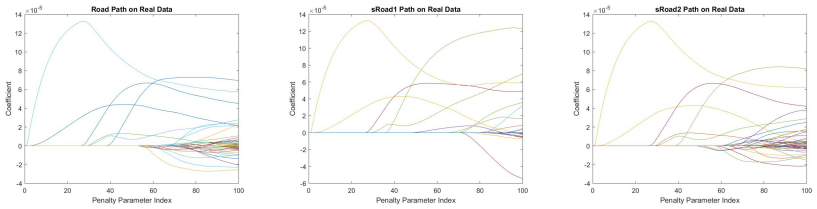


(a) ROAD (b) DROAD

Figure 1: Solution paths for the simulated data set [5]

6.2 Results from real data study

The following simulation is an example of how the ROAD performs when applied to real data. The data comes from a popular gene expression data set[3], leukemia data set [6]. It contains 7129 genes (variables) and only 38 patients (observations), 27 who are diagnosed with acute lmyphoblastic leukemia and 11 diagnosed with acute myeloid leukemia.



(a) ROAD (b) S-ROAD1 (c) S-ROAD2

Figure 2: Solution Paths of ROAD on a Real Data set

Above is the solution path from the leukemia data. This shows that even when the ROAD methods is applied to real data it is able to use variable selection to select features that are most relevant.

Method	Training_Error	Testing_Error	Number_of_Selected_Genes
'ROAD'	1	1	11
'S-ROAD1'	1	1	26
'S-ROAD2'	1	1	46

(a) MATLAB Simulation on Real Data

Method	*Training error*	*Testing error*	*Number of genes selected*
ROAD	0	1	40
S-ROAD1	0	3	49
S-ROAD2	0	1	66
SCRDA	1	2	264
FAIR	1	1	11
NSC	1	3	24
NB	0	5	7129

(b) Researcher's Simulation [3]

Figure 3: Classification Error of ROAD on the Leukaemia Data

Figure 3 (a) shows our test of the different ROAD methods, the classification errors which did not present any interpretive results. Whereas when we look at the researchers' classification errors on Figure 3 (b), we can see that S-ROAD2 outperforms S-ROAD1 and this is due to the incorporation of covariance. From the researchers' results it should also be noted that overall the ROAD method outperforms the other methods.

The table in Figure 4, is to show how different ROAD methods perform in selecting features as we change the correlation setting ρ.

rho_val	ROAD	SROAD1	SROAD2	DROAD
0	10	13	16	10
0.1	93	10	20	10
0.2	99	10	20	10
0.3	136	10	20	10
0.4	111	10	20	10
0.5	136	9	20	10
0.6	153	17	34	10
0.7	234	27	49	10
0.8	186	62	80	2
0.9	149	120	128	2

Figure 4: Number of non-zero coefficients for different correlation settings (rho_val: ρ)

6.3 Conclusions from simulations

From the simulations presented in the paper we have been able to see that the ROAD outperforms other methods in terms of classification error [3] for various real data sets and also for various settings in the simulated data sets.
Even in cases where there are methods that perform just as well as the ROAD [3], it still does a better job as it selects a smaller number of features. Thus making the ROAD a robust classifier for microarray data studies.

7 Conclusion

In this paper we have introduced the ROAD a method that incorporates more features as the penalization term relaxes. We have shown how the ROAD attempts to reduce the classification error by incorporating covariance information without suffering from diverging spectra, accumulation of noise or loss of information, similar to the problems that may be encountered when using the Fisher discriminant rule or the feature annealed discriminant independence rule.

By utilizing the ROAD method we can have better classification results for classification in high dimensional settings.

References

[1] P. Bickel and E.(2004) Levina. *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, 2004.

[2] Fridlyand J. Dudoit, S. and T.P. Speed. *Comparison of discrimination methods for the classification of tumors using gene expression data.* J.Am.Statist.Ass., **97**, 77-87, 2002.

[3] Jianqing F., Yang F., and Xin T. *A road to classification in high dimensional space: the regularized optimal affine discriminant.* Journal of the Royal Statistical Society, 2012.

[4] J. Fan and Y. Fan. *High dimensional classification using features annealed independence rules.* Ann. Statist., **36**, 2605-2637, 2008.

[5] Yang Feng. A road to classification in high dimensional space: the regularized optimal affine discriminant, `https://www.mathworks.com/matlabcentral/fileexchange/40047-a-road-to-classification-in-high-dimensional-space-the-regularized-optimal-affine-discriminant?s_tid=prof_contriblnk`.

[6] Slonim D. K. Tamayo P. Huard C. Gaasenbeek M. Mesirov J. P. Coller H. Loh M. L. Downing J. R. Caligiuri M. A. Bloomfield C. D. Golub, T. R. and E. S. Lander. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 286, 531–537., 1999.

[7] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics, 2001.