Regularization Through Variable Selection and Conditional MLE With Application to Classification in High Dimensions

by

Eitan Greenshtein
SAMSI

Junyong Park
University of Maryland Baltimore County

Guy Lebanon
Purdue University

Technical Report #07-01

Department of Statistics
Purdue University
West Lafayette, IN USA

January 2007

Regularization through Variable Selection and Conditional MLE with application to Classification in High Dimensions

Eitan Greenshtein
SAMSI

Junyong Park
Department of Mathematics and Statistics,
University of Maryland Baltimore County

Guy Lebanon
Department of Statistics, Purdue University

January 23, 2007

Abstract

It is often the case, that high dimensional data consists of only a few informative components. Standard statistical modeling and estimation in such a situation, is prone to inaccuracies due to overfitting, unless regularization methods are practiced. In the context of classification, we propose a class of regularization methods through shrinkage estimators. The shrinkage is based on variable selection coupled with conditional maximum likelihood. Using Stein's unbiased estimator of the risk, we derive an estimator of the optimal shrinkage method within a certain class. A comparison of the optimal shrinkage method in a classification context, with the optimal shrinkage method when estimating a mean vector under a squared loss, is given. The latter problem is extensively studied, but it seems that the results of those studies are not completely relevant for classification. The method is demonstrated and examined on simulated data.

1 Introduction

In this paper we consider the problem of finding a classifier for a response Y, $Y \in \{-1, 1\}$, based on a high dimensional vector $(X_1, ..., X_m)$ of explanatory variables. Here, by high dimensionality we mean $m \gg n$ where n is the size of the training set.

We consider linear classifiers, or predictors $\hat{Y}$ for Y of the form

$$\hat{Y} = \text{sign}\left(\sum_{j=1}^{m} a_j X_j + a_0\right), \tag{1}$$

where $a_0, a_1, ..., a_m$ are constants.

Accurately estimating the constants $a_0, a_1, ...a_m$in high dimensions requires special care. In such cases there is a need for regularization in order to avoid overfitting. The regularization that we suggest involves:
(a) Variable/Model-selection
(b) Correction of selection-bias through conditional maximum likelihood.

Specifically, when a model is selected at stage (a), its parameters are estimated at stage (b) by a conditional MLE, which takes the selection into account. The likelihood is based on assuming a multivariate normal distribution of the vector $(X_1, ..., X_m)$ conditional on the value of Y. Formally, we assume $X_j|Y = 1 \sim N(\mu_j, 1)$ and $X_j|Y = -1 \sim N(\tau_j, 1)$, both independently.

The normality assumption is robust when both m and n are large and when considering linear classifiers that involve linear combinations of many explanatory variables. This follows from the central limit theorem. The role of large m and large n in applying the C.L.T is different. Large m implies that $\sum a_j X_j$ will be close to normal when a_j are comparable in size even if the individual X_j are not normal as in Lindeberg C.L.T; large n implies that Z_j defined in the sequel through averages of independent X_j^i, $i = 1, ..., n$ are close to normal.

When searching for a good set $a_1, ..., a_m$ it is obvious that one may assume w.l.o.g that $\sum_{j=1}^m a_j^2 = 1$. Then the optimal choice is the vector $(a_1, ..., a_m)$, that maximizes $|\sum a_j\mu_j - \sum a_j\tau_j|$. Note that the optimal choice of $a_1, ..., a_m$ is the same regardless of the misclassification loss (or prediction loss). In order to see it observe that $\sum a_j X_j \sim N(\sum a_j\mu_j, 1)$ conditional that $Y = 1$ and $\sum a_j X_j \sim N(\sum a_j\tau_j, 1)$ conditional that $Y = -1$. Hence an optimal choice of $a_1, ...a_m$ is such that $|\sum_j a_j\mu_j - \sum_j a_j\tau_j|$ is maximized. A formal argument showing that the optimal choice of $a_1, ..., a_m$ is the same regardless of the misclassification loss is through the theory of comparison of experiments, implying that the experiment that consists of the distributions $N(\theta_1, 1)$ and $N(\theta_2, 1)$, dominates the experiment that consists of the distributions $N(\theta_1', 1)$ and $N(\theta_2', 1)$ if and only if $|\theta_1 - \theta_2| \geq |\theta_1' - \theta_2'|$. See Lehmann (1986) p-86, for some basic theory on comparison of experiments and some additional references.

Consider first the case where μ_j and τ_j are known, $j = 1, ..., m$. Denoting $\Delta_j = \mu_j - \tau_j$ the optimal choice for a_j, $j = 1, ..., m$ is

$$a_j^0 = \frac{\Delta_j}{\|\Delta\|}, \tag{2}$$

where $\|\Delta\| = \sqrt{\sum_l \Delta_l^2}$ is the l_2 norm of $\Delta = (\Delta_1, ..., \Delta_m)$. In practice Δ_j are unknown, thus we can not find the optimal a_j^0. A natural approach is to estimate a_j^0 through maximum likelihood.

Suppose a training set of size n is available for which $Y = -1$, and a further set of size n for which $Y = 1$ is available. Let $(Z_1, ..., Z_m)$ be the vector obtained when subtracting the mean of those two n-size samples. Then the resulting random vector

$$Z_j \sim N(\Delta_j, 2/n), \ \ j = 1, ...m. \tag{3}$$

is the MLE estimator for Δ_j.

The naive estimation of $\mathbf{a_j^0}$ through m.l.e. The m.l.e estimator of a_j^0 is:

$$A_j^0 = \frac{Z_j}{||Z||},$$

where $Z = (Z_1, ..., Z_m)$. In order to demonstrate its inefficiency consider asymptotics where $n \to \infty$, $m = m(n)$ and $m \gg n$. Denoting $D = \sum \Delta_j^2$, the quantity

$$V \equiv \sum \frac{Z_j}{||Z||} \times \Delta_j \equiv \sum A_j^0 \Delta_j$$

becomes

$$V = \frac{D + O_p(\sqrt{D/n})}{\sqrt{D + O_p(\sqrt{D/n}) + (2m/n) + O_p(\sqrt{m}/n)}}.$$

The above is straightforward when writing $Z_j = (\Delta_j + \epsilon_j)$, where $\epsilon_j \sim N(0, 2/n)$ are independent. For any fixed D or more generally when $D = o(\sqrt{m/n})$, we have $V = o_p(1)$. Note!, $V = o_p(1)$, implies that asymptotically, the corresponding classifier is equivalent to random guessing. Furthermore, the value of V corresponding to the optimal set $a_j = a_j^0 \ j = 1, ...m$, is $\sqrt{D}$. As explained above, given $a_1, ..., a_m$, the larger the value of V, the better the resulting optimal classifier is, when confined to classifiers which are functions of $\sum a_j X_j$ (regardless of the prediction loss).

The above asymptotic consideration, where m increases and D is fixed, or, more generally, when D is small relative to $\sqrt{m/n}$, reflects a situation of sparsity, where most Z_j have mean (nearly) zero. As a result most explanatory variables are (nearly) irrelevant for our classification task. We will elaborate on this point in the next section. More sophisticated estimators of a_j^0, based on careful regularizations, will yield much better classifiers (i.e., with much larger V). Our result, about asymptotic equivalence between random guessing and the resulting estimator when estimating a_j by m.l.e, is very much related to Theorem 1 of Fan and Fan (2007). Our result was obtained independently.

The last example is brought to demonstrate the need for regularization. A popular method of regularization is by model/variable-selection. When no prior information is known a natural model selection procedure will select the variables whose corresponding $|Z_j|$ are large. Such a method, could include all the variables corresponding to j such that $|Z_j| > C$, for an appropriate threshold C. Once a model is selected, a standard routine is to perform m.l.e based on the selected model. In the context of estimating a sparse high dimensional vector of means such a method is also termed 'hard threshold'. The pioneering research, on the magnitude of the threshold in such a context when estimating a vector of means, is Donoho and Johnstone (1994), (1995) and Foster and George (1994). When rescaling so that the variance of Z_j is 1, a 'universal threshold' suggested in those papers is $C = \sqrt{2\log(n)}$. An extensive research on the 'right' threshold has been conducted since those mentioned studies. See recent results and further references in Johnstone and Silverman (2005). Donoho and Johnstone suggested a modification of hard threshold by a soft threshold, i.e., where the mean of Z_j is estimated by $\text{sign}(Z_j)(|Z_j| - C)_+$. One advantage of the soft threshold is that its smoothness enables evaluation of its performance for example through Stein's unbiased estimator of the risk.

This paper develops such concepts in the context of classification. The analog of a hard threshold is the (unconditional) MLE. An analog of the soft threshold is our conditional MLE, which also defines a smooth shrinkage estimator. Shrinkage by conditional MLE is also appealing in non-classification problems, for example estimation of means. We choose to emphasize the classification problem because it is seldom studied in terms of thresholding and shrinkage estimation. We find an advantage of the conditional MLE relative to unconditional MLE as shown in our numerical study in the next section.

As in Donoho and Johnstone (1995) we arrive in Section 3, through Stein's method of unbiased estimator of the risk, to an analog of 'unbiased estimation of the risk'. This is done for various regularizations (i.e., various choices of thresholds), which is helpful in approximating the optimal threshold.

An important message (though not surprising when thinking of it) is that the optimal threshold in classification, may behave dramatically different from the optimal threshold in the context of estimating the mean vector under a square loss. Consequently, many more variables Z_j may be selected, in comparison to the case where the purpose is estimation. Hence the extensive research on the latter problem is not completely relevant for classification. The difference between the optimal thresholds in estimation versus classification may be seen in Table 1 of Section 3. Future research is needed to determine whether model selection and conditional MLE is indeed the right approach in classification problems.

Finally, we mention a recent work on high dimensional classification by Bickel and Levina (2005). Their setup is also of multivariate normal explanatory variables, but under their formulation the major problem becomes the estimation of the unknown covariance matrix, rather than inference concerning the vector of means as in our formulation. They investigated the naive regularization method called 'naive Bayes' which assumes independence of the explanatory variables. They show that such an assumption is often not very harmful and the resulting procedures are not too bad relative to the optimal. Their result suggests that studying a model with independent X_j, as we do, is of interest. Note that when assuming a diagonal covariance matrix, the assumption of known variance (w.l.o.g $\sigma_j^2 = 1$) is mild when n is large, since the variance may be estimated from the data.

2 Regularization through Subset Selection and Conditional MLE

The discussion in the introduction suggests that a reasonable approach is to choose a_j which are estimates of $a_j^0 = \frac{\Delta_j}{||\Delta||}$, $j = 1, ..., m$, under some regularization procedure.

The regularization method suggested next has two components. First, selecting only a size k subset of explanatory variables, out of $X_1, ..., X_m$ thereby decreasing the variability due to estimating the corresponding Δ_j. Second, compensate for the selection-bias, introduced by the above selection.

When there is no prior information about the relevance and importance of the explanatory variables, a most natural subset selection is of the explanatory variables that correspond to indices j for which $|Z_j| > C$. Assuming, w.l.o.g $|Z_1| \geq ... \geq |Z_k| \geq C \geq ... \geq |Z_m|$, we select $X_1, ..., X_k$, as the explanatory variables. Here C is a tuning parameter that defines a collection of regularization methods.

Another related model-selection setup is when the signals are known to be positive, i.e., $\Delta_j \geq 0$. In such a case the natural variable selection is of the form: select variable j, iff the corresponding Z_j, satisfy $Z_j > C$ for an appropriate C.

Notice, that for a sparse setup, as we have in mind, C will be typically large in both models and corresponding variable selection methods. Thus, the MLE when conditioning on the event $\{|Z_j| > C\}$ is practically the same as the MLE when conditioning on $\{Z_j > C\}$; see Figure 2 in the sequel and further elaboration bellow. Thus, using either conditioning and selection methods yield practically the same results.

We will use the latter model and variable selection method, (i.e., model with $\Delta_j \geq 0$ and selection through $Z_j > C$). This is since the resulting equations are simpler and so is the presentation of the ideas. Yet, the same treatment and ideas work for selection through $|Z_j| > C$.

Given a value C, let $\{G_\Delta\}$ be the family of distributions parameterized by Δ and defined by a $N(\Delta, \sigma^2)$ random variable, conditional that it is greater than C. The value of σ^2 that is relevant for us is $2/n$, see (3). Then, $\{G_\Delta\}$ is an exponential family.

Consider first the conditional MLE for Δ_j conditional on $\{Z_j > C\}$, without assuming that $\Delta_j > 0$!

Given an observation Z, distributed G_Δ, it is easy to check that the MLE estimator $\hat{\Delta}$ for Δ is the solution of

$$Z = \hat{\Delta} + \sigma \cdot h((C - \hat{\Delta})/\sigma) \tag{4}$$

where $h(a) = \frac{\phi(a)}{1-\Phi(a)}$ and ϕ and Φ are he density and the cumulative distribution of a standard normal distribution.

For any combination of C and σ, there is a corresponding MLE $\hat{\Delta}(Z, C, \sigma)$, given an observation Z. The following proposition gives the relation between the MLE corresponding to various values of C and σ.

Proposition 1: $\hat{\Delta}(Z, C, \sigma) = C + \sigma\hat{\Delta}(\frac{Z-C}{\sigma}, 0, 1)$.

Proof $\hat{\Delta}(Z, C, \sigma)$ is a solution of $\frac{Z-\hat{\Delta}(Z,C,\sigma)}{\sigma} - h(\frac{C-\hat{\Delta}(Z,C,\sigma)}{\sigma}) = 0$. After slight modifications, we have

$$\frac{Z-C}{\sigma} - \frac{\hat{\Delta}(Z,C,\sigma)-C}{\sigma} - h\left(-\frac{\hat{\Delta}(Z,C,\sigma)-C}{\sigma}\right) = 0.$$

Denote $Z^* = \frac{Z-C}{\sigma}$ and $\Delta^* = \frac{\hat{\Delta}(Z,C,\sigma)-C}{\sigma}$, then the above is $Z^* - \Delta^* - h(-\Delta^*) = 0$, which means Δ^* is the maximum likelihood estimator given Z^* when $C = 0$ and $\sigma = 1$. By strict convexity of conditional likelihood, MLE is unique, i.e., $\hat{\Delta}$ is one to one function of Z. From this uniqueness, $\frac{\hat{\Delta}(Z,C,\sigma)-C}{\sigma} = \Delta^* = \hat{\Delta}(Z^*,0,1) = \hat{\Delta}(\frac{Z-C}{\sigma},0,1)$ and $\hat{\Delta}(Z,C,\sigma) = \sigma\hat{\Delta}(\frac{Z-C}{\sigma},0,1) + C$.

The function $\hat{\Delta}(Z,0,1)$ is illustrated in Figure 1.

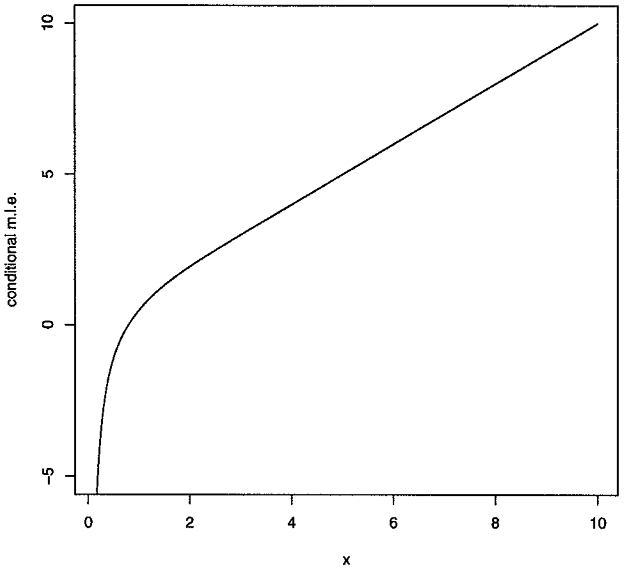

Figure 1: Conditional maximum likelihood estimator when $C = 0$ and $\sigma = 1$.

Now!, consider our model when assuming that $\Delta_j \geq 0$. Then, the conditional MLE for Δ_j conditional on $\{Z_j > C\}$ is:

$$\delta_j = \max(\hat{\Delta}_j, 0). \tag{5}$$

Figure 2 shows the difference between the conditional MLE when conditioning on $|Z| > C$, and the conditional MLE when conditioning on $Z > C$. As C increases, the results become similar.

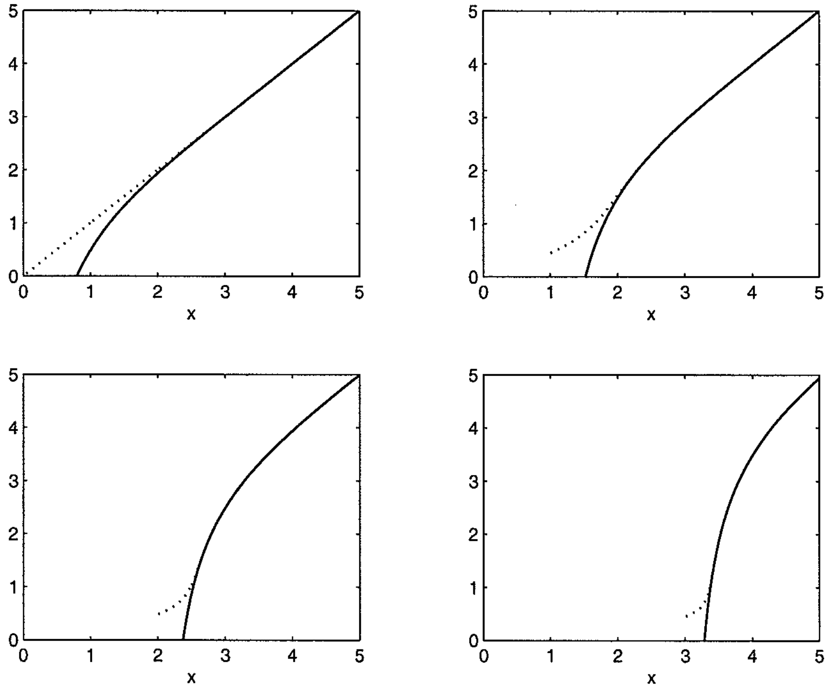

Figure 2: Solid lines represent $\hat{\delta} = \max(\hat{\Delta}, 0)$ for likelihood conditioned on $Z > c$, and real lines represent conditional MLE based on $|Z| > c$. Upper left and right panel are for $c = 0$ and $c = 1$, and lower left and right panel are for $c = 2$ and $c = 3$.

Finally, our suggested conditional MLE estimators for a_j^0 $\hat{a}_j^0$, $j = 1, ..., k$ are:

$$\hat{a}_j^0 = \frac{\delta_j}{||\delta||}, \tag{6}$$

where $\delta = (\delta_1, ..., \delta_m)$.

Similarly to the above one obtain the unconditional MLE as:

$$\hat{\hat{a}}_j^0 = \frac{\mathcal{Z}_j}{||\mathcal{Z}||},$$

where $\mathcal{Z}_j = Z_j \cdot I(|Z_j| > C)$ for an indicator function I. Note that the notations above suppress the dependence of $\hat{\hat{a}}_j^0$ and $\hat{a}_j^0$ on C. We further introduce the notations

$$V(C) = \sum \hat{a}_j^0 \Delta_j,$$

and

$$\mathcal{V}(C) = \sum \hat{\hat{a}}_j^0 \Delta_j.$$

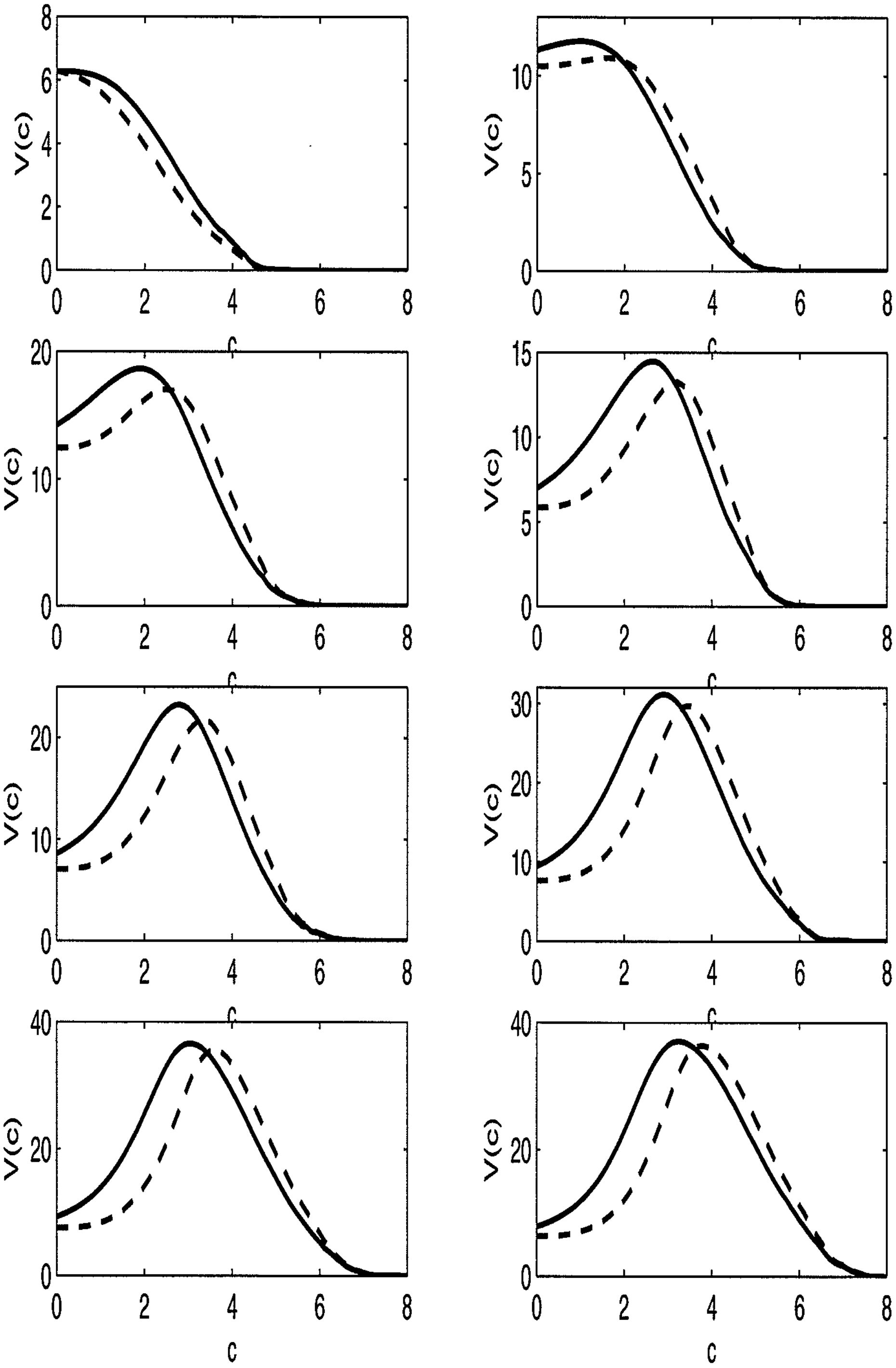

Figure 3: The graphs represent $\mathcal{V}(C)$ (dashed) and $V(C)$ (dotted) for $m = 10^5$ and $\sigma = 1$ v.s. C. Graphs in the first row are for (Δ, l)=(1,2000) and (1.5,1500), the second row for (2,1000) and (2.5, 300), the third row for (3,250) and (3.5,200) and the fourth row for (4,150) and (4.5, 100)

In the following graphs, we demonstrate the performance of conditional MLE and the unconditional MLE in a simulation study, through the pair of functions $V(C)$ and $\mathcal{V}(C)$. We study various combinations of l and Δ, where l variables are normal with mean Δ and variance 1, and $m-l$ variables have mean 0 and variance 1, e.g., $EZ_j = \Delta,\ j = 1, ..., l$, $EZ_j = 0\ j = l{+}1, ..., m$. We present the following combinations of l and Δ where $m = 10^5$. Figure 2 shows the simulation results. Each curve in each graph is based on average of 100 simulations.

Remark 1: We may assume w.l.o.g that by rescaling Z_j have variance 1. Yet, as explained in the introduction, if X_j have variance 1 under the original scale then the variance of Z_j is $2/n$ under that scale. In this section we rescale so that the variance of Z_j is 1 and n is not explicitly mentioned. Yet, given a concrete value of n, the scale of the above graphs should be multiplied by $\sqrt{2/n}$ in order to be interpreted under the original scale for which the variance of X_j is 1.

Discussion. Figure 3 indicates that typically $V_0 = \max_C V(C)$ is larger than $\mathcal{V}_0 = \max_C \mathcal{V}(C)$. Yet, it is larger by 10% or less and sometimes even comparable or (very) slightly smaller.

Assuming, that we may find the optimal C under each procedure and that also the optimal a_0 is found, let us examine the advantage of the conditional MLE procedure when V_0 is 10% larger than $\mathcal{V}_0$. Suppose we want equal misclassification errors, then the values are $1 - \Phi(V_0/2)$ and $1 - \Phi(\mathcal{V}_0/2)$ correspondingly. Suppose under the relevant scale $V_0 = 4.4$ and $\mathcal{V}_0 = 4$, then the advantage in terms of misclassification errors is 1.8% versus 2.3%. It is some advantage but not a major one. Obviously it is less impressive for $V_0 < 4.4$ while it might be more impressive for $V_0 > 4.4$.

In practice however the optimal C is not given. The above graphs indicate that errors in approximating the optimal C will have more severe effect in the case of the unconditional MLE. Moreover, for the conditional MLE, we have a good procedure for approximating the optimal C, as demonstrated in the following section. Those facts make conditional MLE further advantageous when comparing with the unconditional one.

3 Choosing the Appropriate Regularization Parameter C.

In this section we will consider the problem of choosing the regularization parameter C. For every C there are corresponding $\delta_j = \delta(Z_j, C, \sigma)$, for a given data set $Z_1, ..., Z_j$, and a given σ. The optimal C is:

$$\mathrm{argmax}_C \frac{\sum \delta_j \cdot \Delta_j}{\|\delta\|} \equiv \mathrm{argmax}_C V(C).$$

The optimal C can not be found since the values $\Delta_1, ..., \Delta_m$, are unknown. Note that a naive approach where we plug in δ_j for Δ_j trivially yields $C = 0$ as the optimizer. As may be seen in the numerical examples of the previous section, the choice $C = 0$ could be very poor due to overfitting. The formal explanation about naive estimation of a_j^0 is given in Section 1.

We consider the use of Stein's unbiased estimator of the risk in obtaining a good estimation of $V(C)$ for the purpose of approximating the optimal C. A naive way for such an estimation is the use of a validation test. An important advantage of Stein's unbiased estimation method is that it does not use a validation set thereby enabling the use of a larger train set.

Stein's unbiased estimator of the risk. In the expression for $V = V(C)$, the denominator is given and the unknown quantity $\sum \delta_j \cdot \Delta_j$ should be estimated. The method suggested, is based on Stein's unbiased estimation of the risk, applied to the exponential family $\{G_\Delta\}$, see Brown (1986) p-99. The idea is to introduce a function of the data, denoted U, such that:

$$E_{\Delta_j} U(Z_j) \equiv E_{\Delta_j} U_j = E_{\Delta_j} \delta_j \cdot \Delta_j.$$

The obvious advantage of the left hand side over the right hand side is that the expression U involves only the data and does not involve the unknown parameter Δ_j, hence $\sum_j U_j$ is an unbiased estimator of the quantity of interest $\sum \Delta_j \cdot \delta_j$.

We denote, for a fixed C and σ,

$$\delta_j' = \frac{d}{dZ_j} \delta(Z_j, C, \sigma),$$

and similarly denote $\hat{\Delta}'$.

Note that for Z such that $\hat{\Delta}(Z) \geq 0$ we have $\hat{\Delta}'(Z) = \delta'(Z)$, while $\delta'(Z) = 0$ for Z such that $\hat{\Delta}(Z) < 0$.

¿From (5) we get

$$\Delta'(Z) = \frac{1}{1 - h'((C - \hat{\Delta})/\sigma)}.$$

We further denote

$$U_j = \delta_j Z_j - \sigma^2 \delta_j'.$$

Lemma 1: $E_\Delta U_j = E_\Delta \Delta_j \delta_j$

Proof: The proof is straightforward and is based on the principle of Stein's unbiased estimator of the risk, see Brown (1986) p-99. We apply the technique on the exponential family of distributions parameterized by Δ, obtained when conditioning that a $N(\Delta, 1)$ variable is greater than C.

Lemma 1 motivates us to estimate the quantity $V(C) = \sum \delta_j \Delta_i / ||\delta||$ by the following $\hat{V}(C)$, where

$$\hat{V}(C) = \frac{\sum_{[j|Z_j > C]} U_j}{||\delta||}. \tag{7}$$

The relationship between $V(C)$ and $\hat{V}(C)$ is illustrated in Figure 4. Since the agreement seem to be very good it is a good practice to select

$$\hat{C} = \mathrm{argmax}_C \hat{V}(C),$$

as the regularization parameter.

Estimation versus classification. Table 1 below, shows optimal C for different combinations of (Δ, l) under classification, in comparison with those for estimation under a squared loss. Those optimal values are obtained through simulations in various combinations. In both cases of classification and estimation, we examine the class of conditional MLE procedures parameterized by C, and the optimal value of C in each case is reported in Table 1. The values were found through simulation.

It should be emphasized, that the corresponding optimal values of C in classification versus estimation may be very different, e.g., when $(\Delta, l) = (1, 2000)$ and $(1.5, 1500)$. Under square loss, when signals are weak, we choose large C, so most of the estimates for Δ_i are 0, but in classification problem a small C is chosen ($C = 0$ in the case $(\Delta, l) = (1, 2000)$), so many (even all) of the variables are selected.

(Δ, l)	C under classification	C under squared loss
(1,2000)	0.00(0.00)	4.12 (0.28)
(1.5,1500)	0.97(0.19)	4.39 (0.30)
(2, 1000)	1.92(0.15)	4.15 (0.72)
(2.5, 300)	2.59(0.19)	3.55 (0.32)
(3, 250)	2.79(0.10)	3.08 (0.12)
(3.5, 200)	2.92(0.09)	2.98 (0.09)
(4, 150)	3.07(0.08)	3.01 (0.08)
(4.5, 100)	3.25(0.09)	3.16 (0.10)

Table 1: Average of 100 optimal C under $V(C)$, versus the optimal C under a squared loss. The numbers in () indicate standard deviations.

4 Discussion

The proposed method does not really have an iterative training stage as the ML problem is one dimensional which is extremely fast and can also be tabulated. Support vector machines, on the other hand, require quadratic programming with the number of variables scaling up linearly (depending on the number of support vectors) with the number of examples. This means that our method is useful in large scale problems where a fast solution is desirable.

For the case of correlated data, we consider known covariance. Then, if we scale all the variables to have the same variance, we should (i) select all the variables with corresponding $Z_i > c$, estimate their mean using conditional MLE (ii) use the linear transformation on the selected variables, for which the new variables are independent, (iii) estimate the mean of the transformed variables based on the estimates in (i), and proceed as before. The problem of unknown covariance matrix is beyond the scope of this paper.

Some previous studies concerning regularization in discriminant analysis are Campbell(1980), Friedman(1989) and Hastie et al.(1995). The regularization in those papers is based on inverse of the covariance matrix, but under high dimensional problem such as 10,000 variables, this procedure is numerically infeasible. Another regularization direction is assuming independence, namely naive bayes, in Bickel and Levina (2004). This independence assumption as well as variable selection through conditional MLE brings us to our main work on regularization in high dimensional problem.

Conditional mle is a less arbitrary method of soft shrinkage compared to some other shrinkage methods. Our framework is general, it is based on specific modeling assumptions. Other modeling assumption (for example non-normal data) would lead to analogous different shrinkage procedures.

Acknowledgment. Crucial discussions with Anirban DasGupta are gratefully acknowledged.

References

Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives where there are many more variables than observations. *Bernoulli* **10**, No 6. 989-1010.

Brown, L.D. (1986). Fundamentals of statistical exponential families, with applications in statistical decision theory. IMS, Hayward, CA.

Campbell, N.A. (1980). Shrunken Estimation in Discriminant and Canonical Variable Analysis. *Applied Statistics*, **29**, No.1, 5-14

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-455.

Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *JASA* **90**, 4, 1200-1224.

Fan, J. and Fan, Y.(2007). High dimensional classification using features annealed independence rules. Manuscript.

Foster, D.P. and George, E.L. (1994). The risk inflation criterion for multiple regression. *Ann. Stat.* **22**, 1947-1975.

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association, 84*, 165-175

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of the Statistics*, **23**, 73-102

Johnstone, I.M. and Silverman, B.W (2005). Empirical Bayes selection of wavelet thresholds. *Ann.Stat.* **33**, No 4, 1700-1752.

Lehmann, E. L. (1986). Testing Statistical Hypothesis, 2^{nd} edition. Wiley & Sons

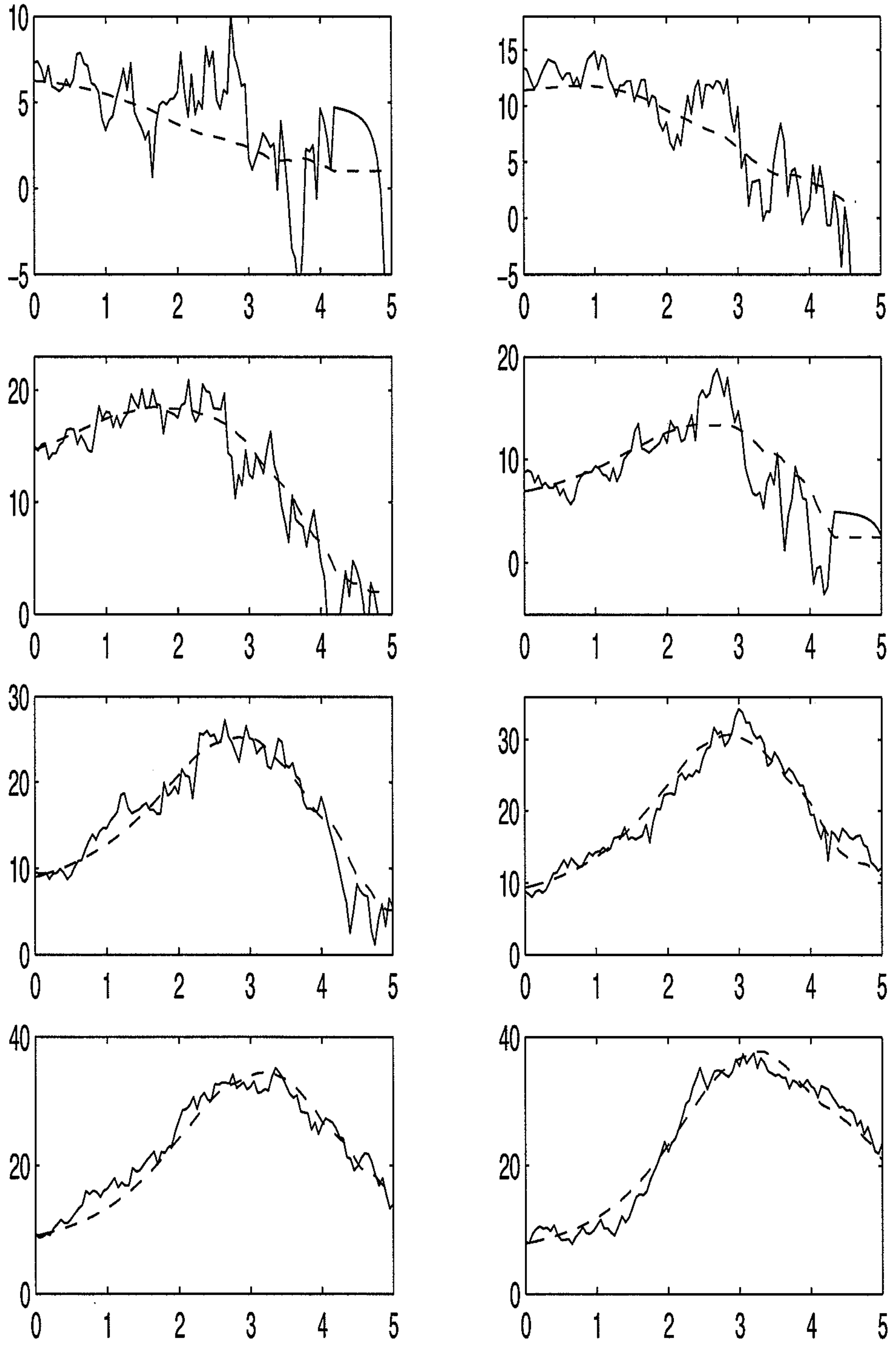

Figure 4: The graphs illustrate $V(C)$ (dashed) vs. Stein's unbiased estimator $\hat{V}(C)$ (solid). Graphs in the first row are for (Δ, l)=(1,2000) and (1.5,1500), the second row for (Δ, l)=(2, 1000) and (2.5,300), the third row for (3,250) and (3.5, 200) and the fourth row for (4,150) and (4.5, 100)