2011

Bayesian Model Selection For Classification With Possibly Large Number Of Groups

Justin Kyle Davis
University of Central Florida

BAYESIAN MODEL SELECTION FOR CLASSIFICATION WITH POSSIBLY LARGE NUMBER OF GROUPS

by

JUSTIN KYLE DAVIS
B.S. Mathematics, B.S. Economics, Duke University, 2004
M.S. Mathematics, University of Central Florida, 2009

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Mathematics
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2011

Major Professor:
Marianna Pensky

ABSTRACT

The purpose of the present dissertation is to study model selection techniques which are specifically designed for classification of high-dimensional data with a large number of classes. To the best of our knowledge, this problem has never been studied in depth previously. We assume that the number of components p is much larger than the number of samples n, and that only few of those p components are useful for subsequent classification. In what follows, we introduce two Bayesian models which use two different approaches to the problem: one which discards components which have "almost constant" values (Model 1) and another which retains the components for which between-group variations are larger than within-group variation (Model 2). We show that particular cases of the above two models recover familiar variance or ANOVA-based component selection. When one has only two classes and features are *a priori* independent, Model 2 reduces to the Feature Annealed Independence Rule (FAIR) introduced by Fan and Fan (2008) and can be viewed as a natural generalization to the case of $L > 2$ classes. A nontrivial result of the dissertation is that the precision of feature selection using Model 2 improves when the number of classes grows. Subsequently, we examine the rate of misclassification with and without feature selection on the basis of Model 2.

For DFE.

ACKNOWLEDGMENTS

I would like to thank Dr. Marianna Pensky for her instruction, both mathematical and otherwise, support, and eternal patience, especially during (what we realize now was) an entirely unnecessary battle with dairy. In particular, I am grateful for the introduction to asymptotic analysis. In addition, she has *tried* to teach me how to navigate the less mathematical aspects of having a mathematical career; what I have not learned is not her fault.

I would also like to thank Dr. Will Crampton, who gave me not only the most interesting computational puzzle I have yet to encounter but also the opportunity to say that I work with electric fish and can be found, if one is patient, on Wikipedia; I can now describe my work without putting laymen to sleep.

I would like to include in general the members of the departments of Mathematics and Statistics at UCF. In particular, Dr. Jian-Jian Ren, whose probability theory course made much of what follows possible. In addition, an offhand comment by Dr. James Schott in a linear algebra class resolved, years later, some calculations below which I thought were impossible. I suppose we should all thank Cholesky.

Of course, my family. I wonder if my parents know that to this day I cannot leave a bookstore without buying *two* books; they made an addict of their son and this is the result. Finally, I need to thank my husband Diego. There is more than can be said here; I have a list, but it is too large even for the margin.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

It is a well-established result that addition of "uninformative" dimensions in data, by which we mean any dimensions which do not improve the accuracy of a generic classifier, eventually makes classification impossible even with just two groups or classes. Additional uninformative dimensions are especially problematic when the total number of dimensions exceeds the number of samples in a data set, since geometrical methods (e.g. the principal component analysis) no longer admit unique solutions. This is called the high-dimension, low sample size (HDLSS) paradigm. The usual goal, then, is the replacement of $\mathbf{D}$ by $\mathbf{D}'$, a new set of dimensions $n \times p'$ with $p' \leq n$. Here, we develop two new models, consider the accuracy of selection in the second model, and consider the classification error of a simple classifier.

1.1 Classification in General

The general problem of classification is to assign a class identity to an individual. We characterize an individual by a p-dimensional vector $\mathbf{d} \in \Re^{\mathbf{p}}$. In practice, one is given $\{\mathbf{d^i}\}_{\mathbf{i=1}}^{\mathbf{n}}$, a set of n observed vectors from L possible classes. Here we organize the vectors into

$$\mathbf{D} = [\mathbf{d^1}, \ldots, \mathbf{d^n}]^\top$$

an $n \times p$ matrix. The columns of $\mathbf{D}$ are indexed by $\mathbf{d_i}$ and the rows, representing individuals, by $\mathbf{d^i}$. For convenience, we will assume that the n_1 individuals from class 1 come first in $\mathbf{D}$,

then the n_2 from class 2, etc., up to the n_l individuals from class L. One then constructs a classifier

$$\hat{c} : \Re^p \rightarrow [1, L]$$

such that $\hat{c}(\mathbf{d^i})$ is by some measure close to l if $\mathbf{d^i}$ is from class l; e.g. one might choose to minimize the misclassification rate. One hopes that the classifier will generalize outside of the training set; i.e. given a new observation $\mathbf{d}$ from independent of $\{\mathbf{d^i}\}_{\mathbf{i=1}}^{\mathbf{n}}$, it would be useful if $\hat{c}(\mathbf{d}) = \mathbf{l}$.

We note that in almost all cases the majority of the work is in the conditioning of $\hat{c}$ on the training set; *evaluation* of $\hat{c}$ will be relatively computationally cheap. Furthermore, it is not necessarily the case that a perfect classifier exists; the individuals might not be discriminable on the basis of the representation chosen or, in fact, discriminable at all. The class structure is taken as given but it is worthwhile to acknowledge that the problem of clustering, or creation of classes, is not itself closed. Finally, even when the individuals *might* admit a perfect classifier on the basis of the representation chosen, there is no guarantee that the training set $\{\mathbf{d^i}\}_{\mathbf{i=1}}^{\mathbf{n}}$ properly represents the classes. Roughly, it is harder to construct a classifier than evaluate one.

1.1.1 Construction of a Simple Classifier

To motivate the ideas of feature selection, we examine a simple classification problem. Recall that we have n training vectors $\{\mathbf{d_i}\}_{\mathbf{i=1}}^{\mathbf{n}}$ organized into $\mathbf{D} = [\mathbf{d_1}, \ldots, \mathbf{d_n}]^\top$. Then let $\mathbf{Y_j} \in \Re^{\mathbf{L}}$

be such that $\mathbf{Y_{ji}} = \mathbf{1}$ if $\mathbf{d_j}$ is in class i, 0 otherwise; i.e. $\mathbf{Y_j}$ is an indicator for the class of individual j. Let

$$\mathbf{Y} = [\mathbf{Y_1}, \ldots, \mathbf{Y_n}]^\top$$

be the $n \times l$ matrix of indicator variables. We consider the regression

$$\mathbf{Y} = \mathbf{DB}$$

where $\mathbf{B}$ is an $p \times l$ matrix of regression coefficients. If $\mathbf{B}$ is found exactly, we can construct

$$\hat{f}(\mathbf{d}) = (\mathbf{dB})^\top$$

$$\hat{c}(\mathbf{d}) = \text{argmax}_{\mathbf{l} \in [\mathbf{1}, \mathbf{L}]} \hat{\mathbf{f}}_\mathbf{l}(\mathbf{d})$$

i.e. $\mathbf{d}$ is assigned to the class for which the predicted indicator is the highest. However, this regression problem rarely admits exact solutions, so we seek to find a matrix $\hat{\mathbf{B}}$ such that

$$\hat{\mathbf{Y}} = \mathbf{D}\hat{\mathbf{B}}$$

approximates $\mathbf{Y}$ in some sense. The most common approximation is the least-squares approximation; i.e. we seek to solve

$$\min_{\hat{\mathbf{B}}} ||\mathbf{Y} - \mathbf{D}\hat{\mathbf{B}}||_\mathbf{2}$$

which, if the problem has a unique solution, yields

$$\hat{\mathbf{B}} = (\mathbf{D}^\top \mathbf{D})^{-\mathbf{1}} \mathbf{D}^\top \mathbf{Y}.$$

This last equation merits examination as it takes us directly into the core of the problem. We note that $\mathbf{D} = [\mathbf{d_1}, \ldots, \mathbf{d_n}]^\top$ is $n \times p$ so that $\mathbf{D}^\top \mathbf{D}$ is $p \times p$ but

$$\text{rank}(\mathbf{D}^\top \mathbf{D}) \leq \min(\text{rank}(\mathbf{D}^\top), \text{rank}(\mathbf{D})) \leq \min(\mathbf{n}, \mathbf{p}).$$

If $n \geq p$, i.e. we have more training individuals than dimensions, then it can occur that $\text{rank}(\mathbf{D}^\top\mathbf{D}) = \mathbf{p}$ and a unique $\hat{\mathbf{B}}$ exists. However, if the columns of $\mathbf{D}$ are linearly dependent, $\text{rank}(\mathbf{D}^\top\mathbf{D}) < \mathbf{p}$ and an infinite number of $\hat{\mathbf{B}}$ minimize the least-squares error.

Finally, if $n < p$, we have that $\text{rank}(\mathbf{D}^\top\mathbf{D}) < \mathbf{p}$ necessarily so that $\mathbf{D}^\top\mathbf{D}$ is never invertible; i.e. if the dimensionality of the data is greater than the number of samples, no unique least-squares approximation to the classification problem exists.

1.1.2 The Necessity of Dimension Reduction

The problem we consider here, namely that many analyses fail when the number of training samples is less than the dimensionality of the data ($n < p$) will be common to all approaches to classification. It manifests in two ways. Above we have seen an algebraic example - a problem no longer admits a unique solution. However, statistical concerns (e.g. estimators which are not consistent if $p \gg n$) are at least equally prohibitive.

As an example, Bickel and Levina[9] demonstrate that a particular Bayesian classifier outperforms the Fisher linear discriminant rule under broad conditions, especially when the dimensionality grows faster than the number of observations.

Specifically, let us try to discriminate between two classes with p-normal distributions $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. If all the parameters are known and we have a new observation $\mathbf{d}$, then the optimal classifier (in the sense of classification error) is given by: $\mathbf{d}$ is assigned

to class 1 if

$$\log \frac{f_1(\mathbf{d})}{f_2(\mathbf{d})} = \boldsymbol{\Delta}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{d} - \boldsymbol{\mu}) > \mathbf{0}$$

where

$$\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \qquad \boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1).$$

In general, however, the parameters will not be known and we will have some estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Delta}}$ of $\boldsymbol{\Delta}$. If we use the usual MLE values, the rule obtained is Fisher's linear discriminant.

However, the authors suggest a new rule, the "naive Bayes" classifier, using $\hat{\boldsymbol{\Sigma}}' = \text{diag}(\hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}}$ is the MLE. Necessarily, $\hat{\boldsymbol{\Sigma}}'$ will not be singular unless one of the dimensions is entirely constant. The authors then demonstrate that asymptotically, specifically as $p \to \infty$, Fisher's linear discriminant is always no better than guessing but the naive Bayes classifier, while never worse, is optimal in certain situations. However, we will examine a result of Fan and Fan[22] who show that any projection method, including this naive Bayes classifier, is eventually no better than guessing since estimates of the classwise means become inconsistent.

These are the problems of high dimension, low sample size (HDLSS) and have become more apparent as our ability to measure and retain large amounts of data has improved. The usual goal, then, is the replacement of $\mathbf{D}$ by $\mathbf{D}'$, a new set of dimensions $n \times p'$ with $p' \leq n$. We call this dimension reduction and find, despite the number and variety of dimension techniques, that there are three major categories of algorithms: synthetic dimension

reduction, subset selection, and feature selection. We review the representative algorithms in these categories and then describe a new method in the third.

1.2 Existing Methods

1.2.1 Synthetic Dimension Reduction

Again, the goal of dimension reduction is the replacement of $\mathbf{D}$ by $\mathbf{D}'$, a data set of more favorable dimensionality. In synthetic dimension reduction, each of the new columns of data is a function of the old columns; none of the original data might be retained. The notable advantage of synthetic dimension reduction algorithms is that the new vectors can be designed to have properties not present in the original data set.

1.2.1.1 Projection Methods

Principal Component Analysis Many synthetic dimension reduction techniques are based on the idea of Pearson's Principal Component Analysis (PCA), first described in 1901 [45]. Given $\mathbf{D}$, where here we assume the empirical mean has been subtracted from each column, we construct the singular value decomposition of $\mathbf{D}$

$$\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$$

where $\mathbf{U}$ $(n \times p)$ is orthogonal and contains the eigenvectors of $\mathbf{DD}^\top$, and $\mathbf{V}$ $(p \times p)$ is orthogonal and contains the eigenvectors of $\mathbf{D}^\top\mathbf{D}$, and $\mathbf{\Lambda}$ $(p \times p)$ is diagonal with *singular values* arranged so that

$$\mathbf{\Lambda_{11}} \geq \mathbf{\Lambda_{22}} \geq \ldots \geq \mathbf{0}.$$

Then the principal components of $\mathbf{D}$ are given by the columns of the $n \times p$ matrix

$$\mathbf{P} = \mathbf{DV} = \mathbf{U\Lambda V}^\top\mathbf{V} = \mathbf{U\Lambda}.$$

We can view each individual's principal components individually; i.e. $\mathbf{P_i} = \mathbf{d^i V}$ yields a $1 \times p$ vector of principal components for individual i. We can interpret this as a projection of $\mathbf{d^i}$ onto a subspace of $\Re^P$ spanned by the columns of $\mathbf{V}$. We note that

$$\mathbf{P}^\top\mathbf{P} = (\mathbf{U\Lambda})^\top\mathbf{U\Lambda} = \mathbf{\Lambda U}^\top\mathbf{U\Lambda} = \mathbf{\Lambda^2}$$

i.e. the principal components of $\mathbf{D}$ are orthogonal, a property very likely not originally present in $\mathbf{D}$ itself. Also,

$$\mathbf{PV}^\top = \mathbf{U\Lambda V}^\top = \mathbf{D}$$

so that the transformation is invertible; i.e. no information has been lost. In fact, since $\mathbf{D}$ is the same size as $\mathbf{P}$ no dimension reduction has taken place.

To assist in dimension reduction we seek an interpretation of the singular values $\mathbf{\Lambda_{ii}}$. We note that $\mathbf{DD}^\top$ is an $n \times n$ covariance matrix for the individuals in $\mathbf{D}$ and have that

$$\mathbf{DD}^\top = (\mathbf{U\Lambda V}^\top)(\mathbf{U\Lambda V}^\top)^\top = \mathbf{U\Lambda V}^\top\mathbf{V\Lambda U}^\top = \mathbf{U\Lambda^2 U}^\top$$

so that

$$\mathbf{\Lambda^2} = \mathbf{U}^\top\mathbf{DD}^\top\mathbf{U}.$$

Now,

$$\text{tr}(\mathbf{\Lambda^2}) = \text{tr}(\mathbf{U}^\top \mathbf{D}\mathbf{D}^\top \mathbf{U}) = \text{tr}(\mathbf{D}\mathbf{D}^\top) = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{d^i})^\top \mathbf{d^i} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \text{Var}(\mathbf{d^i})$$

and

$$\text{tr}(\mathbf{\Lambda^2}) = \sum_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{\Lambda_{ii}^2}$$

so that

$$\sum_{i=1}^{n} \text{Var}(\mathbf{d^i}) = \sum_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{\Lambda_{ii}^2}$$

i.e. we can view each singular value $\mathbf{\Lambda_{ii}}$ as a measure of the variance in $\mathbf{D}$ which is retained by the ith principal component.

Therefore, if we discard those columns of $\mathbf{P}$ which have low $\mathbf{\Lambda_{ii}}$, we can reduce the dimension of $\mathbf{P}$ while retaining most of the variance originally observed in $\mathbf{D}$. In fact, a number of algorithms have arisen which differentiate themselves solely on the number of dimensions retained after a principal component analysis. Most commonly, one might retain p' dimensions, where p' is the smallest which satisfies

$$\sum_{i=1}^{p'} \mathbf{\Lambda_{ii}^2} \geq \alpha \sum_{\mathbf{i=1}}^{\mathbf{n}} \text{Var}(\mathbf{d^i})$$

for some $\alpha \in (0, 1)$.

High-Dimension, Low Sample Size PCA Key in the calculation of the PCA is $\mathbf{\Sigma} = \mathbf{D}^\top \mathbf{D}$, a matrix which, recalling the assumption that each column has mean 0, contains the covariances of the columns of $\mathbf{D}$. For example, the square roots of the eigenvalues of this matrix are the singular values and its eigenvectors are the columns of $\mathbf{V}$. We note that if

$n < p$, i.e. there are fewer individuals in the training set than dimensions in $\mathbf{D}$, then $\mathbf{\Sigma}$ will have 0 as an eigenvalue of multiplicity at least $p - n$ so that $\mathbf{\Sigma}$ is singular. This fact in itself is not problematic as we usually do not need the inverse of $\mathbf{\Sigma}$. Even if this does become necessary, modifications are possible; for instance, Torokhti and Friedland [56] use pseudoinverses to construct a generic PCA, the best weighted linear estimator of the data of a given rank.

Generalization and Alternatives to the PCA Even though PCA is used in dimension reduction, it is not by itself a dimension reduction technique. It can be modified, however, to discard dimensions of $\mathbf{D}$ before returning the principal components. For example, some have used "shrinkage" methods, discounting or discarding the contribution of some of the dimensions in the estimation of the covariance matrix.[17], [38], [50]

Shen and Huan [52] describe a family of algorithms in which low-rank approximations of the SVD are used to construct estimators of the data which are functions of relatively few dimensions. Specifically, one solves

$$\min_{\mathbf{u},\mathbf{v}} ||\mathbf{D} - \mathbf{u}\mathbf{v}^\top|| + \mathbf{P}(\mathbf{v}; \lambda)$$

where $\mathbf{u}$ is $n \times 1$, $\mathbf{v}$ is $p \times 1$, P is a penalty on the number of nonzero components of $\mathbf{v}$, and λ is a tuning parameter. We can view $\mathbf{u}$ as the first principal component and $\mathbf{v}$ the coordinates of $\mathbf{D}$ along that component. Therefore, by penalizing $\mathbf{v}$ we insist on an approximation of $\mathbf{D}$ which is a function of relatively few dimensions. In the HDLSS setting, however, we are still estimating $(p + n) > n$ quantities.

Support Vector Machines, Data Piling, and Fisher's Linear Discriminant We consider the two-class classification problem, where we have the usual $n \times p$ matrix of data $\mathbf{D}$ and a vector $\mathbf{Y} \in \Re^{\mathbf{n}}$ s.t. $\mathbf{Y_i} = \mathbf{1}$ if individual i is in class 1, $\mathbf{Y_i} = -\mathbf{1}$ if class 2. We can describe a hyperplane in $\Re^p$ by $\mathbf{w}$, its normal vector, and $\boldsymbol{\beta}$, its offset. Then we define the vector of residuals

$$\mathbf{r} = \mathbf{Y}(\mathbf{Dw} + \boldsymbol{\beta}\mathbf{e})$$

where $\mathbf{e}$ is a vector of all 1s. We would like to find $(\mathbf{w}, \boldsymbol{\beta})$ so that all the residuals are positive; i.e. the classes have been separated by the hyperplane. This might not be possible, however, so we consider the perturbed residuals

$$\mathbf{r} = \mathbf{Y}(\mathbf{Dw} + \boldsymbol{\beta}\mathbf{e}) + \boldsymbol{\xi}$$

and seek to penalize the error term $\boldsymbol{\xi}$. Specifically, we solve

$$\min_{\mathbf{w},\boldsymbol{\beta},\boldsymbol{\xi}} \mathbf{w}^\top \mathbf{w} + \mathbf{C}\mathbf{e}^\top \boldsymbol{\xi}$$

where C is some constant, subject to $\boldsymbol{\xi} \geq 0$ and

$$\mathbf{Y}(\mathbf{Dw} + \boldsymbol{\beta}\mathbf{e}) + \boldsymbol{\xi} \geq \mathbf{e}.$$

Under some regularity conditions, this problem admits solutions. In practice, it happens that the solution depends only on certain rows of the data; i.e. there is usually a relatively sparse subset of the individuals, which we call *support vectors*, which determine the solution to the Support Vector Machine (SVM) problem. Geometrically, the solution minimizes the distances between the convex hulls of the points when arranged by class and the support

vectors are those which are closest to the dividing boundary. It is worthwhile to note that SVM utilizes a sparse subset of the individuals, not a subset of the dimensions; therefore, it is not necessarily especially suitable for the HDLSS setting.

In fact, Ahn, Marron, and Todd [40] demonstrate that in the HDLSS setting support vectors are relatively numerous and that the projected data show *data piling*; i.e. in the projection, many of the components are exactly the same. Ahn and Marron [3] describe a vector, the maximal data piling direction, onto which projections of two-class data have only two values, one for each class. Specifically, given sample class means $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ and (singular) sample covariance matrix $\hat{\mathbf{\Sigma}}$, they define

$$\mathbf{v_m} = \hat{\mathbf{\Sigma}}^{\dagger}(\hat{\mathbf{x}} - \hat{\mathbf{y}})$$

where $\hat{\mathbf{\Sigma}}^{\dagger}$ is the Moore-Penrose generalized inverse of $\hat{\mathbf{\Sigma}}$. This is a natural generalization of Fisher's linear discriminant[23] in which the vector of projection is given by

$$\mathbf{v_m} = \hat{\mathbf{\Sigma}}^{\mathbf{-1}}(\hat{\mathbf{x}} - \hat{\mathbf{y}}).$$

In simulation studies the method does work well, especially when the data are highly correlated, but it admits no clear generalization to the case of multiple classes.

However, data piling can be an artifact of the HDLSS setting and not a desirable property, especially when considering the classification of new data. In defining Distance Weighted Discrimination Optimization (DWDO) [40] the same authors seek to solve

$$\min_{\mathbf{r},\mathbf{w},\boldsymbol{\beta},\boldsymbol{\xi}} \sum_i \frac{1}{\mathbf{r_i}} + C\mathbf{e}^{\top}\boldsymbol{\xi}$$

subject to $\mathbf{r}, \boldsymbol{\xi} \geq \mathbf{0}$ and $\mathbf{w}^\top \mathbf{w} \leq \mathbf{1}$. The problem no longer admits a single-step solution and the solution the authors describe is computationally complex, but by minimizing the inverses of the residuals they avoid the problem of data piling and improve on their previous result.

Regularized Projections and Truncated Nearest Neighbor For the two-class case with new data $\mathbf{d}$ we can form the regularized classifier

$$\delta(\mathbf{d}) = (\mathbf{d} - \hat{\boldsymbol{\mu}})^\top (\hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I})^{-\mathbf{1}} (\hat{\boldsymbol{\mu}}_\mathbf{1} - \hat{\boldsymbol{\mu}}_\mathbf{0})$$

where λ is a positive constant and we assign $\mathbf{d}$ to class 1 if $\delta(\mathbf{d}) \geq \mathbf{0}$. The addition of the perturbation $\lambda \mathbf{I}$ is akin to a Tikhonov regularization, which we will see again below in the ridge regression. Notably, $\hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}$ is necessarily invertible if we use any nonnegative definite estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$, so the quantity can be computed.

While not specifically a projection method, the nearest shrunken centroid (NSC) method of Tibshirani et al.[55] bears discussion here, as it supersedes the regularized projection in precisely the same way that the lasso superseded ridge regression. Specifically, if $\mathbf{c} \in \Re^\mathbf{p}$ is the centroid of a class, we define the *soft thresholding* rule

$$\mathbf{c}'_\mathbf{i} = \text{sign}(\mathbf{c_i})(|\mathbf{c_i}| - \boldsymbol{\Delta})_+$$

where Δ is some constant. Notably, a number of the components might be set to 0, specifically when $|\mathbf{c_i}| \leq \boldsymbol{\Delta}$. If it happens that two centroids have the same components set to 0 (as will often be the case in practice), then we have essentially excluded the dimension from consideration.

We note that soft thresholding takes its name from the fact that $\mathbf{c_i'}$ is a continuous function of $\mathbf{c_i}$, which is generally a desirable property. The alternative is *hard thresholding*; e.g. we might define $\mathbf{c_i'} = \mathbf{c_i} I(|\mathbf{c_i}| \geq \boldsymbol{\Delta})$ and note that small changes in $\mathbf{c_i}$ can produce large changes in $\mathbf{c_i'}$.

Asymptotic Results for Projection Methods In conclusion, we follow the direction of Fan and Fan [22] who demonstrate that all projection methods are asymptotically no better than guessing. Specifically, we assume that $\boldsymbol{\alpha}$ is a p-dimension uniformly distributed unit random vector on a $(p-1)$-dimensional sphere. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$. We suppose that

$$\lim_{p\to\infty} \frac{1}{p^2} \sum_{j=1}^{p} \lambda_j^2 < \infty$$

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \lambda_j = \tau$$

where τ is some positive constant. Also, assume that $p^{-1}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} \to 0$ as $p \to \infty$. Then, given new data $\mathbf{d}$ we form the classifier

$$\hat{\delta}_\alpha(\mathbf{d}) = (\boldsymbol{\alpha}^\top \mathbf{d} - \boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}})(\boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}_\mathbf{1} - \boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}_\mathbf{2})$$

where $\boldsymbol{\mu}_i$ is the empirical mean for class i, but note that

$$P(\hat{\delta}_\alpha(\mathbf{d}) \leq \mathbf{0} | \mathbf{d} \in \text{class 2}) \to \frac{\mathbf{1}}{\mathbf{2}} \text{ as } \mathbf{p} \to \infty.$$

1.2.1.2 Neural Networks

We avoid too deep a consideration of the history of neural networks, but it is worthwhile to acknowledge the analogy to biological neural networks. To some degree of approximation, a brain can be viewed as a large network of neurons, specialized electrically excitable cells. A neuron receives electrical impulses from other neurons and, depending on the total impulse received, can itself send out an electrical impulse. Though simple to describe, such networks can be large and facilitate complex behaviors; the human brain, for instance, contains on the order of 10^{11} neurons and 10^{15} connections between neurons.[43]

In an artificial neural network, a neuron is a node and represents a function. The impulse received by a neuron is a weighted linear combination of the node or function's inputs; i.e. if the node is connected to m other nodes, each of which contributes an impulse x_i, the total impulse received by the node is

$$\sum_{i=1}^{m} w_i x_i$$

for some set of weights $\{w_i\}_{i=1}^{m}$. In the modeling of biological neural networks, it is most commonly assumed that the firing of a neuron is a binary event; i.e. if the impulse received is above some threshold, then the neuron transmits an impulse, but does not otherwise. In an artificial neural network, however, it is convenient to assume that a node's output is a differentiable function of its inputs. Specifically, the new output is given by

$$x = f\left(\sum_{i=1}^{m} w_i x_i\right)$$

where f is some function that "looks like" the Heaviside function; i.e. f is some differentiable, monotonically increasing and bounded function such as $f(x) = \arctan(x), \tanh(x), \frac{1}{1+e^{-x}}$.

It has been shown recently by Auer et al. [6] that even very simple neural networks can approximate arbitrarily closely any bounded continuous on a compact set. However, finding the approximation is nontrivial. "Learning" for a given set of nodes n is essentially a regression. Given m inputs $\{x_i\}_{i=1}^m$ and targets $\{y_i\}_{i=1}^m$, we most often seek to minimize

$$\epsilon(n, w) = \sum_{i=1}^{m} ||f_{(n,w)}(x_i) - y_i||$$

for some norm $||\cdot||$. It is not necessarily the case that the norm is differentiable with respect to the weights; when neural networks are used for classification, for instance, we could choose to count the number of misclassifications.

However, when the norm and scaling functions are continuously differentiable, gradient methods can be applied; the simplest is gradient descent. We temporarily view n, the node structure as fixed, so that $\epsilon(n, w) = \epsilon(w)$ is a function of the weights alone. Given a set of weights w_0, we define

$$w_{i+1} = w_i - \eta \nabla \epsilon(w_i)$$

where η is some tuning parameter. Then we note that

$$\epsilon(w_{i+1}) - \epsilon(w_i) = -\eta \epsilon_w(w_i) \nabla \epsilon(w_i) + \frac{\eta^2}{2} \nabla \epsilon(w_i)^\top \epsilon_{ww}(\xi) \nabla \epsilon(w_i)$$

for some ξ on the line segment between w_i and w_{i+1}. This suggests that if η is sufficiently small, the first term will dominate the second so that $\epsilon(w_{i+1}) < \epsilon(w_i)$; i.e. we have found an improved set of weights for the network.

Gradient descent is analytically and computationally simple. However, it usually only suffices for the simplest problems. For instance, η is usually taken to decrease monotonically in i so that the weights will converge to, not oscillate around minima. The performance of the training algorithm is heavily dependent on the rate of decrease, yet there is no comprehensive theory of optimum rate. Furthermore, even with appropriate η's, weights can converge to local minima so the selection of initial weights is itself a key problem. Finally, we have treated n, the network structure, as given, but it is known that the performance of a network depends heavily on its structure. In summary, the construction and training of a neural is highly nontrivial. In the following, however, we assume some effective solution exists for the given application.

Autoencoders are a type of neural network pertinent to dimension reduction. Given a set of m individuals d_i of dimension p, we seek a neural network (n,w) which minimizes

$$\epsilon = \sum_{i=1}^{m} ||f_{(n,w)}(d_i) - d_i||_2$$

where we constrain the number of nodes and nonzero weights in (n,w); essentially, we seek a low-rank approximation to the identity. If (n, w) is constructed with a small inner layer, the network can be split in half so that $f_{(n,w)}$ can be factored as

$$f_{(n,w)} = d_{(n,w)} \circ e_{(n,w)}$$

where $p' \ll p$,

$$e_{(n,w)} : \Re^p \to \Re^{p'}$$

is an encoder and

$$d_{(n,w)} : \Re^{p'} \rightarrow \Re^{p}$$

is a decoder. Therefore, an individual d_i can be encoded in only p' dimensions as $d'_i = e_{(n,w)}(d_i)$ and, with error depending on ϵ, decoded as $\hat{d}_i = d_{(n,w)}(d'_i)$.

It has been demonstrated by Hinton and Salakhutdinov [34] that autoencoders can in general reproduce a PCA on HDLSS data yet tend to yield more efficient encodings. In fact, Demartines and Herault [19] showed that neural networks could discover efficient projections onto manifolds, recovering nonlinear generalizations to PCA. Due to their structure, the training of autoencoders is computationally expensive and they typically admit no interpretation. Encodings, for instance, are usually highly nonlinear functions of the inputs. However, when the network can be trained offline and no interpretation is necessary, neural networks present a plausible dimension reduction solution.

1.2.2 Subset Selection

A key in the success of synthetic dimension reduction techniques in the HDLSS setting has been the idea of sparsity. Instead of insisting on extracting all meaningful variance from every dimension in $\mathbf{D}$, some are simply discarded from consideration. Ideally, we might want to examine all 2^p combinations of dimensions to determine which best describes the data but this is computationally intensive; i.e. we would like to find the best subset. The "leaps and bounds" procedure of Furnival and Wilson [24] has for a third of a century made this

feasible for $p < 50$ and, with increases in computing power, this limit has grown. However, we note that in modern applications p regularly exceeds 1,000 and no exhaustive method ever seems plausible. We note, however, that approximations to the best subset can be obtained through penalized regressions.

Specifically, given a set of training data $\{\mathbf{d_i}, \mathbf{y_i}\}_{\mathbf{i=1}}^{\mathbf{n}}$ with $\mathbf{d_i} \in \mathbf{R^p}, \mathbf{y_i} \in \Re$, we seek to solve

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \text{argmin} \left[\sum_{i=1}^{n} \epsilon(\mathbf{d_i}, \mathbf{y_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \mathbf{p}(\boldsymbol{\beta}) \right]$$

where $\boldsymbol{\alpha}$ can be understood as the intercept, $\boldsymbol{\beta} \in \Re^p$ the regression coefficients, ϵ some measure of error, p some penalty term, and λ a tuning parameter. In the two techniques we consider here,

$$\epsilon(\mathbf{d}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = ||\mathbf{y} - (\boldsymbol{\alpha} + \boldsymbol{\beta}^\top \mathbf{d})||_{\mathbf{2}}^{\mathbf{2}}$$

where $|| \cdot ||_p$ is the L^p norm. We consider ridge regression (or Tikhonov regularization) and Tibshirani's lasso [54] in which, respectively,

$$p(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2^2, ||\boldsymbol{\beta}||_1.$$

1.2.2.1 Unpenalized Regression

To understand the influence of the various penalty terms, we first consider an unpenalized regression with intercept 0; i.e. $\boldsymbol{\alpha}, \lambda = 0$. We can rewrite the problem as

$$\hat{\boldsymbol{\beta}}_u = \arg\min_{\boldsymbol{\beta}} ||\mathbf{D}\boldsymbol{\beta} - \mathbf{Y}||_{\mathbf{2}}^{\mathbf{2}}$$

where $\mathbf{D}$ is the $n \times p$ matrix of data and $\mathbf{Y}$ the vector of n targets. We have already noted that when $p < n$ and $\text{rank}(\mathbf{D}) = \mathbf{p}$, then

$$\hat{\boldsymbol{\beta}}_u = (\mathbf{D}^\top \mathbf{D})^{-\mathbf{1}} \mathbf{D}^\top \mathbf{Y}$$

is the unique solution to the unpenalized regression problem. If we assume that the columns of $\mathbf{D}$ form an orthonormal set, i.e. $\mathbf{D}^\top \mathbf{D} = \mathbf{I}$, then

$$\hat{\boldsymbol{\beta}}_u = \mathbf{D}^\top \mathbf{Y}.$$

Notably, $\hat{\boldsymbol{\beta}}_u = 0$ only accidentally, if it occurs that a column of $\mathbf{D}$ is orthogonal to $\mathbf{Y}$. In general we should not expect that many (if any) of the coefficients in such a regression should be 0. It is not immediately clear why this might be problematic.

Accumulation of Errors in Unpenalized Regressions We have observed above that even when the data are observed directly, that is to say, without error, geometrical considerations make the HDLSS case problematic. In practice, there is an additional complication: we do not observe the data directly, but only have perturbed measurements; e.g. instead of $\mathbf{D}$, we observe

$$\mathbf{D}' = \mathbf{D} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is some error term. The more columns of $\mathbf{D}'$ we retain for classification, the more columns of $\boldsymbol{\varepsilon}$ and therefore the more error we introduce into our model; i.e. these errors accumulate. Specifically, Ververidis and Kotropoulos[57] show that the classification rate of any k-means classifier goes to 50% as the dimensionality, p, increases, and attribute this to inaccurate estimates of the Mahalanobis distance between group centroids.

1.2.2.2 Ridge Regression

Using the same hypotheses, including specifically that the columns of $\mathbf{D}$ form an orthonormal set, we seek to solve

$$\hat{\boldsymbol{\beta}}_r = \arg\min_{\boldsymbol{\beta}} ||\mathbf{D}\boldsymbol{\beta} - \mathbf{Y}||_{\mathbf{2}}^{\mathbf{2}} + \lambda||\boldsymbol{\beta}||_{\mathbf{2}}^{\mathbf{2}}.$$

Then we note that

$$\frac{\delta}{\delta\boldsymbol{\beta}}\left[||\mathbf{D}\boldsymbol{\beta} - \mathbf{Y}||_{\mathbf{2}}^{\mathbf{2}} + \lambda||\boldsymbol{\beta}||_{\mathbf{2}}^{\mathbf{2}}\right] = 2(\mathbf{D}\boldsymbol{\beta} - \mathbf{Y})^\top\mathbf{D} + \mathbf{2}\lambda\boldsymbol{\beta}^\top = \mathbf{0}$$

which implies that

$$\hat{\boldsymbol{\beta}}_r = \frac{\mathbf{D}^\top\mathbf{Y}}{1+\lambda}$$

is a critical point of the penalized error function. The second derivative is $2(1+\lambda)\mathbf{I}$ which is positive definite as long as $\lambda > -1$; i.e. $\hat{\boldsymbol{\beta}}_r$ minimizes the penalized error. We then note that

$$\hat{\boldsymbol{\beta}}_r = \frac{\hat{\boldsymbol{\beta}}_u}{1+\lambda}.$$

Then $(\hat{\boldsymbol{\beta}}_r)_i = 0$ iff $(\hat{\boldsymbol{\beta}}_u)_i = 0$, implying that ridge regression gives no dimension reduction; i.e. if a dimension "contributes" to an unpenalized regression, then it will contribute in the ridge regression.

It is a well-established result, beginning in the statistical literature with Hoerl [35], that the ridge regression stabilizes the unpenalized regression. We recall that if $\mathbf{D}$ is not orthonormal, we have the unpenalized regression

$$\hat{\boldsymbol{\beta}}_u = (\mathbf{D}^\top\mathbf{D})^{-\mathbf{1}}\mathbf{D}^\top\mathbf{Y}.$$

If we solve the ridge regression problem without the assumption of orthonormality, we obtain the solution

$$\hat{\boldsymbol{\beta}}_r = (\mathbf{D}^\top \mathbf{D} + \lambda \mathbf{I})^{-\mathbf{1}} \mathbf{D}^\top \mathbf{Y}.$$

We note that the eigenvalues of $\mathbf{D}^\top \mathbf{D}$ are all nonnegative, since if

$$(\mathbf{D}^\top \mathbf{D})\mathbf{u} = \nu \mathbf{u} \Rightarrow$$

$$||\mathbf{D}\mathbf{u}||^{\mathbf{2}} = \mathbf{u}^\top (\mathbf{D}^\top \mathbf{D})\mathbf{u} = \nu \mathbf{u}^\top \mathbf{u} = \nu ||\mathbf{u}||^{\mathbf{2}}.$$

However, it may be the case that an eigenvalue ν of $\mathbf{D}^\top \mathbf{D}$ is near 0, in which case $(\mathbf{D}^\top \mathbf{D})^{-\mathbf{1}}$ will have a large eigenvalue ν^{-1}. Small variations in the data, such as perturbation by measurement error, will have large impacts on the estimates $\hat{\boldsymbol{\beta}}_r$ if they happen to correlate with the eigenvector corresponding to ν. However, we note that the corresponding eigenvalue of $(\mathbf{D}^\top \mathbf{D} + \lambda \mathbf{I})^{-\mathbf{1}}$ will only be $(\nu + \lambda)^{-1}$ so that an error will be magnified by a factor of at most λ^{-1}.

1.2.2.3 LASSO

In Tibshirani's lasso (where it is noteworthy that Tibshirani himself has not settled the question of capitalization) we seek to solve

$$\hat{\boldsymbol{\beta}}_t = \arg\min_{\boldsymbol{\beta}} ||\mathbf{D}\boldsymbol{\beta} - \mathbf{Y}||_{\mathbf{2}}^{\mathbf{2}} + \lambda ||\boldsymbol{\beta}||_{\mathbf{1}}$$

where we note the change in the exponent of the error term. Whereas $||\boldsymbol{\beta}||_2^2$ is a differentiable function of the regression coefficients, $||\boldsymbol{\beta}||_1$ is not, so this problem must in general be solved

by iterative methods. Originally, solution through quadratic optimization was suggested, but later work has shown least-angle regression (LAR) to be far more efficient.[21]

In the lasso, it is often the case that

$$(\hat{\boldsymbol{\beta}}_l)_i = 0 \neq (\hat{\boldsymbol{\beta}}_u)_i$$

i.e. while the ridge regression performs only shrinkage, the lasso performs shrinkage and sets some coefficients to 0; namely those that fall below some threshold depending on λ.

1.2.2.4 Dantzig Selector

Candes and Tao [12] consider the problem

$$\min ||\hat{\boldsymbol{\beta}}_d||_1 \text{ subject to } ||\mathbf{Dr}||_\infty \leq (\mathbf{1} + \mathbf{t^{-1}})\sigma\sqrt{\mathbf{2}\log \mathbf{p}}$$

where $\hat{\boldsymbol{\beta}}_d$ is the vector of regression coefficients, $\mathbf{r}$ the vector of residuals, and t some positive scalar. If the vector $\boldsymbol{\beta}$ is truly sparse and the data obey an uncertainty principle, relating the meaningful variance in the vector to error, then with very large probability

$$||\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}||_2^2 \leq 2C^2 \log p \left(\sigma^2 + \sum_i \min(\boldsymbol{\beta}_i^2, \sigma^2)\right)$$

for a known constant C. Notably, this is not an asymptotic result and the solution can be obtained by linear programming.

Families of Selectors James, Radchenko, and Lv [36] discuss situations in which the lasso and Dantzig selectors are identical, so that the bounds on the error of the Dantzig selector

might be in certain cases extended to the lasso. Specifically, if the data are orthonormal or $\hat{\mathbf{\Sigma}} = \mathbf{D}^\top \mathbf{D} = \rho \mathbf{e}\mathbf{e}^\top$ for some ρ, then the solutions are the same. Meinshausen et al. [41] gave a less strict condition on $(\mathbf{D}^\top \mathbf{D})^{-\mathbf{1}}$, but this quantity cannot be considered in the HDLSS case. These results give the indication that there is a large number of selectors, some of which remain unexplored, which on certain data report the same selections, yet specialize to particular applications; we do not try to make an exhaustive review of the possibilities.

1.2.2.5 Information Criteria

First described by Akaike in 1971 and named in 1974 [4], the Akaike Information Criterion (AIC) can be understood as the goodness of fit of a model. Specifically,

$$\mathrm{AIC} = 2k - 2\ln L$$

where k is the number of parameters in the model and L is the maximized likelihood value for the estimated model; the model with the lowest AIC is considered best. This can be used to search through subsets of the dimensions of $\mathbf{D}$ for the best description of the data. As an example, if we assume that models errors are iid normal, then the AIC becomes

$$\mathrm{AIC} = 2k + n\ln(n^{-1}\mathrm{RSS}) + C$$

where RSS is the sum of squared errors of the model, and C is some constant, the same for all such models. If we assume that the RSS is a decreasing function of k, which is merely

the statement that additional data does not make our model less accurate, then

$$\text{AIC} = 2k + n\ln(\text{RSS}) + C'$$

where C' does not depend on k, might have a global minimum and the AIC can indicate some model with simultaneously small k and RSS.

We also consider the closely related Bayesian Information Criterion (BIC), the most notable proponent of which has been Schwarz [51]. We define

$$\text{BIC} = k\ln n - 2\ln L.$$

The essential difference from the AIC is of course the penalty term; so long as $\ln n > 2$, the BIC places a heavier penalty on the number of free parameters in the model, but this is the statement that there have been more than just 7 observations. Again, we prefer models with smaller BIC. Using the assumptions of the example above, we obtain

$$\text{BIC} = \frac{k}{n}\ln n + \ln(\sigma^2)$$

where σ^2 is the variance of the errors.

One might wonder if these are not a sufficient solution to the problem of model selection. First, with a naive approach the criteria must be applied exhaustively to all 2^p combinations of subsets to find the best subset, but only the best subset according to that IC. The existence of two information criteria differing only in penalty, however, implies that the choice of penalty is not necessarily given by the problem itself and that improvements might yet exist.

1.2.3 Feature Selection

Feature selection differs from subset selection in that feature selection algorithms consider each dimension separately, so that one need only make p rather than 2^p considerations. We recall that the first essential difficulty in dimension reduction was the estimation of the covariance matrix of the dimensions; feature selection, by considering each dimension individually, does away with this at the cost that we might still retain a number of highly correlated vectors after selection.

1.2.3.1 Ad Hoc Feature Selection

Selection by Variance We noted above in the analysis of the PCA that one can often retain p' principle components where p' is the smallest s.t.

$$\sum_{i=1}^{p'} \mathbf{\Lambda_{ii}^2} \geq \alpha \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathrm{Var}(\mathbf{d^i})$$

for some $\alpha \in (0, 1)$; i.e. it is often the case that one tries to retain $(100\alpha)\%$ of the variance of the original data, based on the assumption that the meaningful variance in the data, which can be understood as the signal, will be large compared to the error, or noise. This has motivated a number of ad hoc techniques for the use of variance in feature selection. Essentially, each dimension is retained or discarded based on some function of its variance. The idea suffers from two main flaws. First, variance does not take into account classwise structure. Secondly, variance is necessarily scale variant.

Selection by ANOVA Analysis of variance (ANOVA) is the partitioning of a dimension's variance into components. In the case of classwise data, for instance, we can write

$$\text{Var}(\mathbf{d_i}) = [\text{between-class variance}] + [\text{within-class variance}]$$

where, for example, we might choose to explain variance in human height as dependent upon biological gender, which gives two classes. Then the difference between the average male and female heights would explain a large portion of the variance in the original data while, for example, differences in nutrition and health history might explain the within class variance. We can then create a statistic proportional to the between class variance, e.g.

$$\xi_i = \frac{[\text{between-class variance}]}{[\text{within-class variance}]} \text{ or } \frac{[\text{between-class variance}]}{[\text{total variance}]}.$$

If we have that ξ_i is large for some dimensions and small for others, then we suspect that the dimensions with larger ξ_i are better for classification and can retain a number of these based on some criterion or threshold.

Fan and Fan[22] generalize on the idea of the ANOVA and show that it can select meaningful vectors with probability tending to one. Specifically, we consider a single vector containing data for two classes of sizes n_1, n_2, with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$. We only observe sample means $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$ and sample variances $\hat{\sigma}_1^2, \hat{\sigma}_2^2$. Then we form the statistic

$$T_j = \frac{\hat{\boldsymbol{\mu}}_{1j} - \hat{\boldsymbol{\mu}}_{2j}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_{1j}} + \frac{\hat{\sigma}_2^2}{n_{2j}}}}$$

which, intuitively, represents the ratio of the between-class variance to the within-class variance.

We assume that $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is sparse; i.e. only the first s components are nonzero. This is the statement that after s dimensions, any additional dimension adds only noise; therefore, there are s (out of p) informative dimensions. Then assume that s is such that

$$\log(p - s) = o(n^{\gamma})$$

$$\log s = o(n^{1/2-\gamma}\beta_n)$$

for some $\beta_n \to \infty$, $0 < \gamma < \frac{1}{3}$. These statements assure that the number of informative dimensions, s, remains comparable to n but need not grow linearly. Furthermore, we assume that

$$\min_{1 \le j \le s} |T_j| = n^{-\gamma}\beta_n$$

which is the statement that the signal in the informative components does not vanish with respect to the noise. Finally, if we assume that $x \sim cn^{\gamma/2}$ for some positive constant c, then we have we have

$$P\left(\min_{j \le s} |T_j| \ge x \text{ and } \max_{j < s} |T_j|\right) \to 1 \text{ as } n \to \infty$$

i.e. in the limit, the difference between the t-statistic of an informative and uninformative dimension will be significant. However, we note that this is an asymptotic result and not necessarily applicable to the finite case.

1.2.3.2 The FAIR of Fan and Fan

Fan and Fan[22] define the Features Annealed Independence Rule (FAIR)

$$\hat{\delta}^b(\mathbf{d}) = \sum_{\mathbf{j=1}}^{\mathbf{p}} \hat{\boldsymbol{\alpha}}_\mathbf{j}(\mathbf{d_j} - \hat{\boldsymbol{\mu}}_\mathbf{j})\mathbf{1}\{|\hat{\boldsymbol{\alpha}}_\mathbf{j}| > \mathbf{b}\}$$

where b is some constant and we recall that the vector $\hat{\boldsymbol{\alpha}}$ is an approximation of $\alpha = \mu_1 - \mu_2$, which is supposed to be only sparsely nonzero. Notably, this is a hard thresholding rule, not a soft thresholding or shrinkage rule; vectors are discarded or retained, not transformed. Also, we note that when the data are independent, FAIR is similar to a t-test on each dimension yet selects an optimal number of retained features.

The authors present an upper bound error estimate, which we omit here, but note that in practical studies FAIR produces a sparser and more accurate set of dimensions than does the NSC method of Tibshirani et al.[55]

1.2.4 Bayesian Feature Selection

As an example of the Bayesian approach to dimension reduction, we consider the results of Ambramovich and Angelini[1]. Specifically, we can consider feature selection as a multiple hypothesis test. Given $\mathbf{d_i} \in \Re^\mathbf{p}$ from $i = 1, \cdots, n$ where

$$\mathbf{d_i} \sim \mathbf{f_i}(\mathbf{d_i}|\boldsymbol{\Theta}_\mathbf{i})$$

where Θ_i is some set of parameters, we consider the hypothesis tests

$$H_{0i} : \Theta_i \in \Omega_i \text{ vs. } H_{1i} : \Theta_i \in \Omega_i^c \quad i = 1, \cdots, n.$$

For instance, we may assume that the null hypothesis represents a vector which is not useful for classification; say, perhaps, the vector is constant except for measurement error. However, in what follows we are generally agnostic as to the nature of the parameter spaces and can treat this as a multiple testing problem, not dimension reduction specifically.

1.2.4.1 The Problem of Multiple Testing

We note that in many applications, the goal of hypothesis testing is to control the familywise error (FWE), the probability of making even a single Type I error in a series of n tests. In this case, suppose that $f_i = f, \Omega_i = \Omega$ for all i for some distribution function f, some set of parameters Ω; i.e. the hypothesis test is the same in all cases. Then, if we assume the hypotheses are determined independently, for some ϵ_1 we have

$$P(\text{Type I Error}) = P(\text{reject } H_{0i} | H_{0i}) = \epsilon_1 \quad i = 1, \cdots, n.$$

Then we note that

$$N = [\text{number of Type I errors made}] \sim B(\epsilon_1, n)$$

so that we must have

$$\text{FWE} = P(N \geq 1) = 1 - P(N = 0) = 1 - (1 - \epsilon_1)^n \Rightarrow$$

$$\epsilon_1 = 1 - (1 - \text{FWE})^{1/n}.$$

For instance, if we choose FWE = .75 and have $n = 1000$ hypotheses, insisting that one time in four we expect to have no Type I errors, we must have that $\epsilon_1 \leq .0014$. However, this implies that for each individual test, we can only make a Type I error with probability less than .0014; we will have to accept the null hypothesis nearly indiscriminately to achieve to this rate.

Rather, Ambramovich and Angelini appeal to the false discovery rate (FDR) of Benjamini and Hochberg[7] which controls the expected proportion of Type I errors among the hypotheses being rejected. Suppose in addition to ϵ_1 we have r_1, the probability of rejecting a null when the null is false, and α, the proportion of false nulls among the n hypotheses. Then we can calculate

$$\text{E[number of hypotheses rejected]} = (\alpha r_1 + (1 - \alpha)\epsilon_1)n$$

$$\text{E[number of Type I errors]} = (1 - \alpha)\epsilon_1 n$$

so that, to some degree of approximation,

$$\text{FDR} \approx \frac{\text{E[number of Type I errors]}}{E[\text{number of hypotheses rejected}]} = \frac{(1 - \alpha)\epsilon_1}{\alpha r_1 + (1 - \alpha)\epsilon_1} = \frac{1}{1 + \frac{\alpha r_1}{(1-\alpha)\epsilon_1}}.$$

For instance, if $\alpha = .10$ (the null is usually true), $r_1 = .95$ (we can usually recognize it when it is not), and FDR = .10 (nine out of ten our of discoveries should be true), then any $\epsilon_1 \leq .011$ suffices. While still relatively strict (this corresponds to a p-value of around 1% for a single test), this metric leads to a feasible methodology.

1.2.4.2 Hierarchical Prior Model

Arguing that it is not feasible to calculate the odds on individual tests, especially when the hypotheses might not be determined independently, the authors suggest that there could nevertheless be some intuition on the total proportion of hypotheses that come from the nulls and alternatives. For instance, we could know that the alternatives in the data should be sparse (i.e. the alternative is only rarely true) but we might not be able to calculate the probability that any *specific* alternative might be true.

Therefore, we construct a n-dimensional vector $\mathbf{x}$ where

$$\mathbf{x_i} = \mathrm{I}(\mathbf{H_{1i}} \text{ is true}) \quad \mathbf{i} = \mathbf{1}, \cdots, \mathbf{n}$$

i.e. $\mathbf{x_i} = \mathbf{0}$ if the null is true, 1 if the alternative is true. Again, the authors argue that we cannot usually ascertain the *a priori* probability $P(x_i = 1)$ but that we can discuss the prior distribution $\pi(k)$ of

$$k = \sum_{i=1}^{n} x_i = x^\top x.$$

Then, as we are unable to discriminate between any two $\mathbf{x}$ when they contain the same number of ones, we have

$$P\left(x | \sum_{i=1}^{n} x_i = k\right) = \binom{n}{k}^{-1}.$$

We then assume that

$$(\Theta_i | \mathbf{x_i} = \mathbf{0}) \sim \mathbf{p_{0i}}(\mathbf{\Theta_i}) \quad \text{and} \quad (\mathbf{\Theta_i} | \mathbf{x_i} = \mathbf{1}) \sim \mathbf{p_{1i}}(\mathbf{\Theta_i})$$

for densities on Ω_i and Ω_i^c respectively. The authors then calculate the full model

$$\pi(\mathbf{x}, \mathbf{k}|\mathbf{D}) \propto \binom{\mathbf{n}}{\mathbf{k}}^{-\mathbf{1}} \pi(\mathbf{k}) \mathbf{I}\{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i} = \mathbf{k}\} \prod_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{B_i^{-1}})^{\mathbf{x_i}}$$

where B_i is the Bayes factor of H_{0i}

$$B_i = \frac{\int_{\Omega_i} f_i(\mathbf{d_i}|\mathbf{\Theta_i}) \mathbf{p_{0i}}(\mathbf{\Theta_i}) \mathbf{d\Theta_i}}{\int_{\Omega_i^c} f_i(\mathbf{d_i}|\mathbf{\Theta_i}) \mathbf{p_{0i}}(\mathbf{\Theta_i}) \mathbf{d\Theta_i}}.$$

1.2.4.3 Testing Procedure

A common approach to testing is to select the most likely configuration of true and false null hypotheses according to the posterior mode $\pi(\mathbf{x}, \mathbf{k}|\mathbf{D})$. Normally, this would require a consideration of all 2^n configurations of $\mathbf{x}$, but in this case, as we have the Bayes factors, we only need to consider $n + 1$ configurations.

Specifically, we order the Bayes factors so that $B_1 \leq \cdots \leq B_n$ and find $\hat{k}$ that maximizes

$$\hat{\pi}_k = \pi(k|D) \propto \binom{n}{k}^{-1} \pi(k) \prod_{i=1}^{k} B_i^{-1}.$$

If $\hat{k} = 0$, then accept all the null hypotheses. If $\hat{k} > 0$, accept the first $\hat{k}$ hypotheses corresponding to $B_1, \cdots, B_{\hat{k}}$. Therefore, we retain the $\hat{k}$ vectors for which the null is rejected; i.e. for which $x_i = 1$.

Related Procedures The procedures of Kass and Raferty[37] and Berger and Pericchi[8] consider each of the combinations of hypotheses as disjoint partitions of the larger parameter space $\oplus_{i=1}^{n} \Omega_i$, which requires a generally prohibitive search through 2^n combinations. We also note the stepwise procedure of Sarkar and Chen[49] which differs in the choice of priors.

1.2.5 MCMC Feature Selection and Estimation

Suppose we have a random variable R with a probability distribution function (pdf) f_R or distribution function F_R which is unknown, difficult to integrate (e.g. for the purpose of finding confidence intervals), or difficult to maximize (e.g. for finding MLEs). If we are able to instead generate n independent, random draws $\{r_i\}_{i=1}^n$ from the distribution, we can consider the empirical cumulative distribution function (ecdf) for arbitrary c:

$$\hat{F}_n(c) = \frac{\text{number of } r_i \leq c}{\text{total number of } r_i \text{ generated}}.$$

Under certain circumstances, we expect $F_n(c)$ to converge to the value at c of the actual distribution function of the random variable.

The usual method of generating pseudorandom variables, however, is not useful here. If we are able to generate a sequence of independent pseudorandom variables $u_i \sim U[0,1]$, uniform on [0,1], then we can calculate $r_i = F_R^{-1}(u_i)$ and we have a sequence of pseudorandom, independent draws from R. By assumption, however, neither F_R nor its inverse are available.

A MCMC algorithm is an algorithm which produces a sequence of pseudorandom numbers that, if designed correctly, converges in distribution to a desired target distribution. In our application, the target random variable can be anything from a single parameter to the entire set $\{x_i\}_{i=1}^p$ of indicator variables. We allow the algorithm to produce a large number of samples from which we can draw estimates; e.g. maximum *a posteriori* estimates, confidence intervals, etc.

Whether the ecdf converges to the cdf and what speed at which it does so are discussed at length in [48], though an early result which is sufficient for many applications is given in [42]. We discuss one algorithm below.

1.2.5.1 Random Walk MCMC

Suppose we wish to estimate the distribution of a random variable x, which has pdf function f. Let G_i be a sequence of iid symmetric random variables, with pdf g, its *transition kernel*. Given an estimate x_i, we generate a new proposal $y_i = x_i + g_i$. In other words, the next value we consider is a small perturbation of the previous value of the chain. We then define

$$x_{i+1} = \begin{cases} y_i & \text{with probability } \min\left(1, \frac{f(y_i)}{f(x_i)}\right) \\ x_i & \text{otherwise} \end{cases}$$

The probability that y_i is accepted, i.e. $x_{i+1} = y_i$, is the *acceptance probability*.

This approach is useful for exploring local regions of the space. If, for instance, it is believed that a random variable has the majority of its "mass" in a connected subset of the space, the random walk MCMC will very likely map that random variable well; i.e. its ecdf will converge quickly to the variable's actual cdf. However, for the same reason, if the random variable is for example bimodal, there is a possibility the random walk MCMC will never in finite simulations visit significant portions of the variable's support.

There is also the consideration of step size. In this toy example, we seek to simulate a standard normal variable and use the transition $g_i \sim N(0, \sigma^2)$. Let $x_0 = 0$ be the first value

in the MCMC chain. Then we calculate

$$\log \frac{f(y_0)}{f(x_0)} = \log \frac{f(x_0 + g_0)}{f(x_0)} = \frac{1}{2}(x_0^2 - (x_0 + g_0)^2)$$

where we have substituted the pdf of the standard normal. Multiplying by (-2) and substituting in the value $x_0 = 0$ we obtain

$$-2 \log \frac{f(y_0)}{f(x_0)} = g_0^2 \sim \sigma^2 \chi_1^2$$

since $g_i \sim N(0, \sigma^2)$ by hypothesis.

If this value is large then the probability of accepting the new proposal is small, and vice versa. For example, if $\sigma^2 = 6$, we have

$$E\left[-2 \log \frac{f(y_0)}{f(x_0)}\right] = 6E\chi_1^2 = 6,$$

which implies that

$$\frac{f(y_0)}{f(x_0)} = e^{-3} \approx .05.$$

is roughly the acceptance probability in this case. In other words, by the fifth step of the MCMC algorithm there is a 77% chance we will still be at $x_5 = 0$. The ecdf might converge to the cdf but it will do so slowly. This is the problem of tuning; selection of the distribution used in the MCMC to simulate the target distribution.

1.2.5.2 Simulation Example

Suppose as an example that we have a sequence of $N = 250$ iid random variables

$$(Y_i|\rho^2) \sim (1 + \rho^2)\chi_{10}^2$$

where $\rho^2 = 4$. This example is not arbitrary; c.f. (3.1) below, which is a mixture, but is handled in the same manner. We would like to recover an estimate of ρ^2.

We analyze the data with the random walk MCMC described above, with the gaussian transition kernel $G_i \sim N(0, .05)$. Our initial estimate is $\hat{\rho}^2 = 1$. The variance of the kernel and the initial estimate are both chosen poorly to illustrate what is known as the "burn-in period," as in the figure below. The estimates of ρ^2 are all eventually in a neighborhood of its actual value of 4, but there is a substantial period in which they are not. This period, here approximately 150 samples, is usually discarded and considered the cost of a poor initial guess.

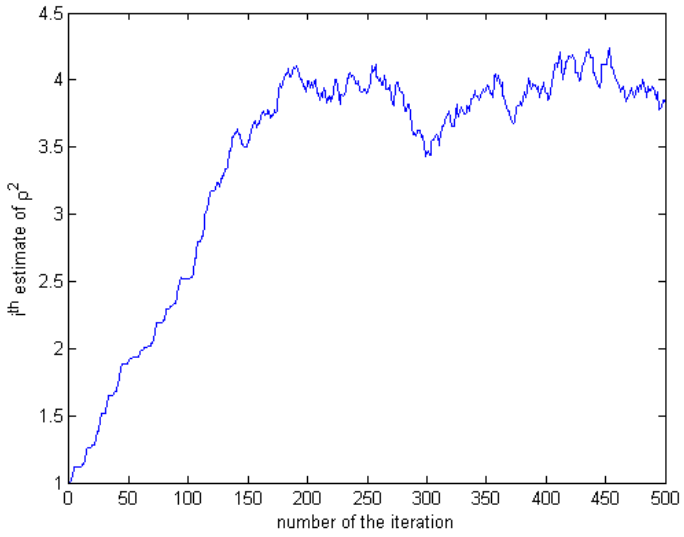

Figure 1.1: Plot of MCMC estimates of ρ^2

Depending on the number of parameters being estimated (we have only considered one here, though MCMC can handle an arbitrary number just as easily) and the tuning of the transition kernel, convergence can take a while. If instead we use the transition kernel

$G_i \sim N(0, .5)$ and first guess $\hat{\rho}^2 = 4.2$, there is no visible burn-in period and we can retain all the generated data. Figure 1.2 shows 5000 iterations of this random walk MCMC and 1.3 its ecdf after those 5000 iterations.

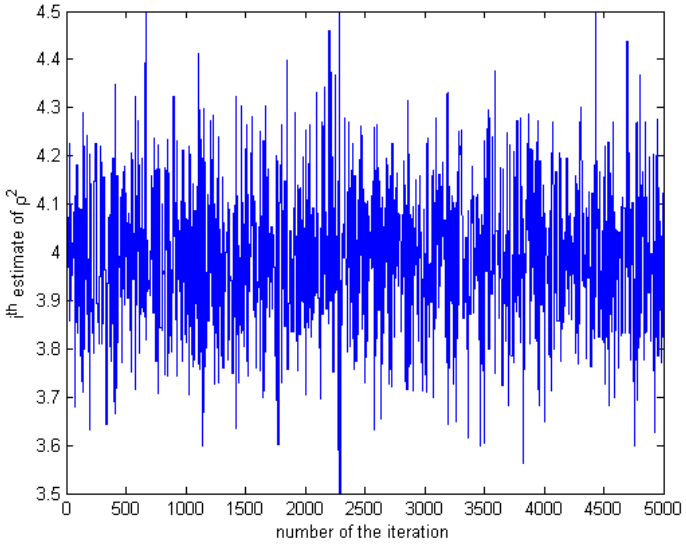

Figure 1.2: Plot of MCMC estimates of ρ^2

There is a subtle difficulty in interpretation. We know from the construction of the data set that $\rho^2 \sim \delta(4)$. However, in construction of the random walk MCMC, we have insisted only that ρ^2 be finite. The MCMC approach creates an *a posteriori* distribution for ρ^2 which is *not* the point mass we might expect and its ecdf will never converge to a δ-distribution. Is this necessarily the wrong approach? In fact, in practice, the *a posteriori* estimates (including not only point estimates but also confidence intervals) might be more robust against measurement error than analytic estimates.

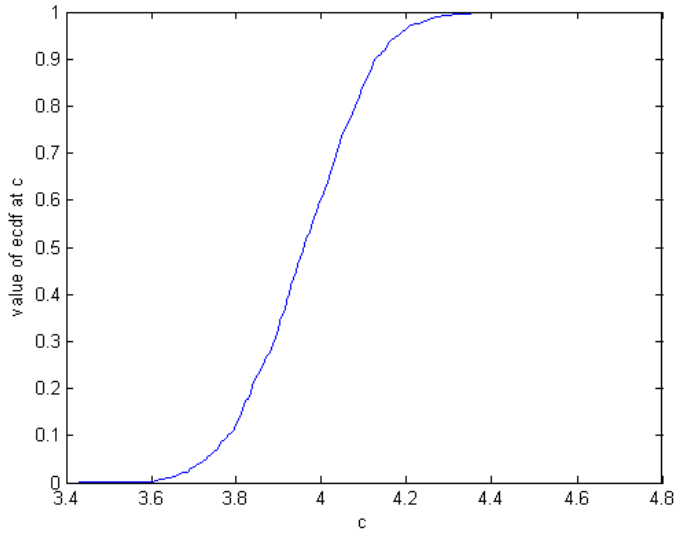

Figure 1.3: Plot of ecdf of estimates of ρ^2

This particular approach is *not* meant to be a recommendation. Obviously, the problem we present here can be solved analytically with a MLE. This is merely a demonstration that a naive MCMC algorithm can be easily coded ($\approx$ 30 lines of code) and casually applied to data with reasonable results. In practice, there can be a variety of reasons to use the MCMC approach over analytic estimates, and there is a large literature around doing this correctly.

1.2.5.3 Notes on MCMC methods

It is worthwhile to note that MCMC algorithms will produce distributions with relatively unusual properties. There is, for example, a nonzero probability that particular values will be reproduced in the sequence, i.e. $x_i = x_{i+1}$, whereas two independent draws from the distribution of x will almost surely not have the same value. In fact, the sequence produced

by the random walk MCMC will be auto-correlated, whereas independent draws from the distribution of x will not be. Finally, if the initial estimate is far from the main body of the distribution, it might take some time, as we saw above, for the estimates to evolve towards the mean of x.

It must be kept in mind that the ecdf of the sequence produced by the algorithm converges pointwise to the cdf of the parameter being estimated; this is one of the weakest forms of convergence. However, it should be remembered that even a poorly considered random walk MCMC can yield useful point estimates of local likelihood maxima. As long as the likelihood under the new proposal is greater than the likelihood under the old value of the parameter, the algorithm will always accept the proposed likelihood. This is not a recommendation of sloppiness, of course, but a note that MCMC algorithms can be more robust than analytic estimates when the underlying structures of the data are not known and therefore cannot be used to justify precise calculations.

1.3 Real Data Application

Subject

Animals use a variety of media to communicate their species, intentions, fitness, etc. Here, our concern is the first; can individuals from a group of electric fish be classified into species based on their signals alone? Specifically, some of the freshwater electric fishes of Africa

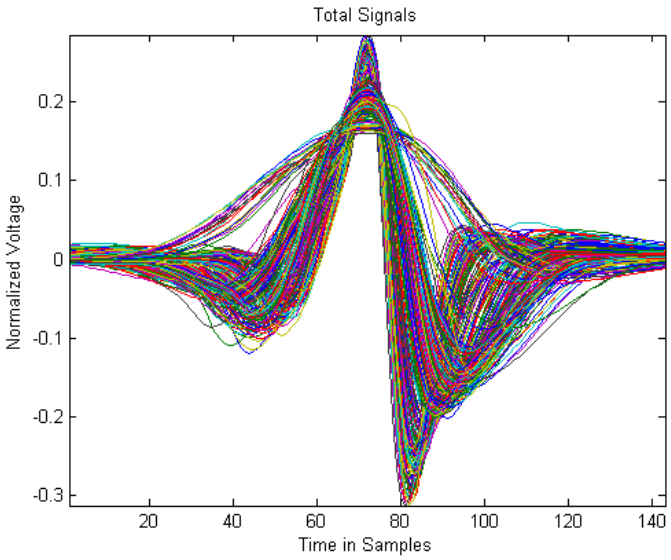

Figure 1.4: Centered, normalized fish signals

(Mormyriformes) and the Neotropics (Gymnotiformes) generate remarkably consistent signals, which they use for navigation and communication.[11] For our purposes, these may be visualized as strings of repeated pulses, such as those portrayed here.

These signals have relatively high dimension. Even when a single pulse is isolated, which is itself a form of dimension reduction, there were 2^{11} samples per pulse after resampling, yet there were only 263 individuals from five species. Therefore, there will be no unique LDA on the data unless the data are subjected to dimension reduction.

Methods

All signals were converted to a common length (2048 samples), sample rate (96 kHz), and root-mean-square normalized amplitude. All pulses were aligned at their global maximum, yielding, for instance, this sample of pulses from various species.

The data were then subjected to a variety of transforms; we recall that the principal component analysis is often used before dimension reduction to reduce the number of correlated variables in the data. Here, noting that because of the dimensionality there exists no unique PCA, we used the Fourier transform, the spectral density, the windowed Fourier transform (gaussian with varying variances), the discrete wavelet transform with a variety of wavelets, and a landmark-based procedure, which located a number of maxima, minima, turning points, etc.

The columns of data were then ranked by four metrics, namely the variance, the coefficient of variation, ANOVA, and pairwise ANOVA and a subset of size p' was chosen. A random subset of the individuals was chosen as a training set to condition a linear discriminant analysis (LDA) and the misclassification rate of the LDA on the whole data set served as the primary metric. Notably, this is not a Bayesian analysis and the analysis recommends no p'. Due to this, p' was varied exhaustively.

Transforms

Here we seek to understand the ideas behind the transforms chosen. While all transforms were applied in their discrete form, here we investigate their continuous forms for convenience.

Fourier Transform Given a signal $f : \Re \to \Re$ we define its Fourier transform

$$[F]f(\omega) = \int_{\Re} f(x)e^{-2\pi i\omega x}dx$$

if the integral exists. Under certain conditions, the transform may be inverted with

$$f(x) = \int_{\Re} [F]f(\omega)e^{2\pi i\omega}d\omega.$$

We can approximate this last integral with a Riemann sum

$$f(x) = \sum_{n=-\infty}^{\infty} [F]f(\omega_n)e^{2\pi i\omega_n}\Delta_n$$

which implies that we may view f as the sum of an infinite number of sinusoids. If $f \in L_1$, then the Riemann-Lebesgue lemma states that $[F]f(\omega) \to 0$ as $\omega \to \infty$, so that f might be well approximated by the inverse transform of the restriction $[F]f(\omega)|_C$ where C is some compact set. This is useful in dimension reduction, as it implies we should be able to "cut off" $[F]f(\omega)$ after some point, ignoring frequencies above some ω_0, yet still retain most of the information in f. This is especially true if f can be represented faithfully as a sum of sinusoids.

Spectral Density The spectral density is defined as $[S]f(\omega) = |[F]f(\omega)|$ and has the same motivations as the Fourier transform. Since the Fourier transform usually yields complex numbers, the spectral density is more easily visualized.

Windowed Fourier Transform We define the Windowed Fourier Transform (WFT)

$$[WF]f(x,\omega) = \int_{\Re} f(t)g(t-x)e^{-2\pi it}dt$$

where $g : \Re \to \Re$ is some window function. Typically, g is nonnegative with $g(x)$ small when x is far from 0. In our application, we used the Gaussian

$$g(x) \propto e^{-\frac{x^2}{\gamma^2}}$$

where γ determines the width of the "window." Essentially, the WFT performs a Fourier transform on small time scales over the length of the signal. The *spectrogram* is defined as the absolute value of the WFT and can be more easily visualized.

Wavelets Wavelets are sets of orthonormal bases of functions $[0,1]$ generated by a *mother wavelet* $\psi : \Re \to \Re$ which has its support on $[0,1]$. Then we define

$$\psi_{a,b} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right).$$

We then define the wavelet transform

$$[W](a,b;\psi)f = \int_{\Re} \psi_{a,b}(x)f(x)dx.$$

As a complete basis, any particular mother wavelet will suffice to completely represent that data, but certain choices may be more amenable to efficient representation; i.e. $[W](a,b;\psi_1)f$

may be zero more often than $[W](a, b; \psi_2)f$. These transforms are similar to the WFT in that they "examine" the function over short periods of time.

Results

While the analysis did recommend particular transform and dimension reduction techniques, these results are secondary to the observation that an exhaustive search through the dimensionality p' will be infeasible for most applications and that a Bayesian approach is recommended. The computations themselves were intensive and occasionally required more than a day to complete; regardless of computing time, we have no precise estimates of the human time involved but believe it to be on the order of hundreds of hours.

This approach was admittedly intended to be excessively exhaustive; we suspected and proved that a particular wavelet respresentation would be most efficient out of the transforms we sampled. It was in addition fruitful; see for example [18], [15], [5]. However, in actual applications that do not intend to show the supremacy of an approach or transform, it can hardly be recommended.

We should however note that selection by (pairwise-)ANOVA showed far better performance, as was expected, than selection by variance. We therefore adopted a Bayesian approach to avoid the difficulties of exploring the dimensionality of the reduced set and adopted something like selection by ANOVA. This, then, was the major result of these studies, as it informs all of what follows.

CHAPTER 2
TWO MODELS

We introduce two Bayesian models for feature selection in high dimensional data, specifically for the purpose of classification. After Bayesian inference is implemented, we use Bayesian multiple testing procedure of Abramovich and Angelini[1]. We point out that although the inference is based on the known testing procedure, the Bayesian model formulation is entirely new.

2.1 General Framework

For convenience, we arrange row vectors

$$\mathbf{D}^\top = [\mathbf{d^1}, \mathbf{d^2}, \cdots, \mathbf{d^n}]$$

yielding the $(n \times p)$ matrix $\mathbf{D}$ and denote its columns by $\mathbf{d_i} \in \Re^{\mathbf{n}}$, $i = 1, \cdots, p$. The objective is to select a sparse subset of these p vectors which enable classification of vectors $\mathbf{d^1}, \mathbf{d^2}, \cdots, \mathbf{d^n}$ into classes $\omega_1, \cdots, \omega_L$. For this purpose, we introduce a binary vector $\mathbf{x} \in \Re^{\mathbf{p}}$ with $x_i = 1$ if vector component $\mathbf{d_i}$ is "informative" and should be retained in subsequent discriminatory analysis, or $x_i = 0$ if $\mathbf{d_i}$ should be discarded. The goal of the analysis, then, is to draw conclusions about vector $\mathbf{x}$ on the basis of matrix $\mathbf{D}$.

We introduce the following notations. Matrices and vectors are denoted in bold while their components are not. Let $\mathbf{e} \in \Re^{\mathbf{n}}$ be a column vector with unit components and $\mathbf{g_l} \in \Re^{\mathbf{n}}$,

$l = 1, \cdots, L$, be column vectors with the j-th components $(g_l)_j = 1$ if it corresponds to class l, i.e. if

$$n_1 + \cdots, n_{l-1} + 1 \leq j \leq n_1 + \cdots, n_{l-1} + n_l,$$

and $(g_l)_j = 0$ otherwise. We also define $\mathbf{G} \in \Re^{\mathbf{n \times L}}$ with columns $\mathbf{g_l}$, $l = 1, \cdots, L$.

In what follows, we consider Bayesian setup. Let $\mathbf{d_i}$ be a noisy measurement of the "true" i-th component $\boldsymbol{\mu}_i$, i.e.

$$\mathbf{d_i} = \boldsymbol{\mu_i} + \boldsymbol{\varepsilon_i} \tag{2.1}$$

where $\boldsymbol{\varepsilon}_i$ are multivariate normal $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2}\mathbf{I_n})$. The distribution of $\boldsymbol{\mu}_i$ depends on whether μ_i is informative ($x_i = 1$) or not ($x_i = 0$). We assume *a priori* that $x_1, \cdots, x_p$ are identically distributed and that the number of informative components

$$X = \sum_{i=1}^{p} x_i$$

is such that

$$P(X = k) = p(k) \geq 0 \quad \text{and} \quad \sum_{k=0}^{p} p(k) = 1.$$

If $x_1, \cdots, x_p$ are also independent with $P(x_i = 1) = \pi$, then X has the binomial distribution with parameters π and p. In general, we assume that $p(k)$ depends on parameter π, i.e. $p(k) = p_\pi(k)$.

2.1.1 VARSEL (Model 1)

In **VARSEL (Model 1)**, we assume that constant vectors are uninformative, i.e. for some scalar values m_i one has

$$\boldsymbol{\mu}_i = n^{-1/2} m_i \mathbf{e} + \mathbf{w_i}, \quad \mathbf{i} = \mathbf{1}, \cdots, \mathbf{p}, \tag{2.2}$$

where $\mathbf{e}^\top \mathbf{w_i} = \mathbf{0}$ and

$$\begin{aligned} m_i &\sim N(0, \sigma_i^2 \tau^2), \\ (\mathbf{w_i} | \mathbf{x_i} = \mathbf{0}) &\sim \delta(\mathbf{0}), \\ (\mathbf{w_i} | \mathbf{x_i} = \mathbf{1}) &\sim N(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2} \boldsymbol{\Sigma}_\mathbf{w}). \end{aligned} \tag{2.3}$$

Matrix $\boldsymbol{\Sigma}_w$ here characterizes correlation among the informative columns. Let $S_C = \text{Span}(\mathbf{e})$ be the linear subspace of constant vectors and $S_N = \Re^n \setminus S_C$ its complement in $\Re^n$. Then, matrices

$$\mathbf{P_C} = \mathbf{n^{-1}} \mathbf{e}^\top \mathbf{e}, \quad \text{and} \quad \mathbf{P_N} = \mathbf{I_n} - \mathbf{P_C} \tag{2.4}$$

are projection matrices for spaces S_C and S_N, respectively, and $\boldsymbol{\Sigma}_w = \mathbf{P_N} \boldsymbol{\Sigma} \mathbf{P_N}$ where $\sigma_i^2 \boldsymbol{\Sigma}$ is the covariance matrix of vector $\boldsymbol{\mu}_i$ given $x_i = 1$, $i = 1, \cdots, p$.

2.1.2 CONFESS (Model 2)

In **CONFESS (Model 2)**, we search for the vectors which are constant within but vary between the classes. In particular, let $S_G = \text{Span}(\mathbf{g_1}, \cdots, \mathbf{g_L})$ and let $S_1 = S_G \setminus S_C$ and $S_0 = \Re^n \setminus S_G$, so that $\Re^n = S_C \oplus S_0 \oplus S_1$. Then

$$\boldsymbol{\mu}_i = n^{-1/2} m_i \mathbf{e} + \mathbf{u_i} + \mathbf{v_i}, \quad \mathbf{i} = \mathbf{1}, \cdots, \mathbf{p}, \tag{2.5}$$

where $\mathbf{u_i} \in \mathbf{S_1}$ and $\mathbf{v_i} \in \mathbf{S_0}$. We assume that

$$\begin{aligned} m_i &\sim N(0, \sigma_i^2 \tau^2), \\ (\mathbf{u_i} | \mathbf{x_i} = \mathbf{0}) &\sim \delta(\mathbf{0}), \\ (\mathbf{u_i} | \mathbf{x_i} = \mathbf{1}) &\sim N(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2} \boldsymbol{\Sigma}_\mathbf{u}) \\ \mathbf{v_i} &\sim N(0, \sigma_i^2 \boldsymbol{\Sigma}_v), \end{aligned} \tag{2.6}$$

where inclusions $\mathbf{u_i} \in \mathbf{S_1}$ and $\mathbf{v_i} \in \mathbf{S_0}$ are enforced by covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_v$. Note that matrix

$$\mathbf{P_G} = \mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-\mathbf{1}} \mathbf{G}^\top$$

projects an arbitrary vector into S_G. Define matrices

$$\mathbf{P_0} = \mathbf{I_n} - \mathbf{P_G}, \ \mathbf{P_1} = \mathbf{P_G} - \mathbf{P_C}. \tag{2.7}$$

Lemma 1 *Matrices $\mathbf{P_C}$, $\mathbf{P_0}$ and $\mathbf{P_1}$ are projection matrices for linear spaces S_C, S_0 and S_1, respectively, so that $\boldsymbol{\Sigma}_u = \mathbf{P_1 \Sigma P_1}$ and $\boldsymbol{\Sigma}_v = \mathbf{P_0 \Sigma P_0}$ where $\sigma_i^2 \boldsymbol{\Sigma}$ is the covariance matrix of the vector $\boldsymbol{\mu}_i$ given $x_i = 1$, $i = 1, \cdots, p$.*

Proof By construction as orthogonal projections, matrices $\mathbf{P_C}$ and $\mathbf{P_G}$ are symmetric, idempotent, and identities on their respective subspaces $\mathbf{S_C}$ and $\mathbf{S_G}$. Then $\mathbf{P_G P_C} = \mathbf{P_C}$ since $S_C \subseteq S_G$. Also,

$$\mathbf{P_1 e} = \mathbf{P_G e} - \mathbf{P_C e} = \mathbf{e} - \mathbf{e} = \mathbf{0}$$

since $\mathbf{e} \in \mathbf{S_C}, \mathbf{S_G}$. Finally,

$$\mathbf{P_1 P_0} = \mathbf{P_0 P_1} = (\mathbf{I_n} - \mathbf{P_G})(\mathbf{P_G} - \mathbf{P_C}) = \mathbf{P_G} - \mathbf{P_C} - \mathbf{P_G P_G} + \mathbf{P_G P_C} = \mathbf{0}.$$

To complete the proof of the lemma, observe that $\mathbf{P_1 u_i} = \mathbf{u_i}$ since $\mathbf{u_i} \in \mathbf{S_1}$. □

2.2 Inference

To select "informative" vectors $\mathbf{d_i}$, one needs to evaluate $P(x_i = 1|\mathbf{D})$, $i = 1, \cdots, p$. However, for each $i = 1, \cdots, p$, only a part of each vector $\mathbf{d_i}$ carries information about x_i, in particular, the part associated with $\mathbf{w_i} \in \mathbf{S_N}$ for VARSEL (Model 1) and with $\mathbf{u_i} \in \mathbf{S_1}$ for CONFESS (Model 2). Therefore, one needs to extract "informative" parts from vectors $\mathbf{d_i}$. For this purpose, one needs to construct matrices $\mathbf{R} \in \Re^{\mathbf{n \times n}}$ and $\mathbf{Q} \in \Re^{\mathbf{n \times n}}$ with the following properties. Matrix $\mathbf{R}$ has $n^{-1/2}\mathbf{e}^\top$ as its first row and matrix $\mathbf{H_N} \in \Re^{\mathbf{(n-1) \times n}}$ as its next $(n-1)$ rows. Matrix $\mathbf{Q} \in \Re^{\mathbf{n \times n}}$ has $n^{-1/2}\mathbf{e}^\top$ as its first row, matrix $\mathbf{H_1} \in \Re^{\mathbf{(L-1) \times n}}$ as its next $(L-1)$ rows and matrix $\mathbf{H_0} \in \Re^{\mathbf{(n-L) \times n}}$ as its last $(n-L)$ rows.

In what follows, we extract $\mathbf{y_i} = \mathbf{H_N d_i}$ Model 1 or $\mathbf{y_i} = \mathbf{H_1 d_i}$ and $\mathbf{z_i} = \mathbf{H_0 d_i}$ for Model 2, and show that model selection is carried out on the basis of vectors $\mathbf{y_i}$ only. For this to

be true, one needs matrices $\mathbf{R}$ and $\mathbf{Q}$ for Models 1 and 2, respectively, with the properties stated below.

Proposition 1 *Let matrix* $\mathbf{R} \in \Re^{\mathbf{n}\times\mathbf{n}}$ *described above be such that* $\mathbf{R}\mathbf{R}^\top = \mathbf{I_n}$ *and also* $\mathbf{H_N^\top H_N} = \mathbf{P_N}$. *Let*

$$y_{i0} = n^{-1/2}\mathbf{e}^\top \mathbf{d_i}$$

$$\mathbf{y_i} = \mathbf{H_N d_i}$$

Then, under conditions (2.1), (2.2) and (2.3), y_{i0} *and* $\mathbf{y_i}$ *are independent with*

$$\begin{aligned} y_{i0} &\sim N(0, \sigma_i^2(1+\tau^2)), \\ (\mathbf{y_i}|\mathbf{x_i} = \mathbf{0}) &\sim N(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2}\mathbf{I_{n-1}}), \\ (\mathbf{y_i}|\mathbf{x_i} = \mathbf{1}) &\sim N(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2}(\mathbf{I_{n-1}} + \mathbf{H_N \Sigma H_N^\top})). \end{aligned} \tag{2.8}$$

Also, y_{i0} *is independent of* $\mathbf{y_i}$, $i = 1, \cdots, p$.

Proof Note that $\mathbf{R}^\top\mathbf{R} = \mathbf{I_n}$ implies that matrix $\mathbf{H_N}$ satisfies (2.12). Then, for any i, one has

$$\mathbf{y_i} = \mathbf{H_N w_i} + \mathbf{H_N}\epsilon_\mathbf{i}$$

where

$$\mathbf{H_N}\epsilon_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2}\mathbf{H_N H_N^\top}) \sim \mathbf{N}(\mathbf{0}, \sigma_\mathbf{i}^\mathbf{2}\mathbf{I_{n-1}})$$

Similarly,

$$y_{i0} = m_i + n^{-1/2}\,\mathbf{e}^\top\epsilon_\mathbf{i}$$

$$n^{-1/2}\,\mathbf{e}^\top \epsilon_{\mathbf{i}} \sim \mathbf{N}(\mathbf{0}, \sigma_{\mathbf{i}}^{\mathbf{2}})$$

Validity of formulae (2.8) can be checked by direct calculations. Also, we have the covariance between y_{i0} and $\mathbf{y_i}$

$$\mathrm{Cov}(y_{i0}, \mathbf{y_i}) = \mathbb{E}(\mathbf{y_{i0}y_i}) = \mathbf{n^{-1/2}}\,\mathbf{e}^\top \mathbb{E}(\mathbf{d_i d_i^\top})\mathbf{H_N} = \mathbf{0},$$

and since the vector $y_{i0}, \mathbf{y_i}$ is normally distributed, y_{i0} and $\mathbf{y_i}$ are independent. □

Proposition 2 *Let matrix* $\mathbf{Q} \in \Re^{\mathbf{n}\times\mathbf{n}}$ *described above be such that*

$$\mathbf{QQ}^\top = \mathbf{I_n}$$

$$\mathbf{H_1^\top H_1} = \mathbf{P_1}$$

$$\mathbf{H_0^\top H_0} = \mathbf{P_0}$$

Let

$$y_{i0} = n^{-1/2}\mathbf{e}^\top \mathbf{d_i}$$

$$\mathbf{y_i} = \mathbf{H_1 d_i}$$

$$\mathbf{z_i} = \mathbf{H_0 d_i}$$

Then, under conditions (2.1), (2.5) and (2.6), one has

$$\begin{aligned} y_{i0} &\sim N(0, \sigma_i^2(1+\tau^2)), \\ (\mathbf{y_i}|\mathbf{x_i} = \mathbf{0}) &\sim N(\mathbf{0}, \sigma_{\mathbf{i}}^{\mathbf{2}}\mathbf{I_{L-1}}), \\ (\mathbf{y_i}|\mathbf{x_i} = \mathbf{1}) &\sim N(\mathbf{0}, \sigma_{\mathbf{i}}^{\mathbf{2}}(\mathbf{I_{L-1}} + \mathbf{H_1 \Sigma H_1^\top})), \\ \mathbf{z_i} &\sim N(\mathbf{0}, \sigma_{\mathbf{i}}^{\mathbf{2}}(\mathbf{I_{n-L}} + \mathbf{H_0 \Sigma H_0^\top})), \end{aligned} \tag{2.9}$$

Also, y_{i0}, $\mathbf{y_i}$ *and* $\mathbf{z_i}$ *are independent,* $i = 1, \cdots, p$.

Proof The proof is very similar to the proof of Proposition 1. □

Note that matrices $\mathbf{R}$ and $\mathbf{Q}$ carry out orthogonal transformation of the data $\mathbf{d_i}$, $i = 1, \cdots, p$, replacing vectors $\mathbf{d_i}$ by y_{i0} and $\mathbf{y_i}$ for Model 1 and y_{i0}, $\mathbf{y_i}$ and $\mathbf{z_i}$ for Model 2, $i = 1, \cdots, p$. It follows from Propositions 1 and 2 that vectors $\mathbf{y_i}$ alone contain information about x_i. With some abuse of notation, denote

$$\mathbf{I} = \mathbf{I_{n-1}}, \quad \mathbf{\Sigma_y} = \mathbf{H_N \Sigma H_N^\top}, \quad \mathbf{z_i} = \mathbf{0}$$

and $\mathbf{\Sigma}_z = 1$ for Model 1 and

$$\mathbf{I} = \mathbf{I_{L-1}}, \quad \mathbf{\Sigma_y} = \mathbf{H_1 \Sigma H_1^\top}, \quad \mathbf{Z} = \{\mathbf{z_1}, \cdots, \mathbf{z_p}\}$$

$$\mathbf{\Sigma}_z = \mathbf{I_{n-L}} + \mathbf{H_0 \Sigma H_0^\top}$$

for Model 2. Let $\mathbf{\Omega}$ be a diagonal matrix

$$\mathbf{\Omega} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2).$$

Then, since all configurations of zeros and ones in vector $\mathbf{x}$ are *a priori* equally likely, the joint pdf of

$$\mathbf{Y_0} = (\mathbf{y_{10}}, \mathbf{y_{20}}, \cdots, \mathbf{y_{p0}})$$

$$\mathbf{Y} = \{\mathbf{y_1}, \cdots, \mathbf{y_p}\}$$

$$\mathbf{Z} = \{\mathbf{z_1}, \cdots, \mathbf{z_p}\}$$

and $\mathbf{x}$ is of the form

$$p(\mathbf{Y_0}, \mathbf{Y}, \mathbf{Z}, \mathbf{x}|\mathbf{\Omega}, \mathbf{\Sigma_y}, \mathbf{\Sigma_z}, \tau, \pi) = \binom{\mathbf{p}}{\mathbf{k}}^{-\mathbf{1}} \mathrm{I}(\mathbf{X} = \mathbf{k})\mathbf{p}_\pi(\mathbf{k})\ (\mathbf{2}\pi)^{-\mathbf{np/2}}(\mathbf{1}+\tau^{\mathbf{2}})^{-\mathbf{p/2}}|\mathbf{\Omega}|^{-\mathbf{n/2}}|\mathbf{I}+\mathbf{\Sigma_y}|^{-\mathbf{k/2}}$$
$$\times\, |\mathbf{\Sigma}_z|^{-p/2} \exp\left\{ -\sum_{i=1}^{p} \frac{1}{2\sigma_i^2} \left[(1+\tau^2)^{-1} y_{i0}^2 + x_i\, \mathbf{y_i}^\top (\mathbf{I}+\mathbf{\Sigma_y})^{-\mathbf{1}} \mathbf{y_i} + (\mathbf{1}-\mathbf{x_i}) \mathbf{y_i}^\top \mathbf{y_i} + \mathbf{z_i}^\top \mathbf{\Sigma_z^{-1}} \mathbf{z_i} \right] \right\}.$$

The posterior distribution of each configuration $\mathbf{x}$ is

$$p(\mathbf{x}, \mathbf{k}|\mathbf{Y}, \mathbf{\Omega}, \mathbf{\Sigma_y}, \pi) \propto \binom{\mathbf{p}}{\mathbf{k}}^{-\mathbf{1}} \mathrm{I}(\mathbf{X} = \mathbf{k})\mathbf{p}_\pi(\mathbf{k})\ |\mathbf{I}+\mathbf{\Sigma_y}|^{-\mathbf{k/2}}\ \exp\left\{ -\sum_{\mathbf{i=1}}^{\mathbf{p}} \frac{\mathbf{1}}{\mathbf{2}\sigma_\mathbf{i}^\mathbf{2}} \mathbf{x_i}\, \mathbf{y_i}^\top (\mathbf{I}+\mathbf{\Sigma_y^{-1}})^{-\mathbf{1}} \mathbf{y_i} \right\},$$

and is independent of $\mathbf{Y_0}$ and $\mathbf{Z}$. Following Abramovich and Angelini[1], we apply a maximum *a posteriori* (MAP) rule to choose the most likely configuration of zeros and ones in vector $\mathbf{x}$. The MAP rule implies that, for a given value of k, $\hat{x}_i = 1$ for the k largest values of Δ_i where

$$\Delta_i = \sigma_i^{-2}\, \mathbf{y_i}^\top (\mathbf{I}+\mathbf{\Sigma_y^{-1}})^{-\mathbf{1}} \mathbf{y_i}, \quad \mathbf{i} = \mathbf{1}, \cdots, \mathbf{p}, \tag{2.10}$$

and $\hat{x}_i = 0$ otherwise. Let $\Delta_{(i)}$ be the i-th largest value, i.e. $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots \geq \Delta_{(p)}$. Then, denoting the chosen configuration by $\hat{\mathbf{x}}(k)$, we derive the MAP value of k:

$$\hat{k} = \arg\max_k \left[2 \left(\ln \binom{p}{k}^{-1} + \ln p_\pi(k) - k \ln |\mathbf{\Sigma}_y + \mathbf{I}| \right) + \sum_{i=1}^{k} \Delta_{(i)} \right]. \tag{2.11}$$

In order to carry out model selection according to (2.10) and (2.11), one needs to construct matrices $\mathbf{R}$ and $\mathbf{Q}$ described above and estimate unknown parameters $\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2, \pi$ and unknown matrix $\mathbf{\Sigma}_y$.

2.2.1 Construction of Matrices R and Q

To construct matrices $\mathbf{H_N}$, $\mathbf{H_1}$ and $\mathbf{H_0}$ satisfying Propositions 1 and 2, we need to construct matrix $\mathbf{H_N}$ for Model 1 and matrices $\mathbf{H_1}$ and $\mathbf{H_0}$ for Model 2 with the following properties

$$\mathbf{H_N e} = \mathbf{0}, \quad \mathbf{H_N H_N^\top} = \mathbf{I_{n-1}}, \quad \mathbf{H_N^\top H_N} = \mathbf{P_N} \tag{2.12}$$

$$\mathbf{H_1 e} = \mathbf{0},\ \mathbf{H_0 e} = \mathbf{0},\ \mathbf{H_1 H_0^\top} = \mathbf{0},\ \mathbf{H_1 H_1^\top} = \mathbf{I_{L-1}},\ \mathbf{H_0 H_0^\top} = \mathbf{I_{n-L}},\ \mathbf{H_1^\top H_1} = \mathbf{P_1},\ \mathbf{H_0^\top H_0} = \mathbf{P_0}, \tag{2.13}$$

Since the first rows of both matrices, $\mathbf{R}$ and $\mathbf{Q}$, is $n^{-1/2}\mathbf{e}^\top$.

For this purpose, we introduce diagonal $n \times n$ matrices $\mathbf{\Lambda}_N$, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_0$ where $\mathbf{\Lambda}_N$ has $(n-1)$ consecutive ones and a zero on the diagonal, $\mathbf{\Lambda}_1$ has $(L-1)$ consecutive ones and $(n-L+1)$ zeros on their diagonal and, finally, $\mathbf{\Lambda}_0$ has $(L-1)$ consecutive zeros followed by $(n-L)$ consecutive ones and then a zero on the diagonal. Introduce also matrices $\mathbf{T_N} \in \Re^{\mathbf{(n-1)\times n}}$ with $\mathbf{I_{n-1}}$ in its first $(n-1)$ columns, the rest being identically zero, $\mathbf{T_1} \in \Re^{\mathbf{(L-1)\times n}}$ with $\mathbf{I_{L-1}}$ in its first $(L-1)$ columns, the rest being identically zero, and $\mathbf{T_0} \in \Re^{\mathbf{(n-L)\times n}}$ with $(L-1)$ first columns being identically zero, then matrix $\mathbf{I_{n-L}}$ in the next $(n-L)$ columns and the last column being zero. By construction, we have

$$\mathbf{T_N T_N^\top} = \mathbf{I_{n-1}}, \quad \mathbf{T_1 T_1^\top} = \mathbf{I_{L-1}}, \quad \mathbf{T_0 T_0^\top} = \mathbf{I_{n-L}},$$

$$\mathbf{T_N^\top T_N} = \mathbf{\Lambda}_N, \quad \mathbf{T_1^\top T_1} = \mathbf{\Lambda_1}, \quad \mathbf{T_0^\top T_0} = \mathbf{\Lambda_0}. \tag{2.14}$$

Now, recall that $\mathbf{P_N}$ is a symmetric, idempotent matrix of rank $(n-1)$; hence, there exists an orthogonal matrix $\mathbf{U}$ such that $\mathbf{P_N} = \mathbf{U}^\top \mathbf{\Lambda_N} \mathbf{U}$. Let

$$\mathbf{H_N} = \mathbf{T_N} \mathbf{U}. \tag{2.15}$$

Then,

$$\mathbf{H_N} \mathbf{H_N^\top} = \mathbf{T_N} \mathbf{U} \mathbf{U}^\top \mathbf{T_N^\top} = \mathbf{I_{n-1}}$$

$$\mathbf{H_N^\top} \mathbf{H_N} = \mathbf{U}^\top \mathbf{T_N^\top} \mathbf{T_N} \mathbf{U} = \mathbf{P_N}$$

due to (2.14). Also, observe that

$$\|\mathbf{H_N e}\|^2 = \mathbf{e}^\top \mathbf{H_N^\top} \mathbf{H_N e} = \|\mathbf{P_N e}\|^2 = \mathbf{0},$$

so that $\mathbf{H_N e} = \mathbf{0}$, and $\mathbf{H_N}$ satisfies all conditions of Proposition 1.

To construct matrices $\mathbf{H_1}$ and $\mathbf{H_0}$, note that according to formula (2.7) and Lemma 1, matrices $\mathbf{P_1}$ and $\mathbf{P_0}$ are symmetric idempotent matrices of ranks $(L-1)$ and $(n-L)$, respectively, and commute pairwise, i.e. $\mathbf{P_1 P_0} = \mathbf{P_0 P_1} = \mathbf{0}$. For this reason, matrices $\mathbf{P_0}$ and $\mathbf{P_1}$ are simultaneously diagonalizable (see for example [46] pg. 192) and there exists an orthogonal matrix $\mathbf{V}$ such that

$$\mathbf{P_1} = \mathbf{V}^\top \mathbf{\Lambda_1} \mathbf{V}, \quad \mathbf{P_0} = \mathbf{V}^\top \mathbf{\Lambda_0} \mathbf{V}.$$

Construct matrices $\mathbf{H_1}$ and $\mathbf{H_0}$ as

$$\mathbf{H_1} = \mathbf{T_1} \mathbf{V}, \quad \mathbf{H_0} = \mathbf{T_0} \mathbf{V}. \tag{2.16}$$

Then, the last four equalities in (2.13) can be verified using (2.14) in a manner similar to the proof for $\mathbf{H_N}$. To show that the first two equalities in (2.13) hold, recall that

$$\mathbf{P_G}\,\mathbf{e} = \mathbf{P_C}\,\mathbf{e} = \mathbf{e}$$

so that

$$\|\mathbf{H_1 e}\|^{\mathbf{2}} = \mathbf{e}^\top \mathbf{P_1 e} = \mathbf{e}^\top \mathbf{P_G}\,\mathbf{e} - \mathbf{e}^\top \mathbf{P_C}\,\mathbf{e} = \mathbf{0}$$

and similarly, $\|\mathbf{H_0 e}\|^{\mathbf{2}} = \mathbf{0}$. Now, the remaining equality $\mathbf{H_1 H_0^\top} = \mathbf{0}$ follows directly from $\mathbf{T_1 T_0^\top} = \mathbf{0}$, and, hence, matrices $\mathbf{P_1}$ and $\mathbf{P_0}$ satisfy all conditions of Proposition 2.

We should mention here that some versions of matrices $\mathbf{R}$ and $\mathbf{Q}$ can be constructed explicitly. In particular, matrix $\mathbf{R}$ can be the n-dimensional Helmert matrix, so that the $\mathbf{H_N} = \mathcal{H}(\mathbf{n})$ with elements $(\mathcal{H}(\mathbf{n}))_{\mathbf{ji}}$ of the form

$$(\mathbf{\Pi}(n))_{ji} = \begin{cases} [j(j+1)]^{-1/2}, & 1 \le j \le n, 1 \le i \le j, \\ -[j/(j+1)]^{1/2}, & 1 \le j \le n, i = j+1, \\ 0, & 1 \le j \le n, j+2 \le i \le n. \end{cases}$$

Matrix $\mathbf{H_0}$ can be constructed as a block matrix with matrix $\mathcal{H}(\mathbf{n_1})$ in the first $(n_1 - 1)$ rows, $\mathcal{H}(\mathbf{n_2})$ in the next $(n_2 - 1)$ rows and so on. Matrix $\mathbf{H_1}$ can be recovered by direct Gram-Schmidt orthogonalization. Set $n_0 = 0$. Then, elements of matrix $\mathbf{H_1}$ are of the form:

$$(\mathbf{H_1})_{\mathbf{ji}} = \begin{cases} 0, & 1 \le i \le n_1 + \cdots + n_{j-1}, \\ h_{jj}, & n_1 + \cdots + n_{j-1} + 1 \le i \le n_1 + \cdots + n_j, \\ h_{j,j+1}, & i > n_1 + \cdots + n_j, 1 \le j \le L-1. \end{cases}$$

Here,

$$\begin{aligned} h_{jj} &= \sqrt{(n-n_1-\cdots-n_j)/[n_j(n-n_1-\cdots-n_{j-1})]}, \quad 1 \le j \le L-1, l=j, \\ h_{j,j+1} &= \sqrt{n_j/[(n-n_1-\cdots-n_j)(n-n_1-\cdots-n_{j-1})]}, \quad 1 \le j \le L-1, j+1 \le l \le L. \end{aligned}$$

Note that formulae (2.15) and (2.16) deliver some orthogonal transformations of the explicit forms mentioned above.

2.3 Estimation of Parameters

Observe that both models reduce to

$$\begin{aligned} y_{i0} &\sim N(0, \sigma_i^2(1+\tau^2)), \\ (\mathbf{y_i}|\mathbf{x_i}=\mathbf{0}) &\sim N(\mathbf{0}, \boldsymbol{\sigma_i^2}\mathbf{I_m}), \\ (\mathbf{y_i}|\mathbf{x_i}=\mathbf{1}) &\sim N(\mathbf{0}, \boldsymbol{\sigma_i^2}(\mathbf{I_m}+\boldsymbol{\Sigma_y})), \\ \mathbf{z_i} &\sim N(\mathbf{0}, \boldsymbol{\sigma_i^2}(\mathbf{I_r}+\boldsymbol{\Sigma_z})), \end{aligned} \tag{2.17}$$

where $m = n-1$, $\boldsymbol{\Sigma}_y = \mathbf{H_N}\boldsymbol{\Sigma}\mathbf{H_N^\top}$, $z_i = 0$ and $r = 0$ for Model 1, and $m = L-1$, $r = n-L$, $\boldsymbol{\Sigma}_y = \mathbf{H_1}\boldsymbol{\Sigma}\mathbf{H_1^\top}$ and $\boldsymbol{\Sigma}_z = \mathbf{H_0}\boldsymbol{\Sigma}\mathbf{H_0^\top}$ for Model 2. To apply the model selection procedure described above, one needs to estimate unknown parameters σ_i^2, $i = 1, \cdots, p$, matrix $\boldsymbol{\Sigma}_y$ and parameter π associated with the prior $p_\pi(k)$. Note that for model selection one does not need to know τ or $\boldsymbol{\Sigma}_z$.

Since vector $\mathbf{x}$ is unknown, a single-step estimation of the parameters is usually intractable. Therefore, we treat vector $\mathbf{x}$ a the latent variable in an EM algorithm, alter-

nating between computing the expectation of log-likelihood, given transformed data $\mathbf{y_i}$, $i = 1, \cdots, p$, and values of parameters (E-step), and estimating parameters by maximizing the expected value of the log-likelihood (M-step). The algorithm begins with initial values $\sigma^2_{i,[0]}$, $i = 1, \cdots, p$, matrix $\mathbf{\Sigma}_{y,[0]}$ and parameter $\pi_{[0]}$.

Then, given values of unknown parameters, $\sigma^2_{i,[h]}$, $i = 1, \cdots, p$, $\mathbf{\Sigma}_{y,[h]}$ and $\pi_{[h]}$ at the h-th iteration of the algorithm, at an E-step, one needs to find the posterior expectation of the latent vector $\mathbf{x}$ given the data $\mathbf{y}$. If the number k of nonzero components of vectors $\mathbf{x}$ has binomial distribution (and, thus, components x_i are independent), then, following Abramovich and Angelini[1], one can find posterior expectations $\hat{x}_i$ of x_i given $\mathbf{y_i}$ as $\hat{x}_i = (1 + B_i(\mathbf{y_i}))^{-\mathbf{1}}$ where $B_i(\mathbf{y_i})$ are Bayes factors

$$B_i(\mathbf{y_i}) = \frac{\mathbf{p}(\mathbf{y_i}|\mathbf{x_i} = \mathbf{0})(\mathbf{1} - \pi_{[\mathbf{h}]})}{\mathbf{p}(\mathbf{y_i}|\mathbf{x_i} = \mathbf{1})\pi_{[\mathbf{h}]}} = |\mathbf{I_m} + \mathbf{\Sigma_{y,[h]}}|^{\mathbf{1/2}} \exp\left\{-\frac{\mathbf{y_i^\top}(\mathbf{I} + \mathbf{\Sigma^{-1}_{y,[h]}})^{-\mathbf{1}}\mathbf{y_i}}{\mathbf{2}\sigma^{\mathbf{2}}_{\mathbf{i},[\mathbf{h}]}}\right\}.$$

Alternatively, if $p_\pi(k)$ is not binomial, following George and Foster[25], one can replace the posterior mean estimators of x_i's by the posterior mode, which leads to choosing $x_i = 1$ for k largest values of Δ_i and then estimating k by (2.11).

At an M-step, one needs to maximize the log-likelihood of the entries σ_i^2, $i = 1, \cdots, p$ of the diagonal matrix $\mathbf{\Omega}$, of matrix $\mathbf{\Sigma}_y$ and parameter π given the data and the latent vector $\mathbf{x}$:

$$l(\mathbf{\Omega}, \mathbf{\Sigma}_y, \pi; \mathbf{Y}, \mathbf{x}) = \text{const} + \log \mathbf{p}_\pi(\mathbf{k}) - \log \binom{\mathbf{p}}{\mathbf{k}} + \log \mathrm{I}(\sum \mathbf{x_i} = \mathbf{k}) - \frac{\mathbf{k}}{\mathbf{2}} \log |\mathbf{I_m} + \mathbf{\Sigma_y}|$$
$$- \sum_{i=1}^{p} \left[\frac{n \log(\sigma_i^2)}{2} + \frac{x_i \mathbf{y_i^\top}(\mathbf{I_m} + \mathbf{\Sigma_y^{-1}})^{-\mathbf{1}}\mathbf{y_i} + (\mathbf{1} - \mathbf{x_i})\mathbf{y_i^\top}\mathbf{y_i}}{2\sigma_i^2} \right]. \tag{2.18}$$

Since the general model is too diverse and includes a very large number of parameters, we consider two special cases of the general model above, Case 1: all σ_i's are equal to each other: $\sigma_i = \sigma$, $i = 1, \cdots, p$, and Case 2: matrix $\mathbf{\Sigma}$ is proportional to identity: $\mathbf{\Sigma} = \rho^2 \mathbf{I_n}$.

Case 1: $\sigma_i = \sigma$, $i = 1, \cdots, p$. If $k << p$, then one can estimate σ by a variety of methods; e.g. the median of the absolute deviations of $\mathbf{y_i}$ divided by 0.6745 (Donoho and Johnstone[20]). If assumption $k << p$ does not hold, then σ^2 can be estimated by

$$\hat{\sigma^2} \;=\; \frac{1}{mp} \sum_{i=1}^{p} \left[\hat{x}_i \mathbf{y_i}^\top ((\mathbf{I_m} + \hat{\mathbf{\Sigma}}_\mathbf{y})^{-1} \mathbf{y_i} \right].$$

Then, matrix $\mathbf{\Sigma}_y$ is estimated by

$$\widehat{\mathbf{\Sigma}_y} \;=\; \left(\frac{1}{k\,\hat{\sigma}^2} \sum_{i=1}^{p} \hat{x}_i \mathbf{y_i} \mathbf{y_i}^\top - \mathbf{I_m} \right)_+ .$$

For now, it can be assumed that $(\mathbf{A})_+$ is a positive semidefinite matrix close to $\mathbf{A}$ in way we will describe later. However, if $\mathbf{\Sigma}_y$ is assumed to be diagonal, then its diagonal entries are estimated by

$$(\widehat{\Sigma_y})_{ii} = (k^{-1} \hat{\sigma}^{-2} \sum_{i=1}^{p} \hat{x}_i \mathbf{y_i}^\top \mathbf{y_i} - \mathbf{1})_+ \tag{2.19}$$

where $t_+ = \max(t, 0)$ for any $t \in \Re$.

Case 2: $\mathbf{\Sigma} = \rho^2 \mathbf{I_n}$. It follows from Propositions 1 and 2 that $\mathbf{I_m} + \mathbf{\Sigma_y} = (\mathbf{1} + \rho^\mathbf{2})\mathbf{I_m}$. Hence, in Model 1, maximization of (2.18) with respect to σ_i^2 and ρ yields

$$\hat{\sigma}_i^2 = \frac{\mathbf{y_i}^\top \mathbf{y_i}}{n-1} \left[\frac{x_i}{1+\hat{\rho}^2} + (1 - x_i) \right]$$

where $\hat{\rho}$ is the solution of the equation

$$\rho^2 = \left(\frac{1}{k} \sum_{i=1}^{p} \frac{x_i (1+\rho^2)}{x_i + (1+\rho^2)(1-x_i)} \;-1 \right)_+ .$$

In Model 2, one can use the separate portion of likelihood associated with $\mathbf{z_i}$, for estimation of $\alpha_i = (1+\rho^2)\sigma_i^2$, $i = 1, \cdots, p$. Since $\mathbf{z_i} \sim \mathbf{N}(\mathbf{0}, \alpha_\mathbf{i}\mathbf{I_{n-L}})$, one can estimate α_i by $\hat{\alpha}_i = (n-L)^{-1}\mathbf{z_i}^\top \mathbf{z_i}$. Then, ρ^2 can be estimated by plugging $\hat{\alpha}_i$, $i = 1, \cdots, p$, into (2.18) and maximizing (2.18) with respect to ρ for a given vector $\mathbf{x}$. Therefore,

$$\hat{\rho} = \left(\sum_{i=1}^{p} \hat{\alpha}_i^{-1}(1-x_i)\mathbf{y_i}^\top \mathbf{y_i} \right) \left((L-1)(p-k) - \sum_{i=1}^{p} \hat{\alpha}_i^{-1}(1-x_i)\mathbf{y_i}^\top \mathbf{y_i} \right)_+ .$$

Then, σ_i^2 are estimated by

$$\hat{\sigma}_i^2 = \left[(1+\hat{\rho}^2)(n-L)\right]^{-1} \mathbf{z_i}^\top \mathbf{z_i}. \tag{2.20}$$

After $\mathbf{\Omega}$ and $\mathbf{\Sigma}_y$ are estimated, parameter π can be found as a solution of the following optimization problem

$$\hat{\pi} = \arg\max_{\pi} \; l(\hat{\mathbf{\Omega}}, \hat{\mathbf{\Sigma}}_y, \pi; \mathbf{y}, \hat{\mathbf{x}}),$$

where $l(\mathbf{\Omega}, \mathbf{\Sigma}_y, \pi; \mathbf{Y}, \mathbf{x})$ is defined in (2.18).

CHAPTER 3
MODEL SELECTION IN CONFESS

Assume we have data generated according to CONFESS, in which $\sigma_i = \sigma_j$ for all i, j, $\mathbf{\Sigma} = \rho^2\mathbf{I}$, and all parameters but x_i are known. Note that the assumption that $\mathbf{\Sigma}$ is diagonal. This is not merely for convenience; we recall the naive Bayes classifier of Bickel and Levina[9]. Let i be some column number. We return to (2.17) and define

$$Y_i = \frac{||y_i||^2}{\sigma_i^2(1+\rho^2)}.$$

Then, since in this model

$$\begin{aligned}(y_i|x_i = 0) &\sim N(0, \mathbf{I_{L-1}}) \\ (y_i|x_i = 1) &\sim N(0, (1+\rho^2)\mathbf{I_{L-1}})\end{aligned}$$

we conclude that

$$\begin{aligned}(Y_i|x_i = 0) &\sim \frac{\chi^2_{L-1}}{1+\rho^2} \\ (Y_i|x_i = 1) &\sim \chi^2_{L-1}. \end{aligned} \tag{3.1}$$

Now, since $\rho^2 \geq 0$, we have $(Y_i|x_i = 1) \geq (Y_i|x_i = 0)$ in the stochastic sense; i.e. for all λ we have

$$P(Y_i \leq \lambda|x_i = 1) \leq P(Y_i \leq \lambda|x_i = 0).$$

If there were to exist a λ for which the supports of Y_i under the two cases $x_i = 0, 1$ were on separate sides of λ, we could immediately recover x_i from an observation and model selection

would be trivial; i.e. if $Y_i \geq \lambda$, then we would correctly conclude that $x_i = 1$. However, we note that under either hypothesis, $x_i = 0, 1$, Y_i has support on all of $[0, \infty)$. This implies that there exists no λ which completely separates the two cases.

3.1 Separability

We consider a set of data sets $\{\mathbf{D_p}\}$ generated according to CONFESS in which $p \to \infty$. All parameters are known but are allowed to evolve with p. We will often suppress the indices for clarity. For each data set we define the random variables

$$\begin{aligned} U_p &= \max\{Y_i | x_i = 0\}_{i=1}^{p_0} \\ V_p &= \min\{Y_i | x_i = 1\}_{i=1}^{p_1}. \end{aligned}$$

Again, U_p and V_p have support on all of $[0, \infty)$, so there is no λ that separates them. However, we call $\{\mathbf{D_p}\}$ *separable* if there exists a sequence $\{\lambda_p\}$ such that

$$\begin{aligned} \lim_{p \to \infty} P(U_p \leq \lambda_p) &= 1 \\ \lim_{p \to \infty} P(V_p \leq \lambda_p) &= 0. \end{aligned}$$

If a set of data sets is separable, the probability of both types of errors goes to 0. In applications no data set will be of infinite dimensionality, but for "large" data sets, which will be quantified later, the idea of separability can imply that the probability of either type of errors is small.

Lemma 2 *The conditions for separability in CONFESS are*

$$\lim_{p\to\infty} \quad p_0(1 - F_{L-1}((1+\rho^2)\lambda)) = 0$$

$$\lim_{p\to\infty} \quad p_1 F_{L-1}(\lambda) = 0$$

where $F_{L-1}(x)$ is the CDF of the χ^2 at x with $L-1$ degrees of freedom written as

$$F_{L-1}(x) = \frac{\gamma\left(\frac{L-1}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{L-1}{2}\right)} \tag{3.2}$$

where $\gamma(.,.)$ is the (lower) incomplete gamma function

$$\gamma(a,z) = \int_0^z t^{a-1}e^{-t}dt$$

Proof Assume first that the data are separable. We note that

$$\begin{aligned} F_{V_p}(\lambda) &= P(V_p \le \lambda) = P(\max\{Y_i|x_i = 0\}_{i=1}^{p_0} \le \lambda) \\ &= \prod_{i=1}^{p_0} P(Y_i \le \lambda|x_i = 0) \\ &= F_{L-1}((1+\rho^2)\lambda)^{p_0}. \end{aligned} \tag{3.3}$$

Similarly,

$$\begin{aligned} F_{U_p}(\lambda) &= P(U_p \le \lambda) = P(\min\{Y_i|x_i = 1\}_{i=1}^{p_1} \le \lambda) \\ &= 1 - P(\min\{Y_i|x_i = 1\}_{i=1}^{p_1} > \lambda) \\ &= 1 - \prod_{i=1}^{p_1} P(Y_i > \lambda|x_i = 1) \\ &= 1 - (1 - F_{L-1}(\lambda))^{p_1}. \end{aligned} \tag{3.4}$$

By the definition of separability, the conditions in this case are:

$$\begin{aligned} \lim_{p\to\infty} \quad & F_{L-1}((1+\rho^2)\lambda)^{p_0} \quad = 1 \\ \lim_{p\to\infty} \quad & (1-F_{L-1}(\lambda))^{p_1} \quad = 1. \end{aligned}$$

We note that these are equivalent to

$$\begin{aligned} \lim_{p\to\infty} p_0 \ln F_{L-1}((1+\rho^2)\lambda) &= 0 \\ \lim_{p\to\infty} p_1 \ln(1-F_{L-1}(\lambda)) &= 0 \end{aligned}$$

i.e. the previous conditions hold iff these conditions hold. Now, we recall that

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}$$

for all $|x| < 1$ and $|F_{L-1}(\lambda)| < 1$ necessarily, so that

$$p_1 \ln(1-F_{L-1}(\lambda)) = -\sum_{n=1}^{\infty} \frac{p_1 (F_{L-1}(\lambda))^n}{n}$$

Now, by hypothesis,

$$p_1 \ln(1-F_{L-1}(\lambda)) \to 0$$

so that

$$\sum_{n=1}^{\infty} \frac{p_1 (F_{L-1}(\lambda))^n}{n} \to 0$$

as well. Now, since all the summands are positive, each summand must also go to 0, so that $p_1 F_{L-1}(\lambda) \to 0$ as $p \to \infty$. The same argument applies when $x_i = 0$, writing

$$F_{L-1}((1+\rho^2)\lambda) = 1 - (1 - F_{L-1}((1+\rho^2)\lambda))$$

where necessary.

Then the formal definition of separability implies that

$$\lim_{p\to\infty} p_0(1 - F_{L-1}((1+\rho^2)\lambda)) = 0 \tag{3.5}$$

$$\lim_{p\to\infty} p_1 F_{L-1}(\lambda) = 0 \tag{3.6}$$

i.e. these two limits are necessary if the data are separable.

Assume next that both (3.5) and (3.6) hold. Necessarily $F_{L-1}(\lambda) < 1$ so that

$$\sum_{n=1}^{\infty} \frac{(F_{L-1}(\lambda))^{n-1}}{n} < \infty$$

since it is dominated by a convergent geometric series. Then we have

$$p_1 F_{L-1}(\lambda) \sum_{i=1}^{\infty} \frac{(F_{L-1}(\lambda))^{n-1}}{n} = \sum_{n=1}^{\infty} \frac{p_1 (F_{L-1}(\lambda))^{n}}{n} \to 0$$

as $p \to \infty$. However, this is the statement that $\lim_{p\to\infty} p_1 \ln(1 - F_{L-1}(\lambda)) = 0$ and we obtain the formal definition of separability by exponentiating. The same argument can be applied to the case in which $x_i = 0$. □

3.1.1 Asymptotic Expansion of the Logarithm of Gamma Func's

It follows from (3.2) that $F_{L-1}(\lambda)$ and $F_{L-1}((1+\rho^2)\lambda)$ can be expressed via the gamma function and the incomplete gamma function. Were either L and λ fixed, we could easily apply the l'Hospital rule (or standard expansion of $F_n(x)$) and obtain some list of situations in which (3.5) and (3.6) held. We need an expansion $F_n(x)$ valid for both $n, x \to \infty$, since p_1, L, λ and ρ *all* evolve with p.

Since (3.2) can be represented as a fraction in recognizable functions, we seek to understand its logarithm. Therefore, in the following sections we consider asymptotic expansions of the incomplete and regular gamma function.

3.1.1.1 Asymptotic Expansion of the Logarithm of the Gamma Function

The denominator in formula (3.2) is expressed via the gamma function of the large argument. An asymptotic expression for $\ln\Gamma(a)$ as $a \to \infty$ is given by formula 8.327.3 of [28]

$$\ln\Gamma(a) \sim \left(a - \frac{1}{2}\right)\ln a - a + \frac{1}{2}\ln(2\pi) + \frac{1}{12a} - \frac{1}{360z^3} + ... \tag{3.7}$$

valid for $a \to \infty, |\arg a| < \pi$.

3.1.1.2 Asymptotic Expansion of the Logarithm of the Gamma Function

The numerator in formula (3.2) is expressed via the incomplete gamma function $\gamma(a, z)$ where both arguments are large. This situation calls for different asymptotic expansions in comparison with the standard case where a is assumed to be fixed and the value of z is growing. Below, we follow the approach of Paris in [44] who studied asymptotic expansions of the incomplete gamma functions

$$\gamma(a, z) = \int_0^z t^{a-1}e^{-t}dt \quad \text{and } \Gamma(a, z) = \int_z^\infty t^{a-1}e^{-t}dt$$

as $a \to \infty$ and $z \to \infty$. These asymptotic expansions depend on the relation between z and a and the variable

$$\chi = (z-a)/\sqrt{z}.$$

In particular, adopting asymptotic expansions in Paris [44] to the case of real z and a, we obtain

$$\begin{aligned}
\Gamma(a,z) &= z^{a-1/2}e^{-z}\left[d_0(\chi) - z^{-1/2}d_3(\chi)\right. \\
&+ \left. O(|z|^{-1})\right], \quad z > a, \\
\gamma(a,z) &= z^{a-1/2}e^{-z}\left[d_0(-\chi) + z^{-1/2}d_3(-\chi)\right. \\
&+ \left. O(|z|^{-1})\right], \quad z < a,
\end{aligned}$$

if $|\chi| \to \infty$, and

$$\begin{aligned}
\Gamma(a,z) &= z^{a-1/2}e^{-z}\left[d_0(\chi)(1 - z^{-1/2}(\chi/2 + \chi^3/6)) + (1/3 + \chi^2/6)z^{-1/2}\right. \\
&+ \left. O(|z|^{-1})\right], \quad z > a, \\
\gamma(a,z) &= z^{a-1/2}e^{-z}\left[d_0(-\chi)(1 - z^{-1/2}(\chi/2 + \chi^3/6)) - (1/3 + \chi^2/6)z^{-1/2}\right. \\
&+ \left. O(|z|^{-1})\right], \quad z < a,
\end{aligned}$$

when $|\chi|$ is bounded. Here, $d_k(\chi)$ is the parabolic cylinder function (see, e.g. Sections 9.24-9.25 of [28]) which has the following asymptotic properties:

1. $d_k(\chi)$ is bounded when χ is bounded;

2. $d_k(\chi) = \chi^{-(k+1)}(1 + O(\chi^{-2}))$ if $\chi \to \infty$

The last property yields

$$d_0(\chi) = \chi^{-1}(1 + O(\chi^{-2})), \chi \to \infty. \tag{3.8}$$

Taking into account the expressions above, we can write

$$\begin{aligned} \ln \Gamma(a,z) &= (a-1/2)\ln z - z + \Delta_\Gamma(a,z), \quad z > a, \\ \ln \gamma(a,z) &= (a-1/2)\ln z - z + \Delta_\gamma(a,z), \quad z < a. \end{aligned}$$

Since $z \to \infty$ and $z^{-1/2} d_3(|\chi|) = o(d_0(|\chi|))$, no matter whether χ is bounded or $\chi \to \infty$, it is easy to show that

$$\Delta_\Gamma(a,z) = O(\ln d_0(|\chi|)), \quad \Delta_\gamma(a,z) = O(\ln d_0(|\chi|)). \tag{3.9}$$

Combination of (3.8) and (3.9) imply that

$$\Delta_\Gamma(a,z) \sim \Delta_\gamma(a,z) \sim -\ln(|\chi|) = -\ln(|z-a|) + 1/2 \ln z.$$

Therefore, as $z \to \infty$ and $a \to \infty$, one has

$$\ln \Gamma(a,z) = a \ln z - z - \ln(|z-a|) + O(1), \quad z > a, \tag{3.10}$$

$$\ln \gamma(a,z) = a \ln z - z - \ln(|z-a|) + O(1), \quad z < a. \tag{3.11}$$

Combining equations (3.10) and (3.11) with the asymptotic expansion (3.7) of $\ln \Gamma(a)$ as $a \to \infty$, we derive

$$\begin{aligned} \ln[\Gamma(a,z)/\Gamma(a)] &= (a-1/2)\ln(z/a) - (z-a) + 1/2 \ln z - \ln(|z-a|) \\ &+ O(1), \ z > a, \end{aligned} \tag{3.12}$$

$$\begin{aligned} \ln[\gamma(a,z)/\Gamma(a)] &= (a-1/2)\ln(z/a) - (z-a) + 1/2 \ln z - \ln(|z-a|) \\ &+ O(1), \ z < a. \end{aligned} \tag{3.13}$$

3.1.2 The Lambert W Function

Since it will be useful in what follows, we introduce here the Lambert W Function, which is most often defined as the principal branch ($W(x) \geq -1$) of the solutions to $x = W(x)e^{W(x)}$. Since this problem is ubiquitous, the solution has arisen in a variety of contexts; it is only recently that a notation has been settled; see for example [16]. For our purposes, it is enough to know that

1. $W(x)$ is defined for all $x \geq -\frac{1}{e}$
2. $W(x)$ increases monotonically in its domain
3. $W(x) \sim \ln(x) - \ln\ln(x)$ as $x \to \infty$

The first property can be understood by inspecting the graph of $f(x) = xe^x$ which attains its global minimum $-\frac{1}{e}$ at $x = -1$. To establish the second property, note that for arbitrary x, we have

$$(f^{-1})'(x) = \frac{1}{f'(f^{-1}(x))}.$$

Since $f'(x) = (1 + x)e^x$, the derivative of the inverse of f (which is $W(x)$) is nonnegative wherever $f^{-1}(x) \geq -1$, i.e. on the branch $W(x) \geq -1$. For the third property, see [16]. Here, we use the approximation $W(x) \sim \ln(x)$ not for actual calculations, but to demonstrate how some of the following cases relate to each other.

Since it will arise several times in the following discussion, we seek to solve the following inequality.

Proposition 3 *Let $\alpha > 0, \beta$ be arbitrary. Then the solution to*

$$\alpha x + \ln x \geq \beta$$

is given by

$$x \geq \alpha^{-1} W(e^{\beta}\alpha)$$

where W is the Lambert W function.

Proof We note that the exponential function is monotonically increasing. Therefore,

$$e^{\alpha x + \ln x} \geq e^{\beta}$$

which implies that

$$\begin{aligned} xe^{\alpha x} &\geq e^{\beta} \\ \implies \alpha x e^{\alpha x} &\geq \alpha e^{\beta}. \end{aligned}$$

Hence, since W is also monotonically increasing

$$\begin{aligned} \alpha x &\geq W(\alpha e^{\beta}) \\ \implies x &\geq \alpha^{-1} W(\alpha e^{\beta}) \end{aligned}$$

where in this last step we note the necessity of the requirement that α be positive. □

3.1.3 Final Separability Requirements

For convenience, in advance we define three terms and two functions that will be useful. Denote

$$\begin{aligned} \nu &= p_1/p && (3.14) \\ \tau &= \frac{L-1}{\lambda} \\ \tau_\rho &= \frac{1+\rho^2}{\tau} = \frac{(1+\rho^2)\lambda}{L-1} && (3.15) \\ F_1(\tau, L, p, \nu) &= (L-1)(\tau^{-1} + \ln\tau - 1) + \ln(L-1) + 2\ln(1-\tau^{-1}) - 2\ln p_1 \\ F_2(\tau, \rho, L, p, \nu) &= (L-1)(\tau_\rho - \ln\tau_p - 1) + \ln(L-1) + 2\ln(\tau_\rho - 1) - 2\ln p_0. \end{aligned}$$

Where appropriate, we will still write p_0, p_1 but it should be understood that $p_1 = \nu p$, $p_0 = (1-\nu)p$.

Now, since the error term in (3.12) is bounded, we can ignore it in the limit:

$$\begin{aligned} \lim_{p\to\infty} \quad & p_1 \frac{\gamma(\alpha, z)}{\Gamma(\alpha)} = 0 \\ \lim_{p\to\infty} \quad & \ln p_1 + a\ln z - z - \ln(a-z) - (a - \frac{1}{2})\ln a + a = -\infty. \end{aligned}$$

We multiply the limit by (-2) and substitute in our own variables:

$$\lim_{p\to\infty} \lambda - (L-1) + (L-2)\ln(L-1) + 2\ln(L-1-\lambda) - (L-1)\ln\lambda - 2\ln p_1 = \infty$$

and substitute in τ:

$$\lim_{p\to\infty} (L-1)(\tau^{-1} + \ln\tau - 1) + \ln(L-1) + 2\ln(1-\tau^{-1}) - 2\ln p_1 = \infty.$$

The requirement becomes

$$\lim_{p\to\infty} F_1(\tau, L, p, \nu) = \infty. \tag{3.16}$$

A similar calculation on (3.5) yields the condition

$$\begin{aligned} &\lim_{p\to\infty} \quad (L-1)(\tau_\rho - \ln\tau_p - 1) + \ln(L-1) + 2\ln(\tau_\rho - 1) - 2\ln p_0 = \infty \\ &\lim_{p\to\infty} \quad F_2(\tau, \rho, L, p, \nu) = \infty. \end{aligned} \tag{3.17}$$

Hence (3.16) and (3.17) are precisely equivalent to (3.5) and (3.6); i.e. these are not approximations of the conditions, even if we obtained them through approximations.

3.2 Finding Separation Constants

To clarify a potentially confusing aspect of the above calculations, we note first that p, p_0, p_1, ρ, and L are given by the data set itself. We note, then, that selection of any one of λ, τ, τ_ρ fixes all the others. It is easiest to accomplish the following calculations in τ. Secondly, we recall that we are seeking a *sequence* of separating λ (though again the calculations will be carried out in τ); we suppress the subscripts here for convenience.

For separability we require both (3.16) and (3.17) to tend to ∞. We define

$$\tau^* = \arg\max_{\tau} \min(F_1(\tau, L, p, \nu), F_2(\tau, \rho, L, p, \nu))$$

i.e. we seek the smallest of the two functions and maximize it; if we can find the conditions under which this function goes to ∞, then the other will as well. We begin by defining $\hat{\tau}$ as

the solution to

$$F_1(\tau, L, p, \nu) - F_2(\tau, \rho, L, p, \nu) = 0$$

which simplifies to

$$(L-1)(\ln(1+\rho^2) - \rho^2\tau^{-1}) + 2\ln\left(\tfrac{\tau-1}{(1+\rho^2)-\tau}\right) + 2\ln\left(\tfrac{1-\nu}{\nu}\right) = 0. \tag{3.18}$$

We note immediately that $1 < \hat{\tau} < 1+\rho^2$ necessarily; any $\hat{\tau}$ that satisfies this requirement will be called *permissible*. The equation does not admit a recognizable solution as it is written. We keep in mind that permissibility is not a necessary condition; there could be τ outside this range for which the data are separable. Here with permissibility we begin seeking sufficient conditions.

To help in the interpretation of what follows, we borrow from signal processing a useful piece of vocabulary. The *signal-to-noise ratio* (SNR) is a relative measurement of the strength of the signal (here, the between-class variation) to the noise (here, the error term, the within-class variation). Here, then, we can refer to ρ^2 as the SNR. We consider two cases: the SNR is small (ρ is fixed) or moderate ($\rho \to \infty$).

We also need the idea of sparsity. We recall the definition (3.15). We say that the data are *sparse* when ν is small. This is shorthand - we really mean that informative columns are relatively few among the uninformative columns. If the data are not sparse, we say that they are *dense*. There are therefore four cases we would like to consider:

1. SNR small, dense data (ρ, ν fixed)

2. SNR moderate, dense data ($\rho \to \infty$, ν fixed)

3. SNR small, sparse data (ρ fixed, $\nu \to 0$)

4. SNR moderate, sparse data ($\rho \to \infty$, $\nu \to 0$)

First, however, it seems intuitive that larger ρ makes separation "easier." The following proposition makes this explicit.

Proposition 4 *Let data be generated according to CONFESS in which ρ_1 is fixed. Let $\{\lambda_p\}$ be a sequence of separating constants; i.e.*

$$\lim_{p\to\infty} P(U_p \leq \lambda_p) = 1$$

$$\lim_{p\to\infty} P(V_p \leq \lambda_p) = 0.$$

Then consider the same data set but with $\rho_2 > \rho_1$. Then this exact $\{\lambda_p\}$ separates this new set as well.

Proof We note that one of the conditions for separability (3.6) remains unchanged, as it does not rely on ρ. However, the other, (3.5), does:

$$\lim_{p\to\infty} p_0(1 - F_{L-1}((1+\rho_1^2)\lambda)) = 0.$$

Necessarily,

$$1 - F_{L-1}((1+\rho_2^2)\lambda)) \leq 1 - F_{L-1}((1+\rho_1^2)\lambda))$$

since λ is nonnegative, as a separating sequence, and F_{L-1} is monotonically increasing, as a cdf. However, this means that

$$\begin{aligned} 0 &= \lim_{p\to\infty} p_0(1 - F_{L-1}((1+\rho_1^2)\lambda)) \\ &> \lim_{p\to\infty} p_0(1 - F_{L-1}((1+\rho_2^2)\lambda)). \end{aligned}$$

Hence,

$$\lim_{p\to\infty} p_0(1 - F_{L-1}((1+\rho_2^2)\lambda)) = 0$$

and this old sequence $\{\lambda_p\}$ separates the new data set. □

If we are only seeking separability conditions, cases 2 and 4 are not strictly necessary, as we can extract results from cases 1 and 3. However, using cases 2 and 4, we can obtain better estimates of the constants when ρ^2 is large in *finite* data sets.

SNR small, dense data

Here, we seek a $\hat{\tau}$ that satisfies (3.18) when ρ, ν are fixed constants. In this case, the first term of (3.18) dominates all other terms as they are logarithmic. We therefore approximate (3.18) by discarding all but the first term and solving

$$(L-1)(\ln(1+\rho^2) - \rho^2\tau^{-1}) \approx 0$$

which implies that

$$\hat{\tau} = \hat{\tau}_1 = \frac{\rho^2}{\ln(1+\rho^2)} \tag{3.19}$$

i.e. we have $\hat{\tau} = \hat{\tau}_1$ in this first case. We know that for all $x > -1$

$$\frac{x}{1+x} \leq \ln(1+x) \leq x$$

so that

$$\frac{1}{1+\rho^2} \leq \frac{\ln(1+\rho^2)}{\rho^2} \leq 1$$

since ρ^2 is nonnegative. This implies, however, that $1 \leq \hat{\tau}_1 \leq 1+\rho^2$; i.e. $\hat{\tau}_1$ is permissible, regardless of the underlying parameters.

Since the SNR is small and the data dense in this case, it is reasonable to assert that ρ^2, ν are simply constant. This is not possible when either of these assumptions is violated; e.g. calculations with $\nu = 0$ are not possible, even if ν does converge to 0 as $p \to \infty$. Since ρ^2 is fixed, $\hat{\tau}_1$ is fixed. We note then that we can view

$$F_1(\hat{\tau}_1, L, p, \nu) \;\; = \;\; (L-1)(\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1) + \ln(L-1) + 2\ln(1-\hat{\tau}_1^{-1}) - 2\ln p_1$$

as a function in L and p.

If $\hat{\tau}_1 = 1$, then $\lambda = L-1$, but comparison to (3.1) shows that this is not a separating λ; i.e. the mean of $(Y_i | x_i = 1)$ is less than the mean of any individual $\chi^2_{L-1} = L-1$, so if $\lambda = L-1$, it is *greater* than the mean of $(Y_i | x_i = 1)$ and cannot possibly separate the two hypotheses. Hence, outright we reject $\hat{\tau}_1 = 1$. Regardless, this $\hat{\tau}_1$ is not even permissible.

We would like to apply Proposition 3 here. We claim that as long as $\hat{\tau}_1 \neq 1$, the coefficient of $(L-1)$ in $F_1(\hat{\tau}_1, L, p, \nu)$, namely

$$\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1$$

is strictly positive. To show this, we first note that its first derivative is

$$\ln \hat{\tau}_1 = \begin{cases} < 0 & \text{if } \hat{\tau}_1 > 0 \\ > 0 & \text{if } \hat{\tau}_1 < 0 \end{cases}$$

i.e. the coefficient is decreasing monotonically in $\hat{\tau}_1$ until it reaches its minimum at $\hat{\tau}_1 = 1$, then increasing monotonically in $\hat{\tau}_1$. The unique global minimum of this coefficient is achieved then at $\hat{\tau}_1 = 1$.

Since we have chosen a $\hat{\tau}$, everything in the model is now fixed. Is there separation in the model? Here, we seek a sufficient condition. Let $c \in [0, 1)$ be arbitrary. If it is eventually true for this c that

$$(L-1)(\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1) + c \ln(L-1) \geq 2 \ln p \tag{3.20}$$

then we have separation; i.e. if as $p \to \infty$ this inequality is violated only a finite number of times, then the data are separable. We demonstrate this as follows. Assume that such a c exists; i.e. we assert that it is eventually true that

$$(L-1)(\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1) + c \ln(L-1) \geq 2 \ln p.$$

Then

$$(L-1)(\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1) + (\ln(L-1) + (c-1)\ln(L-1))$$

$$+2 \ln(1 - \hat{\tau}_1^{-1}) - 2 \ln p_1 \geq 2 \ln(1 - \hat{\tau}_1^{-1}) + 2 \ln p - 2 \ln p_1.$$

However, this implies that

$$F_1(\hat{\tau}_1, L, p, \nu) \geq 2 \ln(1 - \hat{\tau}_1^{-1}) + 2 \ln \nu + (1-c) \ln(L-1).$$

Since $\hat{\tau}_1 \neq 1$ and ν are fixed, the right hand side of the inequality will tend to ∞ if $L \to \infty$.

We apply Proposition 3 to (3.20) and obtain the sufficient condition

$$L - 1 \geq \frac{c}{(\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1)} W\left(\exp\left(\frac{2 \ln p}{\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1} + \ln \frac{\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1}{c}\right)\right)$$

i.e. $L \to \infty$ by hypothesis. We can find a weaker but more comprehensible sufficient condition by setting $c = 0$, in which case we obtain

$$L - 1 \geq \frac{2 \ln p}{\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1} \tag{3.21}$$

i.e. $L - 1$ grows according to $\ln p$ if we are to satisfy the requirements of separability in this case. If L satisfies this inequality, then we have separability.

SNR moderate, dense data

Here, we seek a $\hat{\tau}$ that satisfies (3.18) when ν is a fixed constant but $\rho \to \infty$. Here we note that the second term of (3.18) can be written

$$\begin{aligned} 2 \ln\left(\frac{\tau - 1}{(1 + \rho^2) - \tau}\right) &= -2 \ln\left(\frac{\rho^2}{\tau - 1} - 1\right) \\ &\approx -2 \ln\left(\frac{\rho^2}{\tau - 1}\right) \\ &\approx -2 \ln\left(\frac{1 + \rho^2}{\tau}\right) \end{aligned}$$

and we seek to solve

$$(L - 1)(\ln(1 + \rho^2) - \rho^2 \tau^{-1}) - 2 \ln\left(\frac{1 + \rho^2}{\tau}\right) = 0.$$

Rearranging, we obtain

$$(L-1)\rho^2\tau^{-1} + 2\ln\tau^{-1} = (L-3)\ln(1+\rho^2).$$

By a modification of Proposition 3, replacing the inequality with an equality when necessary, we calculate

$$\hat{\tau}_2 = \frac{(L-1)\rho^2}{2}\frac{1}{W\left(\exp\left(\frac{1}{2}(L-3)\ln(1+\rho^2)\right)\right)}. \tag{3.22}$$

It is not immediately clear how this relates to the $\hat{\tau}_1$ in the case above. If we take the first term of the asymptotic expansion of $W(\cdot)$ from [16], i.e. $W(x) \approx \ln x$, then

$$\hat{\tau}_2 = \frac{L-1}{L-3}\frac{\rho^2}{\ln(1+\rho^2)} \approx \frac{\rho^2}{\ln(1+\rho^2)} = \hat{\tau}_1$$

i.e. it is comparable to the case above. We do not claim that this *should* be done in any sort of application if either W exactly or a more appropriate approximation is available; this is merely a demonstration that allowing $\rho^2 \to \infty$ does not change its effect on $\hat{\tau}$ considerably. However, we do note that when L is large, $\hat{\tau} = \hat{\tau}_1$ is probably acceptable for most applications and Proposition 4 indicates this choice is sufficient.

SNR small, sparse data

Here, we seek a $\hat{\tau}$ that satisfies (3.18) when ρ is a fixed constant but $\nu \to 0$. In this case, the first term and third terms of (3.18) dominate. We also note that $\ln(1-\nu) \approx 0$, so we

can solve directly:

$$\hat{\tau}_3 = \frac{\rho^2}{\ln(1+\rho^2)} \frac{1}{1 - \frac{2\ln\nu}{(L-1)\ln(1+\rho^2)}} = \frac{\hat{\tau}_1}{1 - \frac{2\ln\nu}{(L-1)\ln(1+\rho^2)}}. \tag{3.23}$$

While we note the connection with $\hat{\tau}_1$, we note that $-\ln\nu$ could be very large; i.e. $\hat{\tau}_3 = \hat{\tau}_1$ is not acceptable. For the permissibility of $\hat{\tau}_3$ we should have

$$1 \leq \quad \hat{\tau}_3 \quad \leq 1 + \rho^2.$$

We take reciprocals and distribute:

$$\begin{aligned}
\frac{1}{1+\rho^2} &\leq & \hat{\tau}_3^{-1} & & &\leq 1 \\
\frac{1}{1+\rho^2} &\leq & \tfrac{\ln(1+\rho^2)}{\rho^2}\left(1 - \tfrac{2\ln\nu}{(L-1)\ln(1+\rho^2)}\right) & &&\leq 1 \\
\frac{1}{1+\rho^2} &\leq & \tfrac{\ln(1+\rho^2)}{\rho^2} - \tfrac{2\ln\nu}{\rho^2(L-1)} & &&\leq 1.
\end{aligned}$$

We investigate the upper inequality; we temporarily ignore the lower bound. Multiplying by ρ^2 and rearranging implies we should have

$$0 \leq \frac{-2\ln\nu}{L-1} < \rho^2 - \ln(1+\rho^2)$$

where the lower bound here comes from the fact that L and $-2\ln\nu$ are nonnegative. We recall that in this case, ρ^2 is bounded, which implies that L should grow at a particular rate:

$$L - 1 > \frac{-2\ln\nu}{\rho^2 - \ln(1+\rho^2)} \tag{3.24}$$

We recall the requirement for permissibility (3.24) in this case. For these calculations to even be possible, L must be at least proportional (for a specific constant) to $\ln\nu$. Assume that

$$\frac{-\ln\nu}{L-1} \to k$$

for some constant k. Then from (3.23) we note that

$$\begin{aligned}\hat{\tau}_3 &= \frac{\hat{\tau}_1}{1-\frac{2\ln\nu}{(L-1)\ln(1+\rho^2)}} \\ &\to \frac{\hat{\tau}_1}{1-\frac{2k}{\ln(1+\rho^2)}} \\ &= k'\hat{\tau}_1\end{aligned}$$

for some nonzero constant k' and we recover the calculations for the first case. Hence, it is sufficient that L grow as $\ln\nu$ and $\ln p$, for some known constants of proportionality.

SNR moderate, sparse data

Here, we seek a $\hat{\tau}$ that satisfies (3.18) when $\rho \to \infty$ and $\nu \to 0$. In this case, we can discard no terms of (3.18). The solution is essentially no different than $\hat{\tau}_2$ and we obtain

$$\hat{\tau}_4 = \frac{(L-1)\rho^2}{2}\frac{1}{W\left(\nu^{-1}\exp\left(\frac{1}{2}(L-3)\ln(1+\rho^2)\right)\right)}.$$

This is a useful estimate computationally, but it makes finding separability conditions difficult. However, by Proposition 4, we do not need to; the previous case suffices.

3.3 Summary of Cases

Proposition 5 *Let data be generated according to CONFESS. For every combination of τ, ρ, p, ν there exist constants $c > 0, d \leq 0$ for which the data are separable if*

$$L - 1 \geq c\ln p$$

$$L - 1 \geq d \ln \nu$$

except for a finite number of p.

We have what we need above. In the first case, when ρ, ν are nonasymptotic, $d = 0$ (i.e. there is no growth in $\ln \nu$ since ν is fixed) and the constant c is given by (3.21):

$$c = \frac{2}{\hat{\tau}_1^{-1} + \ln \hat{\tau}_1 - 1}$$

where we recall that $\hat{\tau}_1$ is defined in (3.19) as a function of ρ and is fixed in this case. In the third case, when ρ is fixed and $\nu \to 0$ we appeal to (3.24) and note that

$$d = \frac{-2}{\rho^2 - \ln(1 + \rho^2)}.$$

Here, we invoke Proposition 4 to show that, in fact, these expressions suffice in cases 2 and 4. In other words, if $\rho \to \infty$, it suffices to fix a ρ_0 and calculate and fix the c, d above. As long as L grows accordingly, the data are separable. It can occur, however, that L may grow more slowly and the data still be separable; we recall that we are seeking sufficient conditions.

Simulations

We can also answer the question of separability through simulation. For each value of ρ^2 we generated 1000 data sets according to CONFESS with 10 groups, 10 informative and 10 uninformative vectors; i.e. $\nu = \frac{1}{2}$. We used

$$\hat{\tau}_1 = \frac{\rho^2}{\ln(1 + \rho^2)}$$

as the separation constant. The results are visualized in Figure 3.1.

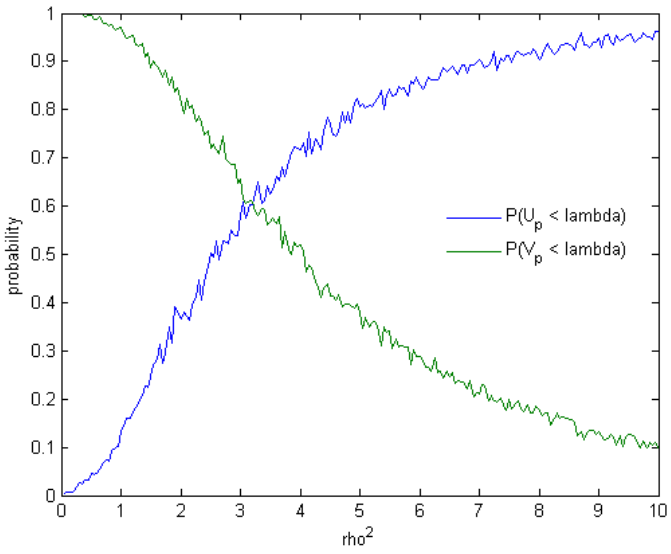

Figure 3.1: Example of separation probabilities in a finite data set, function of ρ^2

Even though the situation is essentially symmetric - there are equal numbers of informative and uninformative vectors, the probability of retaining an uninformative vector is nearly twice the probability of accidentally discarding an informative vector when $\rho^2 = 10$. When $\rho^2 = 5$, we can expect to retain about 8 informative vectors and 4 uninformative vectors; i.e. after dimension reduction we have an effective $\nu \approx \frac{2}{3}$.

For each value of L between 2 and 50, we generated 1000 data sets according to CONFESS with 10 groups, 10 informative and 10 uninformative vectors; i.e. $\nu = \frac{1}{2}$, and $\rho^2 = 5$. We used the same separation constant. The results are visualized in Figure 3.2.

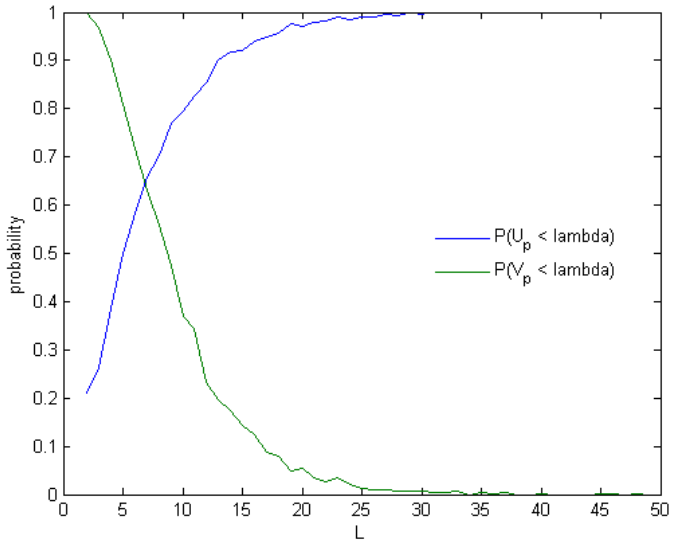

Figure 3.2: Example of separation probabilities in a finite data set, function of L

Of course, as the number of groups increases, the difficulty of classification might increase, regardless of the improved model selection. Below, we develop the framework in which this question can be considered.

CHAPTER 4
MISCLASSIFICATION RATE OF CONFESS

Assume we have data generated according to CONFESS in which all parameters but x_i are known. We have a new observation o which actually resides in the first class. Specifically, let c be an arbitrary column number. We have

$$\begin{aligned} o_c &= n^{-\frac{1}{2}} m_c + u_o + v_o + \epsilon_o \\ &= [\text{constant term}] + [\text{between-class variation}] + [\text{within-class variation}] + [\text{gaussian noise}] \\ d_c^i &= n^{-\frac{1}{2}} m_c e + u_c \\ &= [\text{constant term}] + [\text{between-class variation}] \end{aligned}$$

where the subscripts on o_c indicate that these are the components of the new sample. Compare this for instance to (2.5); in this best-case scenario, we are able to observe class means without any within-class variation. This can be compared to the situation in which the number of observations per class is large.

Here we define the simple classifier

$$\hat{l} = \arg \min_{i=1,\ldots,L} T_i$$

where

$$T_i = ||(\sqrt{\Sigma})^{-1}(o - d^i)||^2$$

and d^i is the actual (row-wise) mean of class i. This is in fact the Mahalanobis distance; see for example [39]. T_1 measures the distance of the new sample from the mean of class 1 and

T_2 measures the distances from the mean of class 2. We classify the sample into whichever class it is closest to.

We recall that the sample to be classified is in the first class, so we can calculate the probability of a misclassification; i.e.

$$P(\text{misclassification}) = P(T_1 > \min\{T_i\}_{i=2}^{L}). \tag{4.1}$$

Let A_i be the event $T_1 > T_i$. Then we can rewrite (4.1) as

$$P(\text{misclassification}) = P(\cap_{i=2}^{L} A_i) \leq \sum_{i=2}^{L} P(A_i)$$

by the subadditivity of the measure. Since all of the classwise contributions are distributed identically in (2.5), the distribution of T_2 is the same as any T_i for $i \neq 1$. Hence, we can rewrite the previous inequality as

$$P(\text{misclassification}) \leq \sum_{i=2}^{L} P(A_2) = (L-1)P(A_2) = (L-1)P(T_2 - T_1 < 0). \tag{4.2}$$

We therefore define the random variable $E = T_2 - T_1$. We note that if $E < 0$, then we have misclassified, since the distance from class 2 is smaller than the distance from the class 1 mean. Hence we need to calculate $F_E(0)$, which is the probability of a misclassification into class 2.

Lemma 3 *The random variable $E = T_2 - T_1$ has characteristic function*

$$\Phi_E(t) = \pi^p \gamma_1^{p_1} \gamma_0^{p_0} [(t - i\alpha_1)^2 + \beta_1^2]^{-\frac{1}{2}p_1} [t^2 + \beta_0^2]^{-\frac{1}{2}p_0} \tag{4.3}$$

where

$$\begin{aligned}
\gamma_1 &= \frac{1}{2\pi\sqrt{4\rho^2+3}} \\
\alpha_1 &= \frac{\rho^2}{2(4\rho^2+3)} \\
\beta_1 &= \frac{\sqrt{(\rho^2+1)(\rho^2+3)}}{2(4\rho^2+3)} \\
\gamma_0 &= \frac{1}{2\pi\sqrt{3}} \\
\beta_0 &= \frac{1}{2\sqrt{3}}.
\end{aligned}$$

Proof We accomplish this in several steps. First, we write E as a sum of E_c over the columns of data. Next, we represent each E_c in a way that facilitates the calculation of its characteristic function; specifically, we consider the product of two correlated normals rather than the difference of two correlated random variables. Finally, we combine the characteristic functions of E_c to obtain the characteristic function of E.

Without loss of generality, we can write $\Sigma = \mathbf{I}$, in which case $(\sqrt{\Sigma})^{-1} = \mathbf{I}$ and

$$\begin{aligned}
E &= T_2 - T_1 = ||(o-d^2)||^2 - ||(o-d^1)||^2 \\
&= \sum_{c=1}^{p} |o_c - d_c^2|^2 - |o_c - d_c^1|^2 \\
&= \sum_{c=1}^{p} E_c.
\end{aligned}$$

We can consider each component E_c of the sum distance separately, since each data column is independent of the rest. Let c be some arbitrary column number. Without loss of generality, assume that $x_c = 1$; we can recover the case $x_c = 0$ by setting $\rho^2 = 0$.

We note that $o_c - d_c^1$ and $o_c - d_c^2$ are correlated as they both include o_c; we cannot simply find the distribution of each and subtract. This is the essential difficulty in this calculation. We note however that

$$\begin{aligned} o_c - d_c^1 &= v_c + \epsilon_{c1} \\ o_c - d_c^2 &= (u_{c1} - u_{c2}) + v_{c0} + \epsilon_{c2} \end{aligned}$$

i.e. the distance from class 2 of the new sample is increased by the between-class variation $u_{c1} - u_{c2}$; compare to (2.5). The larger this between-class variation, the larger is E_c, which increases E, which makes misclassification less likely. Now, we have

$$\epsilon_{c1} \sim N(0,1)$$

$$\epsilon_{c2} \sim N(0,1)$$

$$v_{c0} \sim N(0,1)$$

$$u_{c1} \sim N(0,\rho^2)$$

$$u_{c2} \sim N(0,\rho^2)$$

where all of the above random variables are by hypothesis independent. It is worth noting here that if we had not assumed that $\mathbf{\Sigma}$ had a particular structure, it is at this point that all the calculations would become the same. By normalizing by $\mathbf{\Sigma}$ or by assuming that $\mathbf{\Sigma} = \mathbf{I}$, we obtain these same distributions.

We then note that $o_c - d_c^1$ and $o_c - d_c^2$, as the linear combination of normals, are normal. Hence, stacked together as a vector they are a 2-normal. Specifically, we can write

$$\begin{bmatrix} o_c - d_c^1 \\ o_c - d_c^2 \end{bmatrix} \sim N\left(0, \begin{bmatrix} 2 & 1 \\ 1 & 2(1+\rho^2) \end{bmatrix}\right)$$

by calculating the covariances of each term individually. Call this covariance matrix M.

We note that M admits a Cholesky decomposition $M = UU^\top$ where

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 0 \\ 1 & \sqrt{4\rho^2+3} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 0 \\ 1 & \hat{\rho} \end{bmatrix}$$

where we have defined $\hat{\rho} = \sqrt{4\rho^2+3}$ for convenience. This implies that

$$\begin{bmatrix} o_c - d_c^1 \\ o_c - d_c^2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 0 \\ 1 & \hat{\rho} \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}$$

where N_1, N_2 are iid $N(0,1)$.

It might not be immediately clear what we have accomplished. We have written the two components of this multivariate normal each as a linear combination of two independent standard normals. Then we can write

$$\begin{aligned} E_c &= |o_c - d_c^2|^2 - |o_c - d_c^1|^2 \\ &= \frac{1}{2}(N_1 + \hat{\rho}N_2)^2 - \frac{1}{2}(2N_1)^2 \\ &= \frac{1}{2}[3N_1 + \hat{\rho}N_2][-N_1 + \hat{\rho}N_2] \end{aligned}$$

where we have factored the previous line as a difference of squares. There is no unique factorization; e.g. where do we place the $\frac{1}{2}$? In fact, any choice suffices.

In what follows, we determine the joint distribution of these factors. Next, we find the distribution of their product. Specifically, we define random variables R, S as the two factors

$$\begin{bmatrix} R \\ S \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & \frac{\hat{\rho}}{2} \\ -1 & \hat{\rho} \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}.$$

This transformation has inverse

$$\begin{bmatrix} N_1 \\ N_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{-1}{4} \\ \frac{1}{2\hat{\rho}} & \frac{3}{4\hat{\rho}} \end{bmatrix} \begin{bmatrix} R \\ S \end{bmatrix}.$$

The determinant of this inverse transformation matrix is $\frac{1}{2\hat{\rho}}$; i.e. the Jacobian of this transformation is

$$|J| = \left| \frac{1}{2\hat{\rho}} \right| = \frac{1}{2\hat{\rho}}$$

Then for arbitrary r, s we have the joint distribution

$$\begin{aligned} F_{R,S}(r,s) &= F_{N_1}\left(\frac{1}{2}r - \frac{1}{4}s\right) F_{N_2}\left(\frac{1}{2\hat{\rho}}r + \frac{3}{4\hat{\rho}}s\right) |J| \\ &= \frac{1}{4\pi\hat{\rho}} \exp\left(-\frac{1}{2}\left(\left(\frac{1}{2}r - \frac{1}{4}s\right)^2 + \left(\frac{1}{2\hat{\rho}}r + \frac{3}{4\hat{\rho}}\right)^2\right)\right). \end{aligned} \tag{4.4}$$

Next, we define

$$\begin{bmatrix} E_c \\ S \end{bmatrix} = \begin{bmatrix} RS \\ S \end{bmatrix}$$

where S is to be discarded. This transformation has inverse

$$\begin{bmatrix} R \\ S \end{bmatrix} = \begin{bmatrix} \frac{E_c}{S} \\ S \end{bmatrix}.$$

This inverse has Jacobian

$$|J| = \frac{1}{|S|}$$

Hence, for arbitrary z, w we have

$$\begin{aligned} F_{E_c,Z}(z,s) &= F_{R,S}\left(\frac{z}{s}, s\right)|J| \\ &= \frac{1}{4\pi\hat{\rho}}\exp\left(\frac{4z(\hat{\rho}^2-3)}{32\hat{\rho}^2}\right) \times \frac{1}{|s|}\exp\left(-\frac{1}{32\hat{\rho}^2}\left[\frac{4z^2}{s^2}(\hat{\rho}^2+1)+s^2(\hat{\rho}^2+9)\right]\right) \end{aligned} \tag{4.5}$$

where we have factored the pdf into a product, the first part of which does not depend on s. We then calculate the pdf of E_c by integrating out S:

$$\begin{aligned} f_{E_c}(z) &= g(z)\int_{-\infty}^{\infty} h(z,s)dz \\ &= \frac{1}{2\pi\sqrt{4\rho^2+3}}\exp\left(\frac{\rho^2 z}{2(4\rho^2+3)}\right)K_0\left(\frac{|z|}{2(4\rho^2+3)}\sqrt{(\rho^2+1)(\rho^2+3)}\right) \end{aligned} \tag{4.6}$$

where K_0 is the modified Bessel function of the second kind with $\nu = 0$. The essential calculation is

$$\int_0^{\infty} x^{\nu-1}e^{-\frac{\beta}{x}-\gamma x}dx = 2\left(\frac{\beta}{\gamma}\right)^{\frac{\nu}{2}}K_\nu(2\sqrt{\beta\gamma}).$$

See for example [28], 3.471.9, p.368. We note that this also holds when $x_c = 0$, which we obtain from the above by asserting that $\rho^2 = 0$.

Now, since $E = \sum E_c$, we calculate the characteristic function of E_c. We have

$$\begin{aligned} \Phi_{E_c}(t) &= \int_{-\infty}^{\infty} e^{itz} f_{E_c}(z)dz \\ &= \int_{-\infty}^{\infty} e^{itz} \times \gamma e^{\alpha z}K_0(\beta|z|)dz \\ &= \gamma\int_{-\infty}^{\infty} e^{iz(t-i\alpha)}K_0(\beta|z|)dz \\ &= \frac{\pi\gamma}{\sqrt{(t-i\alpha)^2+\beta^2}} \end{aligned}$$

where we have made the obvious substitutions for clarity; see for example formula 17.34.9 in [28]. There are two cases, $x_c = 1, 0$, and we have the variables $\gamma_1, \alpha_1, \beta_1$ and $\gamma_0, \alpha_0, \beta_0$ for the two cases respectively:

$$\begin{aligned}
\gamma_1 &= \frac{1}{2\pi\sqrt{4\rho^2+3}} \\
\alpha_1 &= \frac{\rho^2}{2(4\rho^2+3)} \\
\beta_1 &= \frac{\sqrt{(\rho^2+1)(\rho^2+3)}}{2(4\rho^2+3)} \\
\gamma_0 &= \frac{1}{2\pi\sqrt{3}} \\
\alpha_0 &= 0 \\
\beta_0 &= \frac{1}{2\sqrt{3}}.
\end{aligned}$$

Now, we recall that E_c is the sum over the c columns of data. Since there are p_1 of these with $x_c = 1$ and p_0 of these with $x_c = 0$, we can calculate the characteristic function of E_c as the product of the respective characteristic functions:

$$\Phi_E(t) = \pi^p \gamma^{p_1} \gamma_0^{p_0} [(t - i\alpha_1)^2 + \beta_1^2]^{-\frac{1}{2}p_1} [t^2 + \beta_0^2]^{-\frac{1}{2}p_0}. \tag{4.7}$$

□

We note that the exponential term in (4.6) provides positive skew of the E_c when $x_c = 1$. As ρ^2 increases, T_2 becomes larger than T_1 and, as far as the contribution of this one E_c is concerned, misclassification becomes less likely. For example, we generated data with $\rho^2 = 5$ in the figure below. Whenever $\rho^2 = 0$, however, the distribution is symmetric around 0 and inclusion only increases the variance of E.

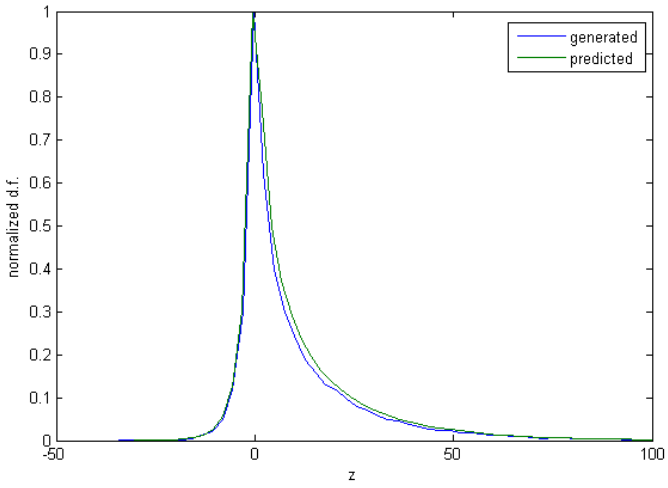

Figure 4.1: Comparison of generated E_c and calculated distribution function

Now that we have the characteristic function of the error of the classifier, presumably we can calculate the misclassification rate. Alternatively, given an acceptable rate of misclassification, call it $\aleph$, using (4.2), one can tolerate up to

$$L - 1 \geq \frac{\aleph}{F_E(0)}$$

groups before expecting to exceed a misclassification rate of $\aleph$ in repeated trials.

4.1 Numerical Inversion of the Characteristic Function

It is true that when p_0, p_1 are even, (4.7) is a rational function with only four singularities. In this case, we can in theory invert the characteristic function by the Cauchy residue theorem. However, this calculation is only tractable for small and definite p_0, p_1; the residues are

rational functions of all the above variables. Series methods and partial fractions suffer from the same limitation. Regardless, we do not need the cdf in general, only $F_E(0)$.

According to the theorem of Gil-Pelaez [26],

$$\begin{aligned} F_E(z) &= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{1}{t}\mathrm{Im}\left[e^{-itz}\Phi_E(t)\right]dt \\ F_E(0) &= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{1}{t}\mathrm{Im}\left[\Phi_E(t)\right]dt. \end{aligned} \tag{4.8}$$

A closed-form integral does not seem possible, especially since the real part of $\Phi_E(0)$ is nonzero; i.e. we cannot swap the order of the operator Im and the integral, as the integral of the real part diverges.

Numerical integration of (4.8), however, can give an idea of the behavior. We used the simple rectangle method with $\Delta_t = 10^{-4}$ to calculate various values of $F_E(0)$; i.e.

$$\hat{F}_E(0) = \frac{1}{2} - \frac{1}{\pi}\sum_{t=0}^{M}\frac{1}{t}\mathrm{Im}\left[\Phi_E(t\Delta_t)\right]$$

for some suitably large M, since the integrand decays quickly. We used $M\Delta_t = 20$, but we note that $\mathrm{Im}\Phi_E(t)$ changes sign frequently near the origin, especially as p_1 grows. More sophisticated methods will be necessary for actual application; see for example [58].

Regardless, here we found this sort of calculation sufficiently close to generated data for small p_1. For instance, when we generated data according to CONFESS with $p_0 = 2, p_1 = 3$ and various values of ρ^2, 2,000 E generated per ρ^2, we find the expected match in Figure 4.2.

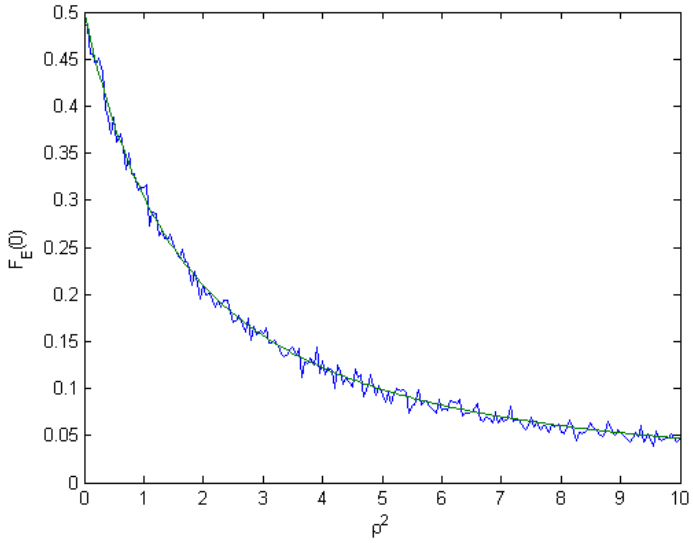

Figure 4.2: Comparison of generated and predicted $F_E(0)$ for various values of ρ^2

4.2 Single Limit Lemmas

Lemma 4 *If ρ, L, p_1 are held constant, then*

$$\lim_{p_0 \to \infty} F_E(0) = \frac{1}{2} \tag{4.9}$$

Proof Consider a data set generated according to CONFESS, with p_0 uninformative variables. For clarity, we rewrite (4.8) as

$$F_E(0) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty I(p_0) dt.$$

Then the claim is equivalent to

$$\lim_{p_0 \to \infty} \int_0^\infty I(p_0) dt = 0.$$

Let p_0 be arbitrary. We note that $\int_0^\infty I(p_0)dt < \infty$ necessarily. Then let m be some natural number and consider the data set with $p_0 + m$ uninformative variables. We calculate

$$\begin{aligned}\int_0^\infty I(p_0+m)dt &= \pi^m\gamma_0^m\int_0^\infty \frac{1}{(t^2+\beta_0^2)^{\frac{m}{2}}}I(p_0)dt \\ \left(\int_0^\infty I(p_0+m)dt\right)^2 &\leq \pi^{2m}\gamma_0^{2m}\int_0^\infty \frac{1}{(t^2+\beta_0^2)^m}dt\int_0^\infty (I(p_0))^2dt\end{aligned}$$

by the Cauchy-Schwarz inequality. Next, by formula 3.241.4 in [28] we can evaluate the first integral exactly:

$$\begin{aligned}&= \pi^{2m+\frac{1}{2}}\gamma_0^{2m}\beta_0^{1-2m}\frac{\Gamma\left(m-\frac{1}{2}\right)}{2\Gamma(m)}\int_0^\infty (I(p_0))^2dt \\ &= \sqrt{\pi}\beta_0\frac{\Gamma\left(m-\frac{1}{2}\right)}{2\Gamma(m)}\int_0^\infty (I(p_0))^2dt.\end{aligned}$$

We next note that

$$\int_0^\infty (I(p_0))^2dt < \infty$$

since $I(p_0)$ is bounded by a constant and eventually small; i.e. $(I(p_0))^2$ is dominated eventually by $I(p_0)$. Next,

$$\lim_{m\to\infty}\frac{\Gamma\left(m-\frac{1}{2}\right)}{\Gamma(m)} = 0$$

by formula 8.328.2 in [28]. Hence,

$$\lim_{m\to\infty}\int_0^\infty I(p_0+m)dt = 0$$

and we have the result. □

The interpretation is straightforward - addition of uninformative variables to a CONFESS model reduces the accuracy of the classifier, and eventually the classification is no better than guessing.

Lemma 5 *If ρ, L, p_0 are held constant and $\rho^2 > 0$,*

$$\lim_{p_1 \to \infty} F_E(0) = 0 \tag{4.10}$$

Proof Due to the difficulty in the Gil-Pelaez inversion, we turn to a less precise method. We note that

$$\Phi_E(t) = \pi^p \gamma^{p_1} \gamma_0^{p_0} [(t - i\alpha_1)^2 + \beta_1^2]^{-\frac{1}{2}p_1} [t^2 + \beta_0^2]^{-\frac{1}{2}p_0}$$

which implies that

$$\begin{aligned}
\mu_E &= \frac{1}{i}\frac{d \ln \Phi_E}{dt}(0) = \frac{\alpha_1 p_1}{(\beta_1^2 - \alpha_1^2)} \\
&= 2p_1\rho^2 \\
\sigma_E^2 &= -\frac{1}{2}\frac{d^2 \ln \Phi_E}{dt^2}(0) = \frac{p_1}{4}\left(\frac{4\alpha_1^2}{(\beta_1^2 - \alpha_1^2)^2} + \frac{2}{\beta_1^2 - \alpha_1^2}\right) + \frac{p_0}{2\beta_0^2} \\
&= p_1(8\rho^4 + 4\rho^2 + 3) + 12p_0.
\end{aligned}$$

Now, for large p_1, we note that we are adding a large number of iid random variables of finite variance; i.e. we can consider a central limit theorem approximation. Instead of formally approximating the actual probability, it suffices to calculate the distance of the mean from 0; i.e. we define

$$\begin{aligned}
z &= \frac{\mu_E - 0}{\sigma_E} \\
&= \frac{2p_1\rho^2}{\sqrt{p_1(8\rho^4 + 4\rho^2 + 3) + 12p_0}}.
\end{aligned} \tag{4.11}$$

The result follows whenever $z \to \infty$, e.g. by Chebyshev's inequality. Note that

$$\begin{aligned} P(E<0) &= P\left(\frac{E-\mu_E}{\sigma_E} < -\frac{\mu_E}{\sigma_E}\right) \\ &\leq P\left(\left|\frac{E-\mu_E}{\sigma_E}\right| > \left|\frac{\mu_E}{\sigma_E}\right|\right) \\ &\leq P\left(\left|\frac{E-\mu_E}{\sigma_E}\right| > z\right) \\ &\leq \frac{1}{z^2} \end{aligned}$$

which converges to 0 as $z \to \infty$, but this is the statement that $\rho^2 > 0$ and $p_1 \to \infty$. □

The interpretation is straightforward - addition of informative variables to a CONFESS model improves the accuracy of the classifier, until it is almost surely able to differentiate between two groups.

We do note that the CLT approximation is not completely inappropriate for calculation of probabilities. For instance, when we generate data according to CONFESS, with $\rho^2 = 2$, $p_0 = 5, p_1 = 20$, we obtain a fairly close match. Keeping in mind that we are mostly interested in distributions when either p_0 or p_1 is large, this gives a good theoretical tool even if certain applications need greater precision.

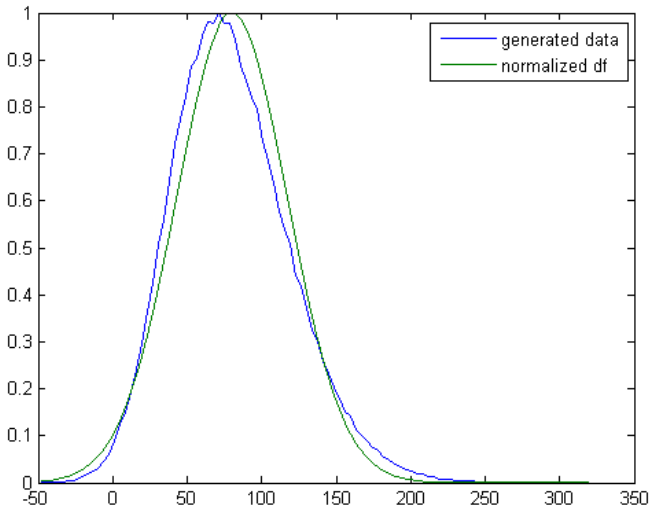

Figure 4.3: Comparison of generated data and CLT approximation

4.3 Discussion

In the present dissertation we considered a novel problem of model selection which is specifically designed for classification of high-dimensional data with a large number of classes. To the best of our knowledge, this problem has never been studied in depth previously and poses new challenges. The problem has been motivated by classification of animal communication signals, in particular, electric signals recorded from 21 groups of tropical South American electric knife fishes.

In the paper, we introduced two Bayesian models for feature selection in high dimensional data, specifically designed for the purpose of classification. We use two approaches to the problem: one which discards the components which have "almost constant" values (Model 1) and another which retains the components for which variations in-between the groups are larger than those within the groups (Model 2). We assume that $p >> n$, i.e. the number of components p is much larger than the number of samples n, and that only few of those p components are useful for subsequent classification. We showed that particular cases of the above two models recover familiar variance or ANOVA-based component selection. When one has only two classes and features are *a priori* independent, Model 2 reduces to to the Feature Annealed Independence Rule (FAIR) introduced by Fan and Fan (2008) and can be viewed as a natural generalization of FAIR to the case of $L > 2$ classes.

One of the nontrivial results of the dissertation is that precision of feature selection using Model 2 improves when the number of classes grows. In particular, it is known that when p

is large and the number of classes is small, e.g., $L = 2$, one needs the difference between the mean vectors of two classes to be large. We showed that when L is also large, separation is possible even if the differences between each of the two classes are relatively small.

Subsequently, we examined the rate of misclassification with and without feature selection on the basis of Model 2; we have only very rough asymptotic results, however. Our study of classification precision is not yet complete: while we have the characteristic function of the random variable associated with misclassification, we have not been able to invert it. Future work will involve investigation of the rate of misclassification when the number of classes grows.

LIST OF REFERENCES

[1] Abramovich, F. and Angelini, C. (2006) Bayesian Maximum a posterior Multiple Testing Procedure. *Sankyaa*, **68**: 436-460.

[2] Abramovich, F., Grinshtein, V., Pensky, M. (2007) On optimality of Bayesian testimation in the normal means problem. *Annals of Statistics*, **35**: 2261-2286.

[3] Ahn, J., Todd, M., and Marron, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, **102**: 1267-1272.

[4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**: 716-723.

[5] Arnegard, M., McIntyre, P., Harmon, L., Zelditch, M., Crampton, W., Davis, J., et al. Sexual Signal Evolution Outpaces Ecological Divergence during Electric Fish Species Radiation. *American Naturalist* **176**.3 (2010): 335-356.

[6] Auer, P., Burgsteiner, H., and Maass, W. (2008). A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Networks*, **21**: 786-795.

[7] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, **57**: 289-300.

[8] Berger, J. and Pericchi, R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of Statistical Planning and Inference*, **91**: 109-122.

[9] Bickel, P, 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.

[10] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**: 373-384.

[11] Bullock, T., Hopkins, C, Popper, A., Fay, R. (2005). Electroreception. Springer, New York.

[12] Candes, E. and Tao, T. (2007). The Dantzig Selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, **35** : 2313-2351.

[13] Chang, W., and Vidakovic, B. (2002). Wavelet estimation of a base-line signal from repeated noisy measurements by vertical block shrinkage. *Computer Statistical Data Analysis*, **40**: 317-328.

[14] Cheng, S. and Higham, N. (1998). A Modified Cholesky Algorithm Based on a Symmetric Indefinite Factorization. *SIAM Journal of Matrix Analysis Applications*, **19**: 1097-1110.

[15] Crampton, W.G.R., Davis, J.K., Lovejoy, N.R., Pensky, M. (2008) Multivariate classification of animal communication signals: a simulation-based comparison of alternative signal processing procedures, using electric fishes. *Journal of Physiology – Paris*, **102**: 304-321.

[16] Corless, R. M. et al. (1996). On the Lambert Function. *Adv. Comput. Math.*, **5**: 329-359.

[17] Daniels, M. and Kass, R. (1999). Nonconjugate Bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models. *Journal of the American Statistical Association*, **94**: 1254.

[18] Davis, J., Pensky, M., and Crampton, W. Bayesian feature selection for classification with possibly large number of classes. *Journal of Statistical Planning and Inference*, **141** (2011): 3256-3266.

[19] Demartines, P. and Herault, J. (1997). Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets. *IEEE Transactions on Neural Networks*, **8**: 148-154.

[20] Donoho, D.L., Johnstone, I.M. (1994) Ideal spatial adaptation via wavelet shrinkage. *Biometrica*, **81**: 425-455.

[21] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, **32**: 407-499

[22] Fan, J. and Fan, Y. High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, **6**: 2605-2637.

[23] Fisher, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**: 179-188.

[24] Furnival, G. and Wilson, R. (1974). Regression by leaps and bounds. *Technometrics*, **16**: 499-511.

[25] George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrica*, **84**: 731-747.

[26] Gil-Pelaez, J. Note on the inversion theorem. *Biometrica*, **38**: 481-382.

[27] Gohberg, I. and Goldberg, S. (1981). *Basic Operator Theory.* Birkhauser, Boston.

[28] Gradshteyn, I.S. and Ryzhik, I.M. (2007). Table of Integrals, Series, and Products. Seventh ed., Academic Press: Amsterdam.

[29] Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*: **82**: 711-732.

[30] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**: 97-109.

[31] Higham, N. (1986). Computing the polar decomposition - with applications. *SIAM Journal of Scientific Statistical Computing*, **7**: 1160-1174.

[32] Higham, N. (1988). Computing the nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, **103**: 103-118.

[33] Higham, N. (2002). Computing the nearest correlation matrix - a problem from finance. *IMA Journal of Numerical Analysis*, **22**: 329-343.

[34] Hinton, G. and Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**: 504-507.

[35] Hoerl, A. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, **58**: 54-59.

[36] James, G., Radchenko, P., and Lv. J. (2009). DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society*, **71**: 127-142.

[37] Kass, R. and Rafferty, A. (1995). Bayes Factors. *Journal of the America Statistical Association*, **98**: 438-455.

[38] Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix: problems in mean-variance optimization. *Journal of Portfolio Management*, **30**: 110-120.

[39] Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, **2**: 49-55.

[40] Marron, J., Todd, M., Ahn, J. (2007). Distance Weighted Discrimination. *Journal of the American Statistical Association*, **102**: 1267-1271.

[41] Meinshausen, N., Rocha, G., and Yu, B. (2007). A tale of three cousins: Lasso, L2Boosting, and Dantzig (discussion on Candes and Tao's Dantzig Selector paper). *Annals of Statistics*, **25**: 2372-2384.

[42] Mengersen, K.L. and Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, **24**: 101-121.

[43] Murre, JM; Sturdy, DP (1995). The connectivity of the brain: multi-level quantitative analysis. *Biological cybernetics*, **73**: 529-45.

[44] Paris, R.B. (2002). A uniform asymptotic expansion for the incomplete gamma function. *Journal of Computation and Applied Mathematics*, **148**: 323-339.

[45] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, **6**: 559-572.

[46] Rao, C.R., and Rao, M,B. (1998). *Matrix Algebra and Its Applications in Statistics and Econometrics.* World Scientific, Singapore.

[47] Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society.* **B, 59**: 731-792.

[48] Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods.* Second ed., Springer-Verlag: New York.

[49] Sarkar, S. and Chen, J. (2004). A Bayesian stepwise multiple testing procedure. Technical Report, Temple University.

[50] Schfer, J. and Korbinian, S. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**: 32.

[51] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**: 461-464.

[52] Shen, H. and Huan, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**: 1015-1034.

[53] Tai, Y. C., and Speed, T.P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Stat.*, **34**: 2387-2412.

[54] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal Royal Statistical Society B*, **58**: 267-288.

[55] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. *Statistical Science*, **18**: 104-117.

[56] Torokhti, A. and Friedland, S. (2009). Towards theory of generic Principal Component Analysis. *Journal of Multivariate Analysis*, **100**: 661-669.

[57] Ververidis, D. and Kotropoulos, C. (2009). Information Loss of the Mahalanobis Distance in High Dimensions: Application to Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**: 2275-2281.

[58] Waller, L.A., Turnbull, B.W. and Hardin, M.J. (1995). Obtaining Distribution Functions by Numerical Inversion of Characteristic Functions with Applications. *The American Statistician*, **49-4**: 346-350.