Improved Linear Discrimination Using Time-Frequency Dictionaries

Jonathan Buckheit and David Donoho
Stanford University, Stanford, CA 94305

Abstract

We consider linear discriminant analysis in the setting where the objects (signals/images) have many dimensions (samples/pixels) and there are relatively few training samples. We discuss ways that time frequency dictionaries can be used to adaptively select a small set of derived features which lead to improved misclassification rates.


1 Linear Discriminant Analysis

1.1 Review of the Classical Setting

Linear Discriminant Analysis (LDA), developed by Fisher in 1936, is perhaps the best known technique for classifying an observation of unknown class
membership into one of two populations on the basis of predictors (that is, covariates, or features). Fisher also had interest in classification into one of several populations, a subject developed further by C.R. Rao [14, 15]. It happens frequently in the literature that “classification” and “discrimination” and used interchangeably, though the former is used sometimes when the problem is one of “clustering.”

Given $G$ groups or classes $\Pi_1, \Pi_2, \ldots, \Pi_G$, we wish to develop a rule for predicting the class membership of an item based on the measurement of $p$ predictors or features $x \in \mathbb{R}^p$. To do so we have available a training set $(x_{i,g} : i = 1, 2, \ldots, N_g, g = 1, 2, \ldots, G)$ consisting of the $p$ features of $N$ items, each with known group membership. Letting $N_g$ denote the number of training items in class $\Pi_g$, we estimate mean vectors for each group, $\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} x_{i,g}$, $g = 1, 2, \ldots, G$. Let $\pi_g$ denote the prior probability of membership in group $g$, which may be estimated by $N_g/N$. Then the overall mean vector $\bar{x}$ with respect to this prior distribution is estimated via $\bar{x} = \sum_{g=1}^G \pi_g \bar{x}_g$.

The variability of samples within each group is measured by the group sample covariance matrices $S_g = \frac{1}{N_g} \sum_{i=1}^{N_g} (x_{i,g} - \bar{x}_g)(x_{i,g} - \bar{x}_g)'$, $g = 1, 2, \ldots, G$; in this way the overall within-group variability is estimated by the within-group covariance matrix $S_w = \sum_{g=1}^G \pi_g S_g$. Similarly, the between-group covariance matrix measures the dispersion of the group mean vectors about the overall mean vector: $S_b = \sum_{g=1}^G \pi_g (\bar{x}_g - \bar{x})(\bar{x}_g - \bar{x})'$.

Now, if $\tilde{X} = A'X$ denotes a linear transformation of the original variables, the between- and within-group covariance matrices for the transformed problem are just $\tilde{S}_b = A'S_bA$ and $\tilde{S}_w = A'S_wA$. Fisher’s idea was to maximize the between to within-group variance ratio, which in this case may be measured by the ratio of the determinants of the preceding two matrices (the determinant, being the product of the eigenvalues, is the product of the variances in the principal directions). The problem is thus

$$\tilde{A} = \arg \max_A \frac{|A'S_bA|}{|A'S_wA|},$$

the solution to which is given by the $\min(p, g-1)$ eigenvectors (called “canonical variates” by Rao) of $S_w^{-1}S_b$. 

2
The problem is then transformed to this “discriminant space” and a new (or test) sample \( x \) is assigned to the group \( g \) for which \( d(A'x, A'ar{x}_g) \) is smallest. Since \( A \) diagonalizes the within-class covariance matrix, the Mahalanobis distance in this space is just the usual Euclidean norm, so \( d(x, y) \) may be taken as \( \|x - y\|_2^2 \).

If the populations are multivariate normal with the same covariance matrix but different mean vectors, all of which are known, then Fisher’s heuristic approach turns out to be both the maximum likelihood rule and the Bayes rule with respect to the empirical prior distribution on the class memberships [14]. When the parameters are not known and need to be estimated, statistical asymptotic theory provides for the convergence (with \( N \)) of the sample estimates to the population parameters; LDA thus retains its optimality in an asymptotic sense. Several authors have discussed the robustness of LDA to departures from the underlying assumptions; see the discussion and references in McLachlan [15], for example.

1.2 Neo-Classical Setting

When we refer to the 

Neo-Classical Setting

in the following we mean: retain exactly the assumptions of the classical setting, but assume that \( p \gg N \). This is meant to model the situation where one is observing signals (images), where each signal (image) has thousands of samples (hundreds of thousand of pixels), and there are only a few hundred training samples. Think therefore of \( p = 10^5 \) or \( 10^6 \), but \( N < 1000 \), and of potential applications ranging from handwritten digit recognition to speaker identification. Authors such as Friedman [9] and Hastie et al. [10, 11, 12] have proposed various techniques to improve the performance of LDA in high-dimensional spaces.

To see why \( p \gg N \) is different from the classical case, assume that we are studying \( G = 2 \) groups with known covariance structure (without loss of generality, the identity covariance). We therefore consider a simplified LDA, where the covariance matrices need not be estimated. In this simplified LDA,
classification of a future observation $x$ is based on $\langle d, x \rangle$, where
\[
d = \tilde{x}_1 - \tilde{x}_2 = \Delta + \frac{1}{\sqrt{N}}\tilde{z}, \quad \tilde{z} \sim N(0, 1)
\] (2)
is the natural estimate of the true group difference vector, $\Delta = \mu_1 - \mu_2 = (\Delta_i)_{i=1}^p$.

Then a new test observation $x$ may be written as
\[
x = \mu + z, \quad \mu = \left\{ \begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right\}, \quad z \sim N(0, 1)
\] (3)
and we can decompose the statistic $\langle d, x \rangle = \langle \Delta, x \rangle + \frac{1}{\sqrt{N}}(\tilde{z}, x)$:

<table>
<thead>
<tr>
<th>Classical</th>
<th>$\langle \Delta, x \rangle$</th>
</tr>
</thead>
</table>
|                       | $= \langle \Delta, \mu \rangle + \langle \Delta, z \rangle$ | (4)
|                       | $= \langle \Delta, \mu \rangle + T_{Class}$ | (5)
|                       | $T_{Class} \sim N(0, \|\Delta\|^2)$. |

<table>
<thead>
<tr>
<th>Neo - Classical</th>
<th>$\frac{1}{\sqrt{N}}(\tilde{z}, x)$</th>
</tr>
</thead>
</table>
|                       | $= \frac{1}{\sqrt{N}}(\tilde{z}, \mu) + \frac{1}{\sqrt{N}}(\tilde{z}, z)$ | (6)
|                       | $= T_{Neo,1} + T_{Neo,2}$ | (7)
|                       | $T_{Neo,1} \sim N(0, \|\mu\|^2/N)$, |
|                       | $T_{Neo,2} \tilde{z} \sim N(0, \|\tilde{z}\|^2/N)$. |

Here the label “Classical” refers to discrimination with the “ideal” discriminator $\langle \Delta, x \rangle$, which has a misclassification rate determined by the signal-to-noise ratio $\rho_{\text{ideal}} = \|\Delta\|^2$.

The key point about the “Neo-Classical” term is that $\text{Var}(T_{Neo,2}) = E \left[ E(T_{Neo,2}^2 | \tilde{z}) \right] = p/N$. In the classical case, $p$ is fixed and $N$ is large, or even tending to infinity, and so this term can be neglected. In the Neo-Classical case, $p/N \gg 1$.

The performance of simplified LDA is based on the average signal-to-noise ratio $\rho_{\text{Actual}} = \|\Delta\|^2/\text{Var}(\langle d, x \rangle)$. We have, approximately, $\rho_{\text{Actual}} \approx$
\[ \| \Delta \|^2 / (1 + p/N) \]. Hence if \( p/N \gg \| \Delta \|^2 \), the noise term \( T_{\text{noise}} \) dominates the simplified LDA statistic and discrimination should appreciably suffer.

Figure 1 illustrates an extreme example of this phenomenon with \( N = 2 \) (one signal from each class) and \( p = 2^{15} \). (The covariance matrix, clearly unestimable, was taken to be the identity.) In a problem for which ideal discrimination would be extremely successful (\( \rho_{\text{ideal}} \gg 1 \)), simplified LDA completely fails on test data, yielding error rates approaching \( 50\% \). Hence for large values of \( p \), the need to estimate the group difference vector may lead to a fundamental breakdown in the LDA method. Although this example is admittedly contrived, such ratios of \( p/N \) would not be uncommon in the analysis of large images.

![Figure 1: Ill-Posed LDA](image)

### 1.3 De-Noising and Discrimination

The analysis of the previous section shows that discrimination with high \( p/N \) and known covariance is essentially an issue of dealing with the noise
in the empirical group-difference vector $d$. This suggests that methods for removing noise from $d$ should yield improvements over straightforward LDA. One would somehow process $d$ and replace it by a vector $\hat{\Delta}$ which was less noisy, doing discrimination based on $\langle \hat{\Delta}, x \rangle$.

There are many methods for removing noise from signals, ranging from simple smoothing techniques to more sophisticated schemes based on de-noising in the wavelet domain [7, 8]. Many such methods could be profitably adapted to improving discrimination in the Neo-Classical setting. We shall describe some below. All such methods operate by, in essence, operating on the noisy signal and downweighting components which are likely to consist of noise rather than signal. For example, smoothing methods keep low frequencies and kill high frequencies. Wavelet DeNoising methods keep a small number of wavelet coefficients and kill the others. The result is an object made up of a small number of “reliable” components, each making a large contribution to the signal and relatively small contributions to the noise. In terms of the previous section, all such methods operate by first discarding many features, arriving at a small set, and thereby effectively reducing $p$ so that it is much smaller than $N$.

An important point: the specific quantitative goal – misclassification rate – is very different from goals like mean-squared error in usual signal de-noising. For example, mean-squared error is not very closely related to discriminant performance. What is important is the effective signal-to-noise ratio. It can turn out that $\hat{\Delta}$ is quite bad as a detailed estimate of $\Delta$, while the signal to-noise ratio is 4 or so, in which case all but a few percent of observations are correctly classified. Because of this remark, existing methods of de-noising will have to be re-thought and/or recalibrated to work well in de-noising an empirical group difference vector.
2 Time-Frequency De-Noising & Discrimination

2.1 Dictionaries, Atoms and Signal Representations

Let \( s = \{s_t : 0 \leq t < n \} \) be a discrete-time signal of length \( n \). Using terminology introduced by Mallat and Zhang [13], define a dictionary as a collection of parametrized waveforms \( \mathcal{D} = (\phi_\gamma : \gamma \in \Gamma) \). The waveforms \( \phi_\gamma \) are discrete-time signals of length \( n \) called atoms.

A dictionary is said to be complete if it contains exactly \( n \) linearly independent atoms; there is then a unique representation of \( s \) with respect to this dictionary. If a dictionary contains more than \( n \) atoms, it is said to be over-complete.

The parameter \( \gamma \) may index:

- frequency, in which case the dictionary is a frequency dictionary (e.g. the Fourier dictionary).
- time and scale jointly, in which case the dictionary is a time-scale dictionary (e.g. wavelet dictionaries).
- time and frequency jointly, in which case the dictionary is a time-frequency dictionary (e.g. wavelet- and cosine-packets dictionaries [5, 18]).

We wish to represent \( s \) as a linear combination of waveforms in a particular dictionary. Given a discrete dictionary of \( p \) waveforms, collect these waveforms as columns of an \( n \) by \( p \) matrix \( \Phi \), say. The decomposition problem is then finding \( \alpha = (\alpha_\gamma) \) such that \( \Phi \alpha = s \). Several methods for performing this decomposition have been proposed in the literature:

- **Best Orthogonal Basis** [6, 18]. Certain special subcollections of the
wavelet-packets and cosine-packets dictionaries amount to orthogonal bases; in fact the standard WP and CP dictionaries for a signal of length \( n \) contain \( \geq 2^n \) such orthogonal bases. Coifman and Wickerhauser have proposed a method of adaptively picking from among these many bases a single orthogonal basis which they term the “best basis,” based on entropy measures of the coefficients. Their algorithm is fast, working in order \( n \log(n) \) time.

- **Matching Pursuit** [13] starts with an initial approximation \( s^{(0)} = 0 \) and residual \( R^0 = s \). At stage \( k \) it identifies that dictionary atom which best correlates with the residual and then adds to the current approximation a scalar multiple of that atom, so that \( s^{(k)} = s^{(k-1)} + \alpha_k \phi_{\alpha_k} \), where \( \alpha_k = \langle R^{(k-1)}, \phi_{\alpha_k} \rangle \) and \( R^{(k)} = s - s^{(k)} \). This algorithm works perfectly for orthogonal dictionaries; for more general dictionaries examples have been constructed which badly foil the method [4].

- **Basis Pursuit** [4] minimizes \( \|\alpha\|_1 \) subject to \( \Phi \alpha = s \) using linear programming methods.

These methods have extensions to producing approximate decompositions; actually Matching Pursuit produces such approximations to start with. They lead to representations of signals using many fewer than the nominally required number of coefficients.

This is potentially connected with the discrimination problem because if applied to the empirical group differences, it would amount to a reduction in the dimensionality \( p \) and an amelioration of the ratio \( p/N \). All three of the above methods have versions adapted to improving the performance of linear discrimination; we describe in detail here only the adaptation of Matching Pursuit.

### 2.2 Discriminant Pursuit

For \( G \)-group discrimination, \( G(G - 1)/2 \) contrasts may be estimated:

\[
\{ d_{ij} = \bar{x}_i - \bar{x}_j : 1 \leq i < j \leq G \}. \tag{8}
\]
It may be appropriate to rescale the contrasts by the pooled estimate of covariance, $S_p = \sum_{g=1}^G (N_g - 1) S_g/(N - G)$, where $S_g$ is the sample covariance matrix of group $g$, i.e., compute $S_p^{-1} d_{ij}$ instead. However, this rescaling may act to downweight variables which effectively discriminate between the groups.

For a given dictionary (in the examples that follow, we use the wavelet- and cosine-packets dictionaries), we may use the matching pursuit algorithm to successively pick out time-frequency features one at a time. This is done by considering the atom that best correlates with the each of the contrast residuals at a given stage of the algorithm, and then among these picking the one with the highest such correlation. This approach gives equal weight to each contrast, assuming they all have comparable discriminatory power. However, in many discrimination problems certain groups are much harder to discriminate between than others, and it makes sense to give preference to contrasts which provide more discriminatory power. One such measure of discriminatory power or group separation is the Mahalanobis distance $D_{ij}$, defined as $D_{ij}^2 = (\bar{x}_i - \bar{x}_j) S_p^{-1} (\bar{x}_i - \bar{x}_j)'$. The algorithm may then weight the best-atom correlations for each contrast by its Mahalanobis distance before it picks a feature.

The obvious question of how many atoms to select may involve a good deal of subjectivity, but the same standards for stopping the matching pursuit algorithm on the analysis of a single signal may be applied to the present situation. These usually involve stopping when the magnitude of the residual falls below a certain value, which may be related to the estimated noise level of the data, but may simply amount to picking a fixed percentage (say 5%) of the original problem dimension $p$. Unlike traditional applications of time-frequency constructions, the goal here is discriminatory power, not reconstruction fidelity; in fact, overfitting the contrasts estimated from the training data may significantly decrease the flexibility in classifying test data. Stopping the matching pursuit algorithm at $k < p$ iterations amounts to an implicit hard-thresholding of the complete MP representation of $p$ basis elements [7, 8]; in this sense, denoising is built into the algorithm.

Assuming $k \ll p$ atoms have been selected, we have constructed a representation that has reduced the dimensionality of the problem — the statistical
consequences of which have been already discussed – and has acknowledged the spatial and/or time-frequency character of the data by its projection onto a low-dimensional basis consisting of atoms that are well-localized in the time-frequency plane.

It should be noted that the classical method of extracting features from a high-dimensional discrimination problem, the Karhunen-Loève expansion [15] (called principal components by statisticians), suffers from the same estimation problems in high-dimensional settings that does LDA itself, a problem which has recently been studied by several authors [16, 19].

2.3 Extensions to Best Orthogonal Basis

We have tried an extension of Best Orthogonal Basis de-noising ideas to improve LDA. In the setting where there are only \( G = 2 \) groups, we have defined the quality of a basis \( B \) by a proxy for its discrimination power. We use a proxy because the actual misclassification rate is not an “additive measure of information” in the Coifman-Wickerhauser sense, while the proxy is. Our proxy is essentially the inverse of a cumulative normal distribution applied to the misclassification rate.

Using this proxy we optimized over all bases in the wavelet packets or cosine packets library to find the basis which is “best” for classification. In the selected basis, the vector \( \hat{\Delta} \) is obtained by thresholding the coefficients, keeping only the (hopefully) few coefficients that convincingly exceed the noise level.

We remark of course that Saito and Coifman [17] have also, in pioneering work, talked about “Best Bases for Discrimination,” and have used wavelet- and cosine-packets libraries successfully for classification. Their ideas seem, to us, to be tied more to quadratic discrimination than linear discrimination. Also, they do not deal directly with operating characteristics of their method (e.g. misclassification rate) but rather on formal analogies of their objective with, say, Kullback-Leibler distances. Our approach is, in contrast, more directly tied to a specific theoretical picture about how to improve discrimini-
nation performance, which we hope to spell out completely elsewhere.

2.4 Extensions to Basis Pursuit

Basis Pursuit is easy to generalize to the case of simplified LDA with several group contrasts. Regarding the contrasts as synthesis operators on the data, $d_{ij} = S\theta$, a variant on Basis Pursuit denoising could be used to select discrimination features:

$$\min_{\theta} \sum_{ij} \lambda \|\theta\|_{\ell_1} + \|d_{ij} - S\theta\|_{\ell_2}.$$  

We are currently experimenting with this approach.

3 Examples

3.1 Speech Recognition

The following data was kindly supplied by Trevor Hastie, having been analyzed in his paper on Penalized Discriminant Analysis [10]. The training set consists of 1633 speech frames of 32 msec duration from the continuous speech of twelve male speakers, extracted from the TIMIT speech database. It is a five-class problem consisting of five phonemes: “sh” as in “she,” “dcl” as in “dark,” “iy” as the vowel in “she,” “aa” as the vowel in “dark” and “ao” as the first vowel in “water.” The phonemes occur with the following frequencies in the training set:

<table>
<thead>
<tr>
<th>phoneme</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>sh</td>
<td>389</td>
</tr>
<tr>
<td>iy</td>
<td>151</td>
</tr>
<tr>
<td>aa</td>
<td>254</td>
</tr>
<tr>
<td>dcl</td>
<td>361</td>
</tr>
<tr>
<td>ao</td>
<td>478</td>
</tr>
</tbody>
</table>
Log-periodograms, a widely used method for representing speech frames for purposes of discrimination, were computed for each training sample. Random samples of ten log-periodograms from each phoneme class in the training sample are plotted in Figure 2.

![Figure 2: Phoneme Data](image)

It is well-known that the periodogram is a poor estimate of the spectral density function unless some type of smoothing is done. Figure 2 clearly shows this erratic behavior. For the situation of linear discriminant analysis, this again manifests itself in the issue of problem dimension versus sample size; indeed, some of the phoneme frequencies in the training set are less than the problem dimension of 256. Figure 3 shows the four LDA canonical variates which, as expected, show little structure beyond noise, except perhaps at the very low frequencies.

The test set consists of 1661 speech frames from twelve different male speakers. LDA yields error rates of 1.4% on the training data and 16.6% on the test data, indicating a fair amount of over-adaptation taking place. Table 1 gives the confusion matrix of true class versus predicted class for LDA on the test data, which indicates that it has a difficult time telling

12
“dcl” apart from “ao” in the test set.

We used the wavelet- and cosine-packets dictionaries generated by the Coiflet 3-tap wavelet and sinusoidal taper, respectively, to extract time-frequency features for this dataset. In this case, the top twenty-five were selected, and (after examining the LDA confusion matrix), covariance rescaling and Mahalanobis weighting were used. The top five WP basis functions are given in Figure 4.

Table 2 summarizes the discrimination performance for LDA, PDA, and LDA on the top twenty-five WP and CP dictionary atoms. WP (surprisingly performing better than CP on this example) narrowed the wide gap between the LDA training and test rates, the latter of which was improved more than 100%.
Predicted Class

<table>
<thead>
<tr>
<th></th>
<th>sh</th>
<th>iy</th>
<th>aa</th>
<th>dcl</th>
<th>ao</th>
</tr>
</thead>
<tbody>
<tr>
<td>sh</td>
<td>374</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>iy</td>
<td>3</td>
<td>147</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aa</td>
<td>0</td>
<td>0</td>
<td>266</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dcl</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>285</td>
<td>64</td>
</tr>
<tr>
<td>ao</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>203</td>
<td>313</td>
</tr>
</tbody>
</table>

Table 1: LDA Confusion Matrix, Phoneme Test Data

Figure 4: Wavelet-Packets Basis Functions, Phoneme Data
Table 2: Phoneme Data

<table>
<thead>
<tr>
<th>Technique</th>
<th>Error Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>LDA</td>
<td>.014</td>
</tr>
<tr>
<td>PDA</td>
<td>.083</td>
</tr>
<tr>
<td>WP</td>
<td>.089</td>
</tr>
<tr>
<td>CP</td>
<td>.094</td>
</tr>
</tbody>
</table>

### 3.2 Waveform Data

Although this example is not one of high-dimensional discrimination, it is included to illustrate that the extraction of time-frequency features can improve the LDA classification rate in the classical setting as well. It is taken from Breiman et al. [1] and is considered a difficult problem in pattern recognition. It is a three-class problem with \( p = 32 \), extended from the original problem dimension of 21 to satisfy the dyadic length requirements of the fast algorithms. The predictors are defined by

\[
  x_i = u h_1(i) + (1 - u) h_2(i) + \epsilon_i \quad \text{Class 1}
\]

\[
  x_i = u h_1(i) + (1 - u) h_3(i) + \epsilon_i \quad \text{Class 2}
\]

\[
  x_i = u h_2(i) + (1 - u) h_3(i) + \epsilon_i \quad \text{Class 3}
\]

where \( i = 1, 2, \ldots, 32 \), \( u \) is uniform on \((0,1)\) and the \( \epsilon_i \) are standard normal variates. The \( h_i \) are the shifted triangular waveforms: \( h_1(i) = \max(6 - |i - 7|, 0) \), \( h_2(i) = h_1(i - 8) \) and \( h_3(i) = h_1(i - 4) \).

Training samples of 500 observations and test samples of 300 observations were generated. Random samples of ten waveforms from each class are plotted in Figure 5. The two canonical variates extracted by LDA (Figure 6) show little structure. To extract time-frequency features, the wavelet-packets dictionary generated by the Coiflet 3-tap wavelet and the cosine-packets dictionary generated by a sinusoidal taper were used, and then the top five discriminatory atoms were selected from each. No covariance rescaling or
Mahalanobis weighting were used. The five wavelet-packets atoms that were selected are plotted in Figure 7, and measure the functions $h_i$ and their derivatives.

![Waveform Data](image)

**Figure 5: Waveform Data**

Because the dataset is generated stochastically, 25 simulations were run in order to accurately assess discrimination performance. Table 3 lists performance for standard LDA, LDA on the top five WP and CP atoms, CART, as well as FDA (using MARS basis functions) and GMDA, taken from Hastie *et al.* [11, 12]. Standard errors are in parentheses.

The best result was obtained using the WP dictionary, achieving an error rate of 15% on the test data. According to Breiman *et al.* [1], the Bayes error rate for this example is 14%; no empirical procedure is likely to do better.
Figure 6: LDA Canonical Variates, Waveform Data

Figure 7: Wavelet-Packets Basis Functions, Waveform Data
4 Discussion

4.1 Reproducible Research

In adhering to our principle of Reproducible Research [3], all the figures and tables in this article may be reproduced using the WAVELAB software package [2], in the directory Papers/DiscrPursuit. In this way the publication of this paper, and especially the figures contained herein, is not merely the advertising of a specific computational result, but the publication of the algorithm, of datasets, and of a series of scripts which invoke the algorithm and the datasets. The figures in this paper are only the visible result of this publication process.

4.2 Acknowledgments

We would like to thank Trevor Hastie for supplying the phoneme data and acknowledge Trevor Hastie, Richard Olshen and Rob Tibshirani for helpful discussions.
References


