
ISSN 0973-1377

International Journal of Applied Mathematics & Statistics



Volume 4

Number J06

June 2006

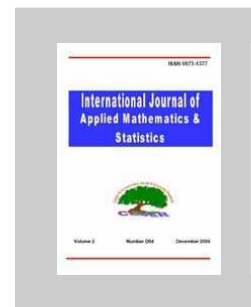
International Journal of Applied Mathematics & Statistics

ISSN 0973-1377

Editorial Board

Editor-in-Chief:

Florentin Smarandache
University of New Mexico
200 College Road
Gallup, NM 87301,
USA.
Email: smarand@unm.edu, ceser_info@yahoo.com
Tel.: (505) 863-7647 (Office)
Tel. & Messages: (505) 726-1720 (Home)
Fax: (505) 863-7532 (Attn.: Prof. Smarandache)



Executive Editor:

Tanuja Srivastava
Department of Mathematics,
Indian Institute of Technology
Roorkee-247667,
INDIA
Email: tanujfma@iitr.ernet.in

Editors

| | | |
|-----------------------------|--|-----------|
| John Michael Rassias | University of Athens | Greece |
| Akca Haydar | United Arab Emiartes University | UAE |
| R. K. S. Rathore | Indian Institute of Technology, Kanpur | India |
| Larissa Borissova | Institute of Theoretical & Exper. Biophysics | Russia |
| Alexandru Murgu | British Tel. Net. Research Centre | UK |
| Hans Gottlieb | Griffith University, | Australia |
| Somesh Kumar | Indian Institute of Technology, Kharagpur | India |
| Edward Neuman | Southern Illinois University | USA |
| Dmitri Rabounski | Institute of Theoretical & Exper. Biophysics | Russia |
| S. P. Sharma | Indian Institute of Technology, Roorkee | India |

Associate Editors:

| | | |
|---------------------------------|---|-----------|
| W.B.Vasantha Kandasamy | Indian Institute of Technology | India |
| Sukanto Bhattacharya | Alaska Pacific University | USA |
| M. Khoshnevisan | Griffith University | Australia |
| Marcin Kozak | Warsaw Agricultural University | Poland |
| Alexander Grigorash | University of Ulster | U.K. |
| Ferhan Atici | Western Kentucky University | USA |
| Karen Yagdjian | University of Texas-Pan American | USA |
| Rui Xu, | University of West Georgia | USA |
| Abdollah Khodkar | University of West Georgia, | USA |
| Bogdan G. Nita | Montclair State University | USA |
| Weijiu Liu | University of Central Arkansas | USA |
| Alexander A. Katz | St. John's University | USA |
| Wiesław A. Dudek | Wrocław University of Technology | Poland |
| Ki-Bong Nam | University of Wisconsin, Whitewater, | USA |
| Diego Ernesto Dominici | State University of New York, New Paltz | USA |
| Wen-Xiu Ma | University of South Florida | USA |
| Ming Fang, | Norfolk State University | USA |
| Hemant Pendharkar | Worcester State College | USA |
| Delfim F. M. Torres | University of Aveiro | Portugal |
| Tianzi Jiang | Chinese Academy of Sciences | PR China |
| Rodrigo Capobianco Guido | University of S.ao Paulo. | Brazil |
| V. Ravichandran | Universiti Sains Malaysia | Malaysia |

Assistant Editors:

| | | |
|----------------------------------|--|---------|
| Irene Sciriha | University of Malta | Malta |
| Răzvan Răducanu | Al. I. Cuza University | Romania |
| Anahit Ann Galstyan | University of Texas-Pan American | USA |
| Oliver Jones | California State University, | USA |
| Doreen De Leon | California State University, | USA |
| Jose Almer T. Sanqui | Appalachian State University | USA |
| Miranda I. Teboh-Ewungkem | Lafayette College | USA |
| Guo Wei | University of North Carolina at Pembroke | USA |
| Michael D. Wills | Whitman College | USA |
| Alain S. Togbe, | Purdue University North Central | USA |
| Rogemar S Mamon | Brunel University, | UK |
| Anna Karczewska | University of Zielona Góra | Poland |
| Bixiang Wang | New Mexico Inst. of Mining & Technology | USA |
| Andrei Volodin | University of Regina | Canada |
| Samir H. Saker | Mansoura University, | Egypt. |
| Eduardo V. Teixeira | Rutgers University, | USA |
| Ashwin Vaidya | Florida State University | USA |
| Ganatsiou V. Chrysoula | University of Thessaly | Greece |
| Xiaoli Li | University of Birmingham | UK |
| Guangyi Chen | University of Montreal | Canada |

International Journal of Applied Mathematics & Statistics

ISSN 0973-1377

Contents

| Volume 4 | Number J06 | June 2006 |
|---|-------------------|------------------|
| Regularization of Projection Directions via Best Basis Selection Approach | | 1 |
| Inna Stainvas and Nathan Intrator | | |
| A Simple Proof of the Constancy of the Pontryagin Hamiltonian for Autonomous Problems | | 23 |
| Delfim F. M. Torres | | |
| Fitting Smooth Paths on Riemannian Manifolds | | 25 |
| Luís Machado and F. Silva Leite | | |
| Retraction of Simplicial Complexes | | 54 |
| M. El-Ghoul, A. E. El-Ahmady and T. Homoda | | |
| Normal Mode Expansion Method for Generalized Thermoelastic Lamb Waves in Transversely Isotropic Thin Plates | | 68 |
| K.L. Verma | | |
| A Common Fixed Point Theorem in Fuzzy Metric Spaces | | 84 |
| Reza Saadati and Shahriar Yousefi | | |
| A Remark on Noether's Theorem of Optimal Control | | 88 |
| Ilona A. Dzenite and Delfim F. M. Torres | | |

Regularization of Projection Directions via Best Basis Selection Approach

Inna Stainvas¹ and Nathan Intrator²

¹ FPD, Orbotech.Ltd
P.O. Box 215, Yavne
Israel
inna-s@orbotech.com

² Computer Science Department
Sackler Faculty of Exact Sciences
Tel-Aviv University
Israel

Abstract

Classification and recognition of high-dimensional data is difficult due to the “curse of dimensionality” problem, i.e. it is not enough data to robustly train an estimator. The problem may be overcome by dimensionality reduction. Many statistical models, such as linear discriminant analysis (LDA) and neural networks (NNs), for example, include dimensionality reduction as an implicit preprocessing step. However, such projection onto discriminant directions is not sufficient since the number of direction parameters still remains large (proportional to dimensionality of the data); and models persist to be many parameter models and require regularization.

In this work, we propose to regularize the low-dimensional structure of the projection parameter space based on compression concepts. We assume that an intrinsic dimensionality of the discriminant space spanned by projection directions is essentially small and the latter may be sufficiently well represented as a linear superposition of a small number of wavelet functions in the wavelet packet basis. We further, introduce a simple incremental way to increase the dimensionality of the parameter space using hypothesis testing and apply the technique to logistic regression and to Fisher linear discrimination.

Three benchmark data-sets: triangular waveforms (Breiman 1984), the vowel data-set (CMU repository) and a letter data set (DELVE) are used to demonstrate the proposed method. We show that this approach leads to significant classification improvement.

Keywords: Dimensionality reduction, Projection methods, Best-Basis wavelet packet.

2000 Mathematics Subject Classification: 62H30, 65T60, 55R15.

1 Introduction

It is well known that for a proper regularization, the complexity of a model should be matched to the complexity of the data and to the number of training patterns. Often, this is done by adding a

cost-complexity term to the optimization process so that both terms are concurrently optimized. Some common cost terms include smoothness (which leads to splines) (Wahba, 1990), entropy related terms or simple parameter shrinkage which is also called weight decay (Krogh and Hertz, 1992). Characterizing model complexity by its description length or some other indirect measure such as smoothness, is the most common regularization.

Assuming some kind of structure in the data is a useful regularization and capacity control technique. Hinton and Nowlan (1992) first extended this idea to search for structure in the parameter space as well. They looked for a mixture of Gaussians weight distribution. In a recent work, it was argued that a useful way to characterize the capacity of a model is by description length; i.e. they suggest to optimize concurrently the model description and description length of the residuals (Hinton and Zemel, 1994).

Characterizing model complexity by its description length or some other indirect measure such as smoothness, does not directly constrain the internal structure of the model, namely, this is a *variance* constraint only. The effects of constraining only variance vs. providing constraints that affect both variance and bias were first discussed in (Intrator, 1993) and later in (Intrator, 2000). To distinguish between constraints that affect also the variance, one should turn off the main optimization goal and see whether the remaining optimization goal (the additional constraint or penalty) takes the set of parameters into a meaningful solution or otherwise, take the set of parameters to a zero or fixed solution. For example, exploratory projection pursuit constraint (Intrator, 1993) fall into the bias constraints category, as well as reconstruction constraints (Stainvas, 1999).

We propose here to use a more detailed constraint than the description length and to actually measure the dimensionality of the space where the parameters reside. It is not entirely clear that characterizing a model by the dimensionality of the parameter space is at all useful. We intend to demonstrate this in this paper. We further, introduce a simple incremental way to increase the dimensionality of the parameter space and apply the technique to logistic regression and to Fisher linear discriminant method (Fisher, 1936).

Several challenging data sets are used to demonstrate the usefulness of the proposed method. In a subsequent paper, the potential of this method will be explored on a feed-forward non-linear neural network architecture.

The rest of the paper is organized as follows: Section 2 presents our method and its subsections review a necessary background on the relevant discriminant techniques and wavelet based compression techniques. Three benchmark classification problems are results using the proposed methodology are presented in Section 4. Finally, the paper is concluded by discussion in Section 5.

2 Methodology

The proposed approach is quite simple. One first extracts model parameters in the 'classical' way, namely using a regular parameter optimization of a model and a cost complexity term. In a simple case, one can extract the Fisher linear discriminant directions for a multi-class problem (Fisher, 1936). Then, a post-model processing is performed by constraining the model

parameters into a nested sequence of subspaces. Although the nested sequence is in space of parameters and not data points, still the theory of likelihood ratios holds so that the significance of adding another dimension can be tested using the χ^2 test.

A nested sequence of subspaces can be naturally defined via compression of the model parameters using Fourier or wavelet transforms and such advanced techniques as matching pursuit (MP) (Mallat and Zhang, 1993) or best basis approach (Coifman and Wickerhauser, 1993). These techniques allow us to decompose discriminant directions into linear combinations of specific basis functions with most of coefficients close to zero. Based on the values of the decomposition coefficients basis functions are ordered by significance and nested subspaces are spanned by an increasing number of basis functions.

2.1 Methodological details

We consider parameter spaces spanned by a set of projection directions extracted by linear discriminant analysis (LDA) (Fisher, 1936; Duda and Hart, 1973; Fukunaga, 1990) or its simplified version called contrast vectors (Buckheit and Donoho, 1995). Two algorithms – best basis (BB) (Coifman and Wickerhauser, 1992) and matching pursuit (MP) (Mallat and Zhang, 1993) – are applied to compress discriminant directions considered as a set of signals. We demonstrate that the best basis approach better fits the data and thus gives better results for a small number of basis functions.

Instead of classifying the original data, they are first projected either onto: (i) a parameter subspace spanned by a small number of leading wavelet basis functions, (ii) reconstructed discriminant directions. Secondly, the projected data are classified using two methods: (i) LDA and (ii) the first nearest neighbor classification rule. A number of basis wavelet functions is defined either empirically or using χ^2 hypothesis test. Another possibility is to minimize misclassification rate using bootstrap or cross-validation methods.

2.2 Linear Discriminant Analysis

LDA was originally proposed by Fisher to find an orientation w in which data coming from two classes are best separated (Fisher, 1936). This was expressed as a maximization of the Fisher linear discriminant function:

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

where \tilde{m}_i , $i = 1, 2$ and \tilde{s}_i , $i = 1, 2$ are class means and variances of the data projected onto w direction. This measure was generalized for a multi-class case to be large when data points belonging to the same class are well concentrated, and class centers are as far as possible from each other. Several ways to introduce separability measures were proposed (Fukunaga, 1990). The most used separability measures are given by:

$$J_1 = tr(\tilde{S}_w^{-1} \tilde{S}_b) \tag{2.1}$$

$$J_2 = det(\tilde{S}_w^{-1} \tilde{S}_b) = \frac{det(\tilde{S}_b)}{det(\tilde{S}_w)} \tag{2.2}$$

where within class scatter matrix \tilde{S}_w and between class scatter matrix \tilde{S}_b of the projected data onto the space spanned by the columns of matrix $W_{[p \times M]}$ are given by:

$$\tilde{S}_b = W^T S_b W, \quad \tilde{S}_w = W^T S_w W.$$

Matrix S_w of size $[p \times p]$ is a within class scatter matrix of an original data, that is an average of the class covariance matrices and S_b of size $[p \times p]$ is a between class scatter matrix of the original data:

$$S_w = \sum_{g=1}^G \pi_g Cov_g$$

$$Cov_g = \frac{1}{N_g} \sum_{x_i \in g} (x_i - m_g)(x_i - m_g)^T$$

$$S_b = \sum_{g=1}^G \pi_g (m_g - m^0)(m_g - m^0)^T$$

where $m^0 = \sum_{g=1}^G \pi_g m_g$; m_g and π_g are mean values and prior probabilities of the classes, respectively.

The best feature space of the chosen dimensionality m , maximizing J_1 and J_2 , is spanned by the first m eigenvectors of the matrix $S_w^{-1} S_b$, called *canonical variates*. In addition, any nonsingular transformation of this space does not change separability measures (2.1–2.2). The rank of the between scatter matrix S_b does not exceed the minimum between data dimensionality p and $(G - 1)$: $rank(J_1) \leq \min(p, G - 1)$; this leads to the same upper bound on the allowed feature space dimensionality M . When the dimensionality M grows, both separability measures grow as well.

In this reduced feature space, classification problem is usually completed by assigning a sample to a class that minimizes an Euclidean distance between the sample and class centroids. However, other classification methods, such as nearest neighbor classification rules or neural network methods may be applied¹.

2.3 Contrast vectors

Contrast vectors were introduced considering a simplified version of LDA for a two-class case and with an isotropic within-scatter matrix S_w . In a two-class case, the first separability measure (2.1) exactly deduces to the Fisher discriminant measure (Duda and Hart, 1973) and a unique canonical variate satisfies to $S_b w = \lambda S_w w$. The between-scatter matrix may be represented as $S_b = \pi_1 \pi_2 (m_1 - m_2)(m_1 - m_2)^T$ and thus $S_b w$ is always oriented in the same direction as a difference between class means $m_1 - m_2$. Fisher solution may be written as: $w = S_w^{-1} (m_1 - m_2)$. For a very high dimensional data the number of samples is not enough to robustly estimate the within-scatter matrix and it may be singular. In this case, it is often regularized by considering a regularized within scatter matrix $S_w^R = S_w + \lambda I$ (I - is a spheric isotropic matrix). When λ is large, this indeed ignores the data covariance and leads to a

¹Theoretically, the best classifier, i.e. Bayesian, depends on the posterior probabilities of the projected data

solution that is a difference between class means $m_1 - m_2$ that requires less training data for estimation.

A simplified version of LDA generalized for a multi-class problem considers all possible differences between class means taken pairwise to be discriminant projection directions² (Buckheit and Donoho, 1995). The obtained discriminant projections are referred to as contrast vectors. We simplify this further on considering only a subset of contrast vectors that are linear independent.

2.4 Best Basis (BB) Algorithm

The best-basis algorithm was proposed by Coifman and Wickerhauser (1992) for signal compression. This method expands a signal into a library of orthonormal bases that has a binary tree structure. The nodes of the wavelet packet tree represent subspaces with different time-frequency localization characteristics. For each node of the tree (j, k) (where j is a depth of the node in the tree) its two children nodes $(j + 1, 2k)$ and $(j + 1, 2k + 1)$ represent subspaces that are mutually orthogonal and basis vectors of each tree node $B_{j,k}$ are mutually orthogonal as well:

$$\begin{aligned}
 W_{j,k} &= W_{j+1,2k} \oplus W_{j+1,2k+1} \\
 j &= 0, \dots, J ; k = 0, \dots, 2^j - 1 ; J = \log_2(p) \\
 B_{j,k} &= \{w_{j,k}^l\}, l = 0, \dots, 2^{p-j}, \quad \langle w_{j,k}^l, w_{j,k}^m \rangle = \delta_{lm}
 \end{aligned}$$

This representation is essentially redundant and provides more than 2^p complete orthogonal bases (p - is the number of signal sampling points). The algorithm chooses a complete orthogonal basis from a library of orthonormal bases $\{B_{j,k}\}$ that optimizes the signal *representation cost*. This representation cost is the entropy of the vector of the signal energies in the wavelet basis directions. In other words, a basis in which a normalized signal f has a minimum entropy is selected, i.e.

$$f = \sum_{i=1}^p \alpha_i \phi_i \quad \sum_{i=1}^p \alpha_i^2 = \|f\|^2 = 1$$

where ϕ_i are the orthonormal waveforms from the wavelet packet and

$$H_\phi(f) = H(\{\alpha_i\}) = - \sum_{i=1}^p \alpha_i^2 \log_2 \alpha_i^2 \tag{2.3}$$

is minimal. This measure is large when all α_i are about the same and small when a lot of $\alpha_i = 0$, thus it is a good measure of concentration or efficiency of an expansion. Such basis selection allows a good compression rate when basis functions corresponding to small coefficients α_i are neglected. This measure is an additive measure in the sense that $H(f) = H(\{\alpha_i\}) = \sum_i h(\alpha_i)$ (this definition of additiveness and method may be easily generalized to $H(\{\alpha_i\}) =$

²Buckheit and Donoho show that due to noise in a contrast vector, a variance of the data projected onto contrast vector contains an additional term that is proportional to the ratio between data dimensionality and the number of training samples. This term may be very large for high dimensional data and thus contrast vectors as well as canonical variates have to be regularized.

$\sum_i h_i(\alpha_i)$). Due to “additiveness” the algorithm proceeds bottom-up comparing the goodness of each subspace to that of union of its two children nodes and decides whether to keep the node or replace it by its children. In other words, by induction on j in the inverse direction: $j = J - 1, \dots, 1, 0$; the best basis $A_{j,k}$ ($k = 0, \dots, 2^j - 1$) that spans the space $W_{j,k}$ is defined by:

$$A_{j,k} = \begin{cases} B_{j,k}, & \text{if } H(B_{j,k}f) < H(A_{j+1,2k}f) + H(A_{j+1,2k+1}f) \\ A_{j+1,2k} \oplus A_{j+1,2k+1}, & \text{otherwise} \end{cases}$$

where initial $A_{J,k} = B_{J,k}$. In the end of the procedure, we get the best basis $A_{0,0}$ that spans the space $W_{0,0}$ that is the space of the original signal.

For a collection of N signals $\{s^n\}$ a basis $\{\phi_i\}$ that is good for all signals' representation is considered, i.e. that minimizes (2.3) with α_i^2 being an average energy of all signals in the direction ϕ_i :

$$\alpha_i^2 = \frac{1}{N} \sum_{n=1}^N \frac{(\langle s^n, \phi_i \rangle)^2}{\|s^n\|^2}$$

The algorithm permits a fast search with $O(pN \log p)$ steps.

2.5 Matching Pursuit (MP) Algorithm

Matching pursuit (MP) was proposed in (Mallat and Zhang, 1993) to decompose any signal or a set of signals into a linear expansion of waveforms that are selected from a redundant dictionary of basis functions (orthogonal wavelet packet in our consideration). These waveforms are chosen to best match signal structures. An algorithm proceeds iteratively starting from an initial approximation $s^{l,0} = 0$ and residuals $R^{l,0} = s^l$. For each approximation k , an atom that best correlates (matches) with one of the residuals $R^{l,k-1}$ is sought and next signals' approximations $s^{l,k}$ and residuals $R^{l,k}$ are evaluated:

$$\begin{aligned} \gamma_k &= \arg \max_{\gamma \in \Gamma} |\langle R^{l,k-1}, \phi_\gamma \rangle| \\ R^{l,k} &= R^{l,k-1} - \langle R^{l,k-1}, \phi_{\gamma_k} \rangle \phi_{\gamma_k} \\ s^{l,k} &= s^{l,k-1} + \langle R^{l,k-1}, \phi_{\gamma_k} \rangle \phi_{\gamma_k} \end{aligned}$$

The algorithm stops when residuals are enough small. This algorithm does not guarantee the orthogonality of the decomposition atoms.

2.6 Linear regression analysis and an optimal number of basis functions

LDA and an adapted to a classification problem linear regression analysis (called “optimal scoring” OS) produce a set of discriminant directions that coincide up to scalars (Hastie, Tibshirani and Buja, 1994b). In OS class labels are transformed to continuous variables that are fitted by linear regression. Each class label j is replaced by a vector of scores $\theta^j = \{\theta_{1j}, \theta_{2j}, \dots, \theta_{Mj}\}$ and discriminant directions $w_m \in \mathcal{R}^p$, $m = 1, \dots, M$ are optimized to minimize the average squared residuals:

$$R = \frac{1}{N} \sum_{m=1}^M \sum_{i=1}^N (\theta_{mg_i} - x_i^T w_m)^2,$$

where g_i is a class label of sample i . In addition, it is assumed that scores have zero means, unit variances and are uncorrelated for N observations: $(Y\theta)^T(Y\theta)/N = I_M$, where $Y_{[N \times M]}$ is an indicator matrix with the element $y_{im} = 1$ if the i th observation falls in class m , and 0 otherwise.

Herein, we interpret the score variables as class probabilities, so that θ_{mg_i} is considered to be a posterior probability of class m given an observation x_i and the number of scores M is the same as the number of classes ($M = G$). We don't optimize on score values (i.e. θ is considered to be fixed $\theta_{mg_i} = \delta_{mg_i}$) and carry out a multivariate linear regression of the indicator matrix Y on the observations $x_i \in \mathcal{R}^p$: $\hat{Y} = XW$, where $X = \{x_1, x_2, \dots, x_N\}^T$. Each i th observation is classified by the class having the largest value of \hat{y}_{im} . We aware that this estimation is unrealistic since there is no guarantee that an estimate lies in $[0, 1]$; however, it enables us to choose easily a number of wavelet basis functions ϕ_i using the χ^2 test (the F-test may be used as well (Draper and Smith, 1998), though the χ^2 is more general and is used here).

One can see that χ^2 test on the number of wavelet basis functions is equivalent to the χ^2 on the number of new predictor variables \tilde{X} , which are obtained by projecting the vector of predictor variables x onto the wavelet basis functions ϕ_j . Indeed, we seek W in the specific form with columns w_m to be linear combinations of wavelet basis functions: $w_m = \sum_{j=1}^J a_{jm}\phi_j$, where ϕ_j are extracted using either best basis (BB) or matching pursuit (MP) algorithms on the set of $\{w_m\}$, and $J = p$ for BB and some suitable J for MP algorithm. In other words, $W = \Phi_{[p \times J]}A_{[J \times M]}$, where $A = [a_{jm}]$ is a matrix of wavelet coefficients for the discriminant directions. Therefore, we get a modified linear regression problem of Y on new predictor variables $\tilde{X}_{[N \times J]}$: $\hat{Y} = \tilde{X}A$, where $\tilde{X} = X\Phi$ (note, that $XW = X(\Phi A) = (X\Phi)A$) and the matrix A plays the same role as the matrix W earlier. Thus, we show that χ^2 test in the parameter space is equivalent to a feed-forward model selection (or to a stepwise regression for a linear model). The main difference, however, is the assumption that the significance order of the new predictor variables $\tilde{x}_1, \dots, \tilde{x}_p$ is given in advance by the significance order of wavelet basis functions ϕ_i .

A test checks if to increase a number of basis functions by one comparing iteratively on $j = 2, \dots, J$ a current null hypothesis:

$$H_0^{j-1} : a_{jm} = 0, m = 1 \dots M$$

against an alternative (that is considered as a current full model):

$$H_0^j : \text{arbitrary } a_{rm}, r \leq j.$$

The ratio of the maximum likelihood $2 \log \lambda = \frac{p_j}{p_j - 1}$ is approximately distributed as $\chi^2(r)$ with a degree of freedom $r = M$ (a number of constraints) (Silvey, 1975); the maximum likelihood p_j is the maximum likelihood of Y given a truncated set of variables \tilde{X} with only j first columns. The χ^2 test with significance level α stops the model growing if $\sum_{m=1}^M \frac{\Delta rse_m}{\sigma_m^2} \leq \Xi_r^{-1}(1 - \alpha)$, where Ξ_r^{-1} is an inverse function of the cumulative density function of $\chi^2(r)$ with r degrees of freedom, $\Delta rse_m = rse_m^j - rse_m^{j-1}$ is a difference of residual square errors for the regression variable $y_{(\cdot), m}$ with j and $j - 1$ basis functions in W representation and $\sigma_m = rse_m/M$ are estimated variances of an independent Gaussian noise in the full model $Y = XW + \epsilon$. In addition, in order to achieve robust results, we estimate rse_m using cross-validation technique while σ_m^2 and coefficients of

linear regression using jackknife technique (Efron and Bradley, 1982).

3 Robust estimation

Real data sets frequently contain outliers, which may lead to large classification errors. One of the intuitive ways to test robustness is to add outliers to the data and test the classifier performance versus data contamination. We show by simulation that our regularization approach leads to estimation that is less sensitive to outliers.

4 Results

Three classification problems are described and results for them are presented below in Sections 4.1–4.3.

4.1 Waveforms Classification

This is a three class recognition problem based on the three shifted triangular waveforms $h_i(t)$, $i = 1, 2, 3$ (Breiman, Friedman, Olshen and Stone, 1984):

$$h_1(t) = \max(6 - |t - 7|, 0)$$

$$h_2(t) = h_1(t - 4)$$

$$h_3(t) = h_1(t - 8)$$

Each class consists of a random convex combinations of two of these waveforms (u is a random variable uniformly distributed on $(0,1)$) sampled at the integers and with standard normal noise ϵ added:

$$x_1(j) = uh_1(j) + (1 - u)h_3(j) + \epsilon(j)$$

$$x_2(j) = uh_1(j) + (1 - u)h_2(j) + \epsilon(j)$$

$$x_3(j) = uh_3(j) + (1 - u)h_2(j) + \epsilon(j)$$

$$j = 1 \dots 32$$

The number of samples is taken to be dyadic for implementation of BB and MP algorithms and equals 32. The triangular waveforms and several examples of waveforms from three classes are presented in Figure 1.

Different experimental setting with 300 (500) observations for training and 500 (300) for testing were used by Hastie et al. 1993 (Buckheit and Donoho, 1995). We use 187 examples per class for training and 100 examples per class for testing and run all algorithms 25 times to get average results and standard deviations.

The LDA vectors (canonical variates) and their reconstructed versions obtained using BB and MP algorithms are presented in the top plots of Figure 2. Projections of the testing data onto these directions are presented in the bottom plots of Figure 2 and show that classes lie very close to the edges of the triangle with the vertices that are projections of the triangle waveforms

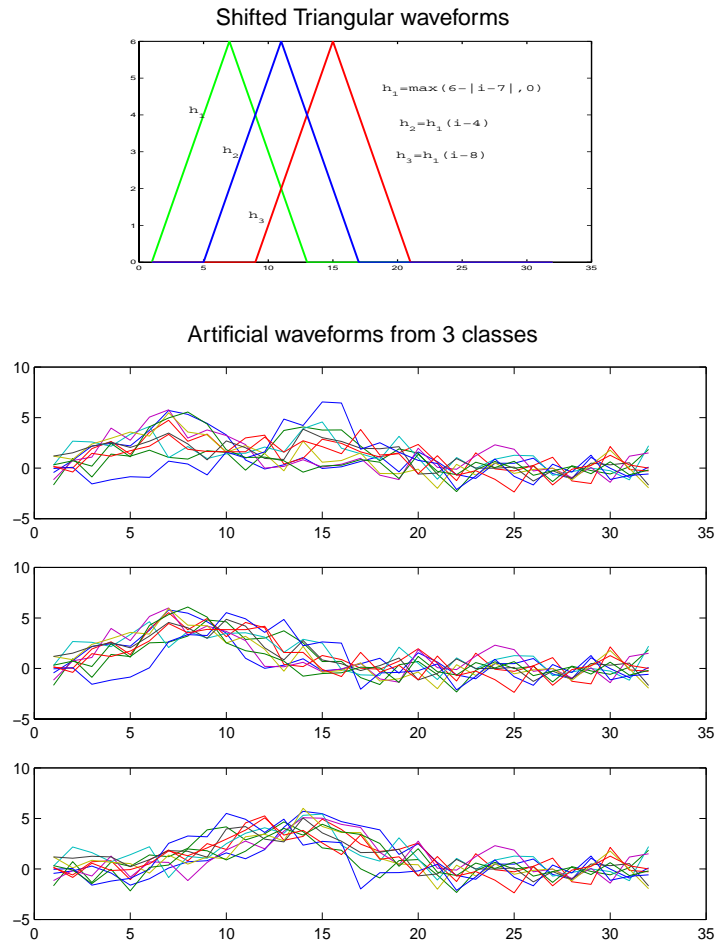


Figure 1: **Waveform data:** Top figure: Shifted triangular waveforms; Bottom figure: three classes of the waveform data.

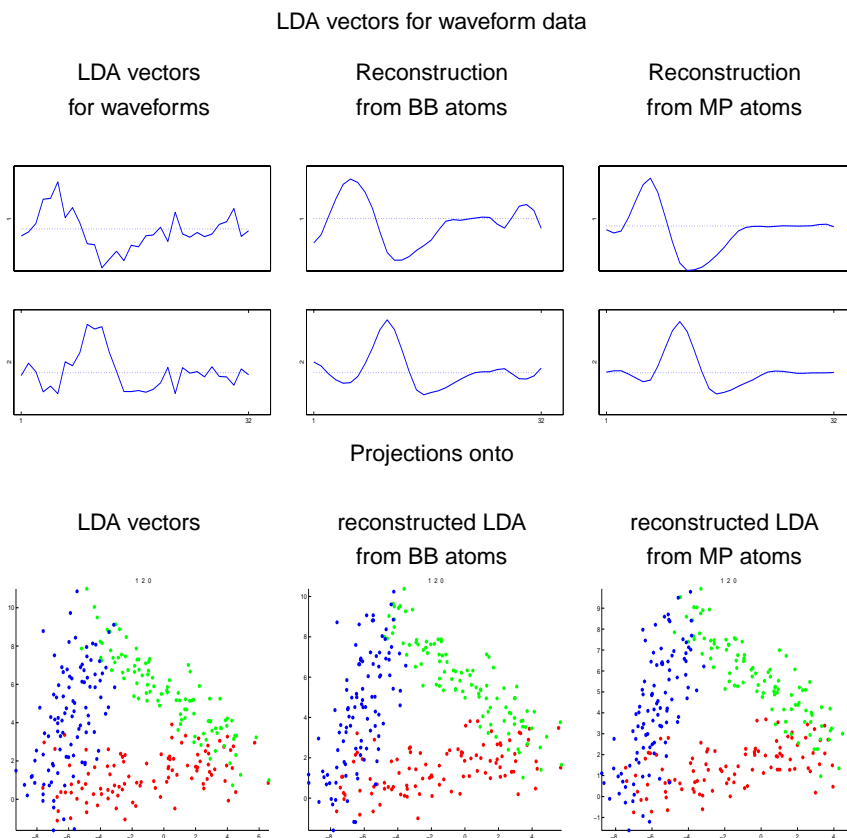


Figure 2: Top plot: Canonical variates and their reconstructions from a set of wavelet atoms extracted using BB and MP algorithms; Bottom plot: Data projected onto the space of canonical variates and their reconstructions

h_i onto the 2-D plain of canonical variates. Due to the symmetry in the data formation process, LDA algorithm projects the data to form an equally sided triangle. The classes are generally pairwise mixed near the vertices.

The same plots for contrast vectors are shown in Figure 3. We observe that the canonical variates are very noisy while contrast vectors are more robust. Using BB or MP algorithms we not just smooth discriminant directions but regularize their structure constraining them to lie in the space spanned by a small set of wavelet atoms (Figure 4).

We also see that BB algorithm is less sensitive than MP to a choice between canonical variates and contrast vectors. Moreover, MP extracts completely different wavelet atoms in these two cases. Compression via BB algorithm is more efficient for contrast vectors than for canonical variates (Figure 5).

Percent misclassification results averaged over 25 runs and their standard deviations are presented in Table 1. In most of the cases, results are slightly improved and the best result with

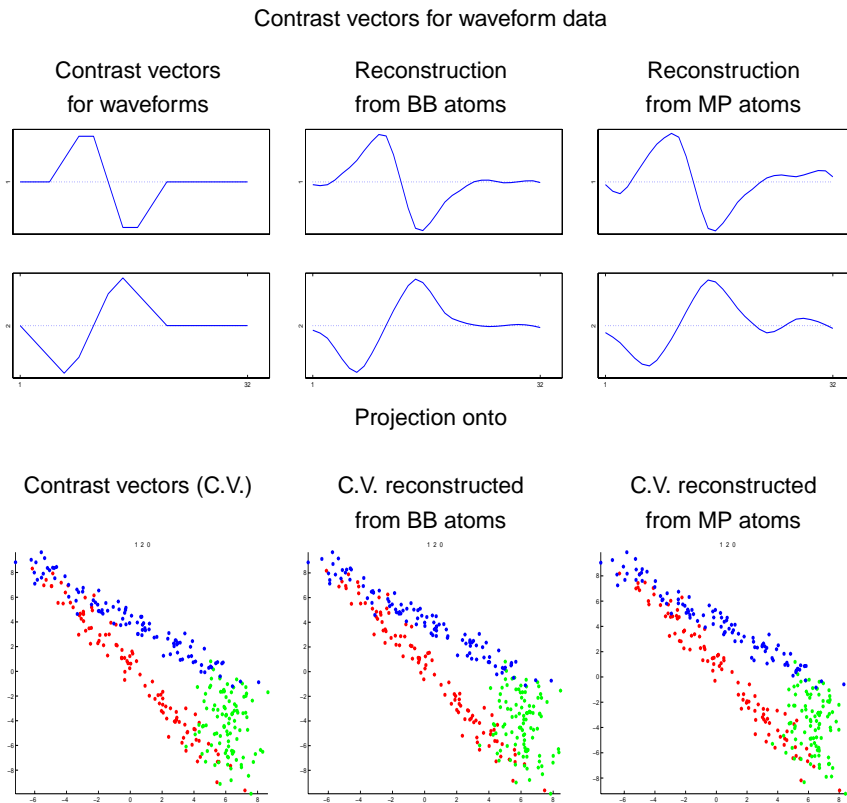


Figure 3: Top plot: Contrast vectors and their reconstructions from a set of wavelet atoms extracted using BB and MP algorithms; Bottom plot: Data projected onto the space of contrast vectors and their reconstructions

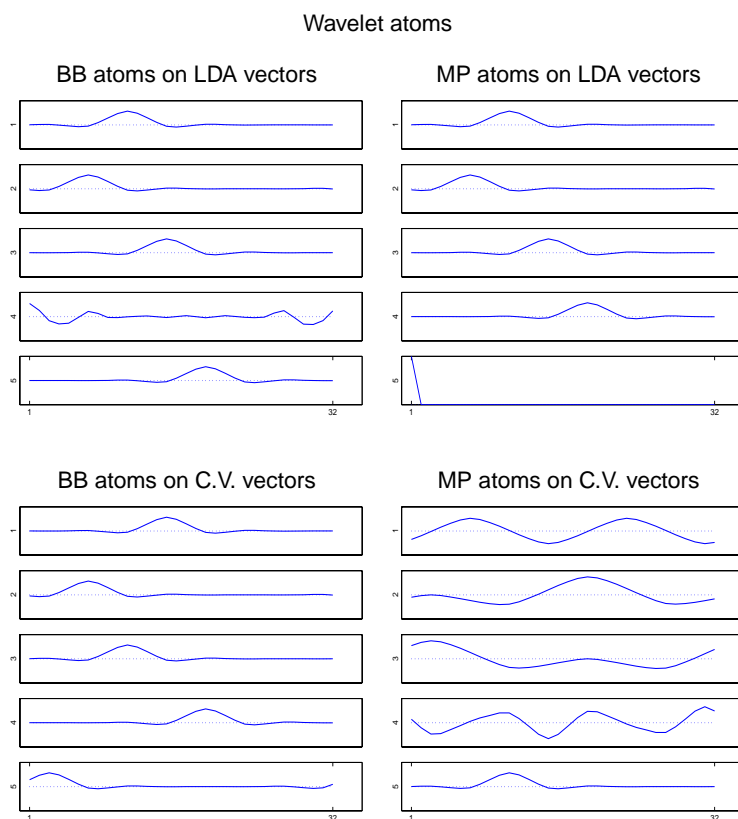


Figure 4: Five ordered wavelet atoms extracted using BB and MP algorithms on LDA and contrast vectors. Pay attention that MP is more sensible than BB to a choice of discriminant projections.

Compression of the discriminant directions

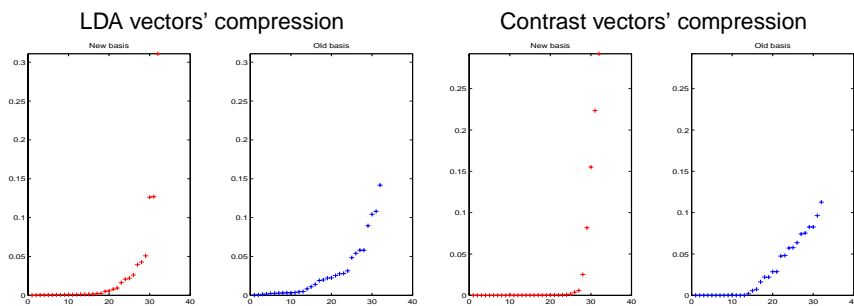


Figure 5: Plots of logarithmically scaled averaged signal energies in the best basis and in the original basis of signals extracted from LDA and contrast vectors. Compression in the case of contrast vectors is more efficient than in the case of canonical variates.

five wavelet atoms is achieved for LDA classification of the data projected on the wavelet atoms extracted using MP algorithm on contrast vectors. We emphasize that though the number of wavelet atoms (new projection directions) is greater than the number of canonical variates, their description is more compact.

Classification results for waveform data

| | | | | |
|-------------|-----------------------------------|------------|-------------------|-------------|
| Error types | Original data representation | | | |
| LDA test | 19.8 (3.0) | | | |
| PDA test | 19.1 (2.5) | | | |
| | LDA discriminative vectors | | | |
| | 5 basis functions | | | |
| Error types | BB-atoms | BB-reconst | MP | MP-reconst. |
| 1-nn test | 19.4 (2.4) | 18.9 (3.2) | 19.8 (2.7) | 19.5 (3.0) |
| LDA test | 19.0 (2.9) | 18.7 (3.0) | 19.3 (2.6) | 19.2 (3.0) |
| | Contrast vectors | | | |
| | 5 basis functions | | | |
| Error types | BB-atoms | BB-reconst | MP | MP-reconst. |
| 1-nn test | 19.4 (2.6) | 18.7 (2.1) | 19.8 (3.1) | 19.5 (2.6) |
| LDA test | 18.9 (3.0) | 24.4 (4.4) | 17.4 (2.2) | 25.6 (4.5) |

Table 1: Percent classification error and its std for waveform data. Experiments are run 25 times.

4.1.1 Selection of the number of basis functions

Here we present results of the χ^2 test described in Section 2.6. We use 165 samples per class in the training stage and 500 samples per class in the test. The cross-validation and jackknife groups are the same and their number is set to 11. We use best basis algorithm to extract wavelet basis and the number of basis wavelets are defined using χ^2 test. The results of the χ^2 test with significance levels of $\alpha = 0.1$, $\alpha = 0.05$ and $\alpha = 0.01$ averaged over 25 runs are presented in Table 2. When the α decreases the price for rejection the null hypothesis is so high that an “optimal” model with a smaller number of wavelet atoms “wins”. The cross-validation test for a misclassification error is presented in the last row of Table 2 and shows that both χ^2 test and cross validation test on misclassification error are compatible. Misclassification rate of the reduced models falls both on the training and testing sample sets and standard deviations remain about the same ones.

Discriminant directions of the full linear regression, regularized discriminant directions (i.e. coefficients of the reduced linear regression model) and extracted wavelet atoms are shown in Figure 6.

Averaged over 25 runs classification results for linear regression

| Models | Train error | Test error | N of wavelet atoms | | | | | |
|--|-------------|------------|--------------------|---|---|---|---|---|
| | | | 3 | 4 | 5 | 6 | 7 | 9 |
| Full model | 19.3 (1.5) | 18.8 (1.8) | * | * | * | * | * | * |
| Reduced model with $\alpha = 0.1$ | 16.1 (1.8) | 16.9 (1.9) | 4 | 7 | 7 | 6 | 0 | 1 |
| Reduced model with $\alpha = 0.05$ | 16.3 (1.6) | 16.6 (1.8) | 8 | 9 | 6 | 2 | 0 | 0 |
| Reduced model with $\alpha = 0.01$ | 16.4 (1.8) | 16.8 (1.7) | 14 | 8 | 2 | 1 | 0 | 0 |
| Reduced model with cross-validation | 15.9 (2.0) | 17.1 (1.4) | 10 | 6 | 4 | 3 | 2 | 0 |

Table 2: The number of basis functions is chosen by χ^2 significance test. The six last columns for reduced models show the frequency of number of atoms in 25 runs of χ^2 test. It is easy to see that when the α decreases the model growth is constrained and the "optimal" model has a small number of wavelet atoms. Misclassification train and test error are given in percents and standard deviation in percent are given in parenthesis. Nonrelevant information is marked with *.

4.1.2 Robustness issue

We artificially added outliers to a training sample set:

$$\begin{aligned}
 x_1(j) &= uh_1(j) + (1-u)h_3(j) + \epsilon(j) + Av_j\epsilon_{out}(j) \\
 x_2(j) &= uh_1(j) + (1-u)h_2(j) + \epsilon(j) + Av_j\epsilon_{out}(j) \\
 x_3(j) &= uh_3(j) + (1-u)h_2(j) + \epsilon(j) + Av_j\epsilon_{out}(j) \\
 j &= 1 \dots 32,
 \end{aligned}$$

where v_j are independent binomial random variables with probability of success $\pi_j = \pi/p$, π is a proportion of signals that are contaminated at least in along one component; $\epsilon_{out}(j)$ are distributed by independent normal distribution with a large standard deviation Σ and A is an amplification coefficient. Table 3 presents averaged over 25 runs classification results for a full linear regression model and our regularization method and for parameters $\pi = 0.5$ (50% of outliers), $A = 10$ and $\Sigma = 10$. The test sample set is taken both with and without outliers.

Theoretically, a breakdown point of the linear regression estimator is equal to $1/N$ (N is a size of the training sample set) (Draper and Smith, 1998). Indeed, even if one observation is replaced by an unbounded one an estimator is also carried over all bounds. Since our regularization method is a postprocessing of the linear regression its breakdown point is also expected to be $1/N$. Nevertheless, as show our experiments it is less sensitive to outliers.

4.2 Vowel Recognition

This example is a popular benchmark for neural network algorithms. The task is a speaker independent recognition of the eleven steady state vowels of British English. The data were

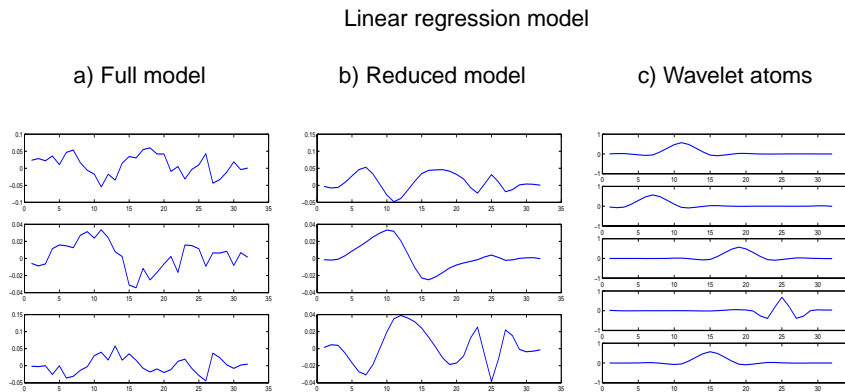


Figure 6: Discriminant directions found by a) a full linear regression and b) a reduced optimal linear regression model; c) five optimal wavelet atoms.

Averaged over 25 runs classification results for linear regression with outliers

| Models | Train error | Test error without outliers | Test error with outliers | N of wavelet atoms | | | | |
|------------------------------------|-------------|-----------------------------|--------------------------|--------------------|---|----|---|---|
| | | | | 3 | 4 | 5 | 6 | 7 |
| Full model | 22.8 (1.7) | 17.6 (1.3) | 21.7 (1.1) | * | * | * | * | * |
| Reduced model with $\alpha = 0.05$ | 20 (2) | 16.4 (1.3) | 19.9 (1.2) | 3 | 7 | 11 | 3 | 1 |

Table 3: The number of basis functions is chosen by χ^2 significance test. We consider two test misclassification errors for contaminated and “clean” test sample sets. It is a surprise, that the most frequent optimal number of wavelet atoms (5 basis atoms) coincides with a number of wavelet atoms that was chosen empirically in the experiments with LDA.

contributed by Anthony Robinson and obtained from the benchmark collection at CMU (White, 1997).

An ascii approximation to the International Phonetic Association symbol and the words in which the eleven sounds were recorded are presented in Table 4.

Each word is uttered 6 times by each of the 15 speakers. Four male and four female speakers were used to get 528 training samples, and the other four male and three female were used to get 462 testing samples. We repeated this procedure ten times to get average classification errors and standard deviations.

The signals were specially processed and transformed to 10-dimensional feature vectors (Hastie, Tibshirani and Buja, 1994a, see description). Since dyadic data are needed to use our method, we stretch the data to 16-dimensions by resampling data using linear approximation.

Classification results for vowel data averaged over 10 runs are presented in Table 5. The best classification result (39.8% misclassification error and 4.6% standard deviation) is achieved by the first nearest neighbor classifier applied to the data projected onto wavelet atoms extracted using BB algorithm on the contrast vectors. The best known to us result (averaged over 5 runs) regarding this data is obtained using classification by pairwise coupling (Hastie and Tibshirani,

| vowel | word | vowel | word |
|-------|------|-------|-------|
| i | heed | o | hod |
| I | hid | C: | hoard |
| E | head | U | hood |
| A | had | u: | who'd |
| a: | hard | 3 | heard |
| Y | hud | | |

Table 4: Vowel data: Words used in vowel recording

1996) and constitutes 47.3% misclassification error and 2.% standard deviation.

Vowel data classification results averaged over 10 runs

| Error types | Original data representation | | | | | |
|-------------|---------------------------------|-------------|-------------|----------------------------|-------------|-------------|
| | Signal dimensionality (10) | | | Signal dimensionality (16) | | |
| 1-nn test | 50.4 (5.3) | | | 50.1 (5.2) | | |
| LDA test | 52.2 (4.5) | | | 51.1 (4.6) | | |
| Error types | LDA discriminative vectors | | | | | |
| | 10 basis functions | | | 11 basis functions | | |
| | BB-atoms | BB-reconst. | MP-reconst. | BB-atoms | BB-reconst. | MP-reconst. |
| 1-nn test | 41.2 (4.4) | 49.9 (3.5) | 54.0 (4.1) | 40.6 (4.7) | 49.9 (4.6) | 53.2 (4.0) |
| LDA test | 51.2 (4.6) | 51.8 (4.8) | 57.4 (8.6) | 51.0 (4.8) | 51.1 (4.8) | 55.1 (9.0) |
| Error types | Contrast discriminative vectors | | | | | |
| | 10 basis functions | | | 11 basis functions | | |
| | BB-atoms | BB-reconst. | MP-reconst. | BB-atoms | BB-reconst. | MP-reconst. |
| 1-nn test | 40.5 (4.2) | 44.5 (7.2) | 44.9 (5.9) | 39.8 (4.6) | 43.4 (6.3) | 44.9 (5.9) |
| LDA test | 53.6 (5.7) | 56.7 (7.1) | 60.6 (8.9) | 51.0 (4.6) | 57.7 (13.4) | 60.6 (8.9) |

Table 5: Percent classification error avergaed over 10 runs.

4.3 Letter Recognition

This dataset is obtained from both DELVE (Rasmussen, Neal, Camp, Revow, Ghahramani, Kustra and Tibshirani, 1996a) and the UCI repository of machine learning database (Blake and Merz, 1998) and was originally submitted by D. Slate (Frey and Slate, 1991). The objective is to identify a large number of binary character images as one of the 26 capital letters in the English alphabet. The characters are presented in 20 different fonts and each letter within these 20 fonts is randomly distorted to produce a set of 20,000 samples. Thus, the data are, in fact, a conglomeration of 520-subclasses. Each image was converted into 16 primitive numerical attributes (statistical features of the pixel distribution) which were then scaled to fit into a range of integer values from 0 through 15. Two setting of experiments have been used for this data: (i) using the first 16,000 samples for training and remaining 4,000 for testing (Frey and Slate, 1991; Fogarty, 1992) (ii) proposed by DELVE environment (Rasmussen et al., 1996a). Results for the first experimental setting (i) are presented in Table 6. The best result obtained

Classification methods versus different data representations

| Error types | Original data representation | | | |
|-------------------|-----------------------------------|-------------|--------------------|-------------|
| | Signal dimensionality (16) | | | |
| (Frey & Slate 91) | 17.3 | | | |
| (Fogarty 92) | 4.3 (1.2) (averaged over 10 runs) | | | |
| 1-nn test | 5.1 | | | |
| LDA test | 31.5 | | | |
| Error types | LDA discriminant vectors | | | |
| | 10 basis functions | | 16 basis functions | |
| | BB-atoms | BB-reconst | BB-atoms | BB-reconst |
| 1-nn test | 19.2 | 17.2 | 4.2 | 5.1 |
| LDA test | 46.7 | 45.1 | 31.5 | 31.5 |
| Error types | Contrast discriminant vectors | | | |
| | 10 basis functions | | 16 basis functions | |
| | BB-atoms | BB-reconst. | BB-atoms | BB-reconst. |
| 1-nn test | 3.5 | 6.0 | 4.3 | 7.3 |
| LDA test | 35.8 | 34.9 | 31.5 | 38.4 |

Table 6: In the experiments with LDA discriminant vectors, the best result corresponds to the number of basis functions $m=16$. Therefore, in this case, compression is harmful, but new data representation leads to slightly better classification results. The best result is achieved for contrast discriminant weights with the number of basis functions $m=11$, i.e. compression is useful. The smallest test misclassification error is only 3.5% for one nearest neighbour rule (1-nn) and wavelet basis extracted by BB algorithm.

by Frey and Slate is 17.3% (the first line). This result was substantially improved using the first nearest neighbor (1-nn) classification rule (Fogarty, 1992). The average performance of the 1-nn classification rule over 10 runs for a random splitting to 16,000 train and 4,000 test samples is 4.3% with a standard deviation of 1.2%. The 1-nn rule classification for the same partition as in (Frey and Slate, 1991) gives a misclassification error 5.1%. The LDA method leads to a high misclassification error – (31.5%) on the test set. The LDA deficiency may be explained by a very large number of classes which themselves are mixtures of font sub-classes. We treat this data as signals to use our method. Our best results for 1-nn in the space of wavelet atoms extracted using BB algorithm on LDA vectors is 4.2% and on contrast vectors is 3.5%. For letter data the number of LDA vectors (16) is less than the number of contrast vectors (25), since the number of the classes exceeds the dimensionality of the input space.

We have also conducted and compared our results with the already available in DELVE environment³. There are 3 groups of experiments in DELVE with the training data sizes increasing

³Eight methods were tested for this data, six of which are presented in Figure 7(a): 1) *cart-1* stands for a basic version of CART (Breiman et al., 1984); 2) *hme-el-1* is an hierarchical mixtures-of-experts trained with Bayesian methods (ensemble learning) 3) *hme-ese-1* is a committee of hierarchical mixtures-of-experts trained by early stopping; 4) *hme-grow-1* is a committee of hierarchical mixtures-of-experts grown via early stopping; 5) *knn-class-1* is a K nearest neighbor algorithm for classification, with K chosen on the basis of leave-one-out cross-validation on the training set; 6) *me-el-1* is a committee of mixtures-of-experts trained by ensemble learning

by factor 2: 390, 780 and 1560 training samples; the size of testing data is 1777 samples. Six runs are done per each group, and data are split so that testing and training data are disjoint for each run and over different runs (Rasmussen, Neal, Camp, Revow, Ghahramani, Kustra and Tibshirani, 1996b, see Chapter 4). We use BB algorithm on LDA and contrast vectors and change a number of wavelet atoms, increasing them from 8 till 16; a classification rule is the nearest neighbor rule (1-nn). Comparison with DELVE results (Figure 7) shows that our methods are comparable with available in DELVE and are superior to them for (i) the wavelet atoms extracted using BB algorithm on contrast vectors for training data of size 780 and 1560, (ii) reconstructed LDA vectors from the large number of wavelet atoms extracted by BB algorithm.

5 Summary and Discussion

We proposed a novel method to regularize the structure of many parameter models by compressing projection directions via small linear combinations of wavelet packet atoms. The method was demonstrated on two discriminant directions: on LDA projections and contrast vectors. We compressed the set of discriminant parameters using best-basis algorithm. Classification was done from the data projected onto: (i) the small collection of wavelet atoms, (ii) the reconstructed projections. Results were given for three data-sets and showed that this approach leads to improved classification.

Our method is closely related to Minimum Description Length (MDL) principle (Rissanen, 1985). In a supervised learning, to communicate data labels to a receiver, a sender composes a message describing classification errors and model parameters (Nowlan and Hinton, 1992b). Optimization of the message description length leads to a tradeoff between these two parts of the message – a short description of the model parameters leads to an increase in the classification error. In our model, the number of wavelet atoms (taken from the wavelet packet) controls a description length of the model. An optimal number of the wavelet atoms depends on the “mother function” of the wavelet packet, that defines the structure of the model parameter space and is chosen heuristically. We are going to address a question of an on-line search of a set of wavelet atoms (from the wavelet packet) regularizing hidden weights of feed-forward Neural Networks in our future work.

Error misclassification bars and standard deviations for letter data

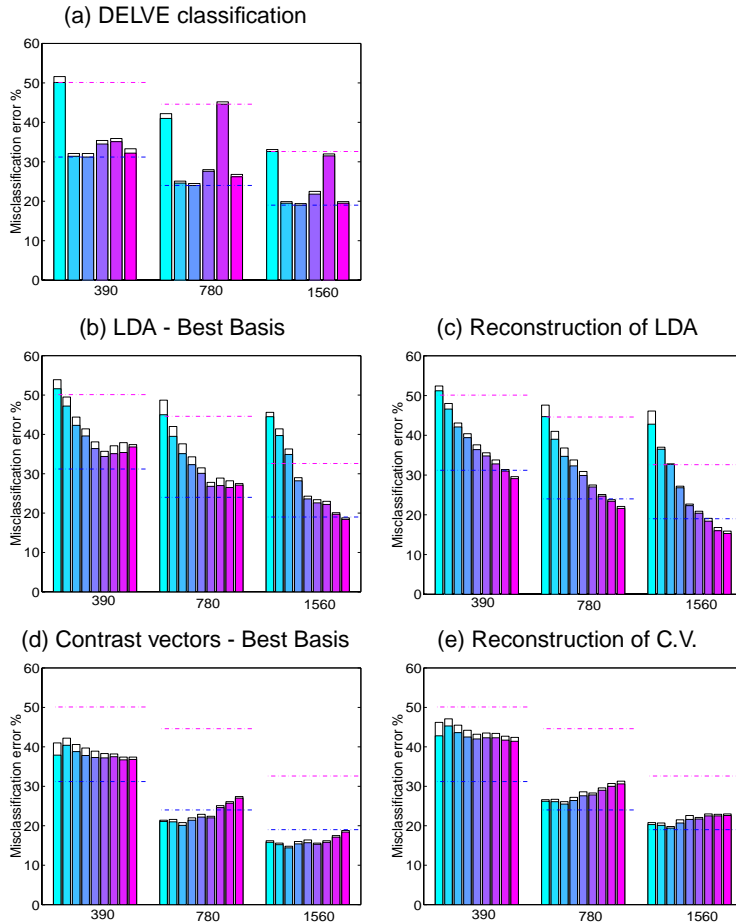


Figure 7: (a) - Classification results taken from DELVE: bars from left to right correspond to classification using *cart-1*, *hme-el-1*, *hme-ese-1*, *hme-grow-1*, *knn-class-1*, *me-el-1*; Other plots present classification results of the first nearest neighbor rule (1-nn) for data projected onto: b) - wavelet atoms extracted by BB algorithm on LDA vectors; c) - reconstructed from atoms (b) LDA vectors; d) - wavelet atoms extracted by BB algorithm on contrast vectors; e) - reconstructed from atoms (d) contrast vectors. The number of atoms in (b-e) increases from 8 to 16. In all plots (a-e), pink and blue horizontal lines show the worst and best classification results for DELVE classification schemes.

References

- Blake, C. L. and Merz, C. J. 1998. UCI repository of machine learning databases.
*<http://www.boltz.cs.cmu.edu/>
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. 1984. *Classification and Regression Trees*, The Wadsworth Statistics/Probability Series, Belmont, CA.
- Buckheit, J. and Donoho, D. L. 1995. Improved linear discrimination using time-frequency dictionaries, *Technical Report*, Stanford University.
*<http://stat.stanford.edu/reports/donoho/>
- Coifman, R. R. and Wickerhauser, M. 1992. Entropy-based algorithms for best basis selection, *IEEE Trans. Info. Theory* **38**(2): 713–719.
*<http://wuarchive.wustl.edu/doc/techreports/wustl.edu/math/papers/entbb.ps.Z>
- Coifman, R. and Wickerhauser, M. V. 1993. Wavelets and adapted waveform analysis. a toolkit for signal processing and numerical analysis, in I. Daubechies (ed.), *Different perspectives on wavelets, I*, American Mathematical Society, Providence, pp. 119–145.
- Draper, N. R. and Smith, H. 1998. *Applied Regression Analysis*, Willey series in probability and statistics, USA. third edition.
- Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*, John Wiley, New York.
- Efron and Bradley 1982. *The jackknife, bootstrap and other resampling plans*, Society for Industrial and Applied Mathematics, Philadelphia.
- Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**: 179–188.
- Fogarty, T. C. 1992. First nearest neighbour classification on frey and slate's letter recognition problem, *Machine Learning* **9**: 387–388.
- Frey, P. W. and Slate, D. J. 1991. Recognition using holland-style adaptive classifiers, *Machine Learning* **6**: 161–182.
- Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*, Academic press, London.
- Hastie, T. and Tibshirani, R. 1996. Classification by pairwise coupling, *Preprint* **10**: 1–38. Also a short version in NIPS 10, 1998.
*<http://www-stat.stanford.edu/hastie/Papers/>
- Hastie, T., Tibshirani, R. and Buja, A. 1994a. Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association* **89**: 1255–1270.
*<http://stat.stanford.edu/reports/hastie/fda.ps.Z>

- Hastie, T., Tibshirani, R. and Buja, A. 1994b. Penalized discriminant analysis, *Paper*, Department of Statistics, Stanford University.
*<http://stat.stanford.edu/reports/hastie/pda.ps.Z>
- Hinton, G. E. and Zemel, R. S. 1994. Autoencoders, minimum description length and helmholtz free energy, in J. Cowan, G. Tesauro and J. Alspector (eds), *Advances in Neural Information Processing Systems*, Vol. 6, MIT Press, Cambridge, MA, pp. 3–10.
- Intrator, N. 1993. Combining exploratory projection pursuit and projection pursuit regression with application to neural networks, *Neural Computation* 5(3): 443–455.
*<ftp://cns.brown.edu/nin/papers/epp-ppr.ps.Z>
- Intrator, N. 2000. Robust prediction in many parameter models: Specific control of variance and bias, in J. W. Kay and D. M. Titterton (eds), *Statistics and Neural Networks: Advances at the Interface*, Oxford University Press, pp. 97–128.
- Krogh, A. and Hertz, J. A. 1992. A simple weight decay can improve generalization, in J. Moody, S. Hanson and R. Lippmann (eds), *Advances in Neural Information Processing Systems*, Vol. 4, Morgan Kaufmann, San Mateo, CA, pp. 950–957.
- Mallat, S. and Zhang, Z. 1993. Matching pursuit in a time-frequency dictionary, *IEEE Transactions on Signal Processing* 41: 3397–3415.
- Nowlan, S. J. and Hinton, G. E. 1992a. Adaptive soft weight tying using gaussian mixtures, in J. Moody, S. Hanson and R. Lippmann (eds), *Neural Information Processing Systems 4*, Morgan Kaufmann, San Mateo, CA.
- Nowlan, S. J. and Hinton, G. E. 1992b. Simplifying neural networks by soft weight-sharing, *Neural Computation* 4: 473–493.
- Rasmussen, Neal, Camp, Revow, Ghahramani, Kustra and Tibshirani 1996a. Data for evaluating learning in valid experiments (delve).
*<http://www.cs.utoronto.ca/delve>
- Rasmussen, Neal, Camp, Revow, Ghahramani, Kustra and Tibshirani 1996b. The delve manual.
*<http://www.cs.utoronto.ca/delve>
- Rissanen, J. 1985. Minimum description length principle, *Encyclopedia of Statistical Sciences* pp. 523–527.
- Silvey, S. D. 1975. *Statistical Inference*, Chapman and Hall, New York.
- Stainvas, I. 1999. *Trade-off between recognition and reconstruction: Application of Neural Networks to Robotic Vision*, PhD thesis, Engineering Faculty, Tel-Aviv University.
*<http://www.math.tau.ac.il/stainvas>
- Wahba, G. 1990. *Splines Models for Observational Data*, Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

White, M. 1997. CMU repository of neural network benchmarks.
*<http://www.ics.uci.edu/~mlearn/MLRepository.html>