University of Tennessee, Knoxville

Trace: Tennessee Research and Creative Exchange

Doctoral Dissertations

Graduate School

5-2010

Kernel-Based Data Mining Approach with Variable Selection for Nonlinear High-Dimensional Data

Seung Hyun Baek

University of Tennessee - Knoxville, seunghyunbaek@hotmail.com

To the Graduate Council:

I am submitting herewith a dissertation written by Seung Hyun Baek entitled "Kernel-Based Data Mining Approach with Variable Selection for Nonlinear High-Dimensional Data." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial Engineering.

Alberto Garcia, Yuanshun Dai, Major Professor

We have read this dissertation and recommend its acceptance:

Xiaoyan Zhu, Hamparsum Bozdogan, Adam M. Taylor

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Seung Hyun Baek entitled “Kernel-Based Data Mining Approach with Variable Selection for Nonlinear High-Dimensional Data.” I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial Engineering.

Alberto Garcia
Co-Advisor

Yuanshun Dai
Co-Advisor

We have read this dissertation
and recommend its acceptance:

Xiaoyan Zhu

Hamparsum Bozdogan

Adam M. Taylor

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and
Dean of the Graduate School

(Original signatures are on file with official student records.)

Kernel-Based Data Mining Approach with Variable Selection for Nonlinear High-Dimensional Data

A Dissertation
Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Seung Hyun Baek
May 2010

Dedication

This dissertation is dedicated to my late father Tae Ho Baek (1945-2009), who passed away in a young age without seeing my success in finishing my degree. I also wish to dedicate this dissertation to my mother, Hee Suk Lim, who continually encouraged me to pursue my education and sacrificed many things for me so that one day I would become an honorable and educated citizen of this world. I have learned much about life from them. They have been my role-models for hard work, persistence and instilled in me the inspiration to set high goals and the confidence to achieve my future goals. I am dedicated to deliver their expectations in my life throughout my future.

ACKNOWLEDGEMENTS

I owe thanks to many people, whose assistance was indispensable in completing this dissertation. I wish to thank Dr. Alberto Garcia, and Dr. Yuanshun Dai, who served as my co-advisors and Ph.D. committee co-chairs, who helped me during the writing of my dissertation and provided support throughout in finishing my dissertation. I thank my mentor, Dr. Hamparsum Bozdogan for his cooperation and passion in providing knowledge and improving my research in the area of statistical modeling, especially model selection with his own development of information measure of complexity criterion. I thank Dr. Adam Taylor for his consistent help in providing NIR data sets. I thank to my sisters, Ji Yeon Baek, Su Yeon Baek, and Mi Yeon Baek for giving me warm-hearted support during the years of my graduate studies. I thank to my peers and friends in supporting me to complete my doctoral degree. Without such support and feedback process or loop, it would have been impossible to maintain and fulfill my long journey to get my doctoral degree.

ABSTRACT

In statistical data mining research, datasets often have nonlinearity and high-dimensionality. It has become difficult to analyze such datasets in a comprehensive manner using traditional statistical methodologies. Kernel-based data mining is one of the most effective statistical methodologies to investigate a variety of problems in areas including pattern recognition, machine learning, bioinformatics, chemometrics, and statistics. In particular, statistically-sophisticated procedures that emphasize the reliability of results and computational efficiency are required for the analysis of high-dimensional data.

In this dissertation, first, a novel wrapper method called SVM-ICOMP$_{\text{PERF}}$-RFE based on hybridized support vector machine (SVM) and recursive feature elimination (RFE) with information-theoretic measure of complexity (ICOMP) is introduced and developed to classify high-dimensional data sets and to carry out subset selection of the variables in the original data space for finding the best for discriminating between groups. Recursive feature elimination (RFE) ranks variables based on the information-theoretic measure of complexity (ICOMP) criterion.

Second, a dual variables functional support vector machine approach is proposed. The proposed approach uses both the first and second derivatives of the degradation profiles. The modified floating search algorithm for the repeated variable selection, with newly-added degradation path points, is presented to find a few good variables while reducing the computation time for on-line implementation.

Third, a two-stage scheme for the classification of near infrared (NIR) spectral data is proposed. In the first stage, the proposed multi-scale vertical energy thresholding (MSVET) procedure is used to reduce the dimension of the high-dimensional spectral data. In the second stage, a few important wavelet coefficients are selected using the proposed SVM gradient-recursive feature elimination (RFE).

Fourth, a novel methodology based on a human decision making process for discriminant analysis called PDCM is proposed. The proposed methodology consists of three basic steps emulating the thinking process: perception, decision, and cognition. In these steps two concepts known as support vector machines for classification and information complexity are integrated to evaluate learning models.

Table of Contents

Chapter Page

List of Tables

Table Page

List of Figures

Figure Page

Chapter 1 Introduction

This chapter provides an introduction of this dissertation research. Section 1.1 presents the motivation for the research. The contributions of the research are presented in Section 1.2. The organization of the rest of this dissertation is outlined in Section 1.3.

1.1 Motivation

Machine learning plays an important role in a variety of scientific fields including text mining, machine vision, pattern recognition, medical diagnosis, bioinformatics, and chemometrics. Practical problems arising in these fields require an approach built on innovative analytical methods. Two particularly important problems are (i) the presence of nonlinearities in available data; and (ii) the high-dimensionality of available data. In order to overcome these problems, kernel-based methods have been developed by several machine learning researchers. These methods are an effective alternative to increase computational power by first nonlinearly mapping the data into a high-dimensional space to avoid nonlinearities and then applying learning machines (modeling procedures). The objective of this dissertation is to develop innovative and effective analytical methods to increase computational power and improve scalability of complex data structures by (i) nonlinearly mapping the data into a high-dimensional space avoiding nonlinearities; and (ii) selecting the most relevant and informative variables.

Kernel-based methods exploit both the geometric and regularizing properties of a high-dimensional reproducing kernel Hilbert space. Since the early 1990s, kernel-based methods have been built in several developments, including (a) support vector machine for both classification and regression (Boser *et al.* 1992; Vapnik 1995); (b) kernel principal component analysis (Schölkopf *et al.* 1999); and (c) kernel fisher discriminant analysis (Mika *et al.* 1999). Perhaps the best-known kernel-based method is the support vector machine, which has been successfully applied in a diverse range of domains. Several recent publications describe the application of kernel-based methods and address their overall performance in terms of computational requirements and ability, for both classification and regression (Cristianini and Shawe-Taylor 2000, Herbrich 2002, Schölkopf and Smola 2002, Vapnik 1995). The properties of a support vector machine are (i) managing large input spaces powerfully with kernel-based methods; (ii) dealing with noisy samples in a robust way; and (iii) producing sparse solutions (Chistianini and Shawe-Taylor 2000). Support vector machine can be incorporated with the scheme of the kernel-based methods. The kernel-based methods are based on mapping data from the original input space to a kernel space with high-dimensionality and then solving the problem in that space which is nonlinearly related to the input space. A kernel is a function *K*, such that for all $\mathbf{x}, \mathbf{y} \in X$ satisfies that $K_{\Phi}(\mathbf{x}, \mathbf{y}) = <\Phi(\mathbf{x}), \Phi(\mathbf{y})>$, where Φ is a mapping from X to an inner product feature space F . The purpose of using the kernel function is as follows: (i) it provides the connection between the data and the modeling method; (ii) it can influence the performance of the modeling method by incorporating prior knowledge about the problem domain; and (iii) its evaluation might be computationally advantageous compared to an explicit construction of the feature space (Bloehdorn and Sure 2008).

Variable selection is an important area of research in machine learning, pattern recognition, statistics, and related fields. The key idea of variable selection is to find input variables which have predictive information and to eliminate non-informative variables. Variable selection identifies a small subset of variables so that the classifier constructed with the selected variables minimizes error and the selected variables also better explain the data (Koller and Sahami 1996). The use of variable selection techniques is motivated by three reasons: (i) to improve discrimination power; (ii) to find fast and cost-effective variables; and (iii) to reach a better understanding of the application process (Guyon and Elisseeff 2003). In the case of high-dimensionality data, variable selection plays a crucial role because of four challenges (Theodoridis and Koutroumbas 2006): (i) large sets of variables; (ii) existence of irrelevant variables; (iii) presence of redundant variables; and (iv) data noise.

1.2 Contributions of the Dissertation

Based on the motivations in Section 1.1, the contributions of this dissertation are as follows:

1. **Hybridized support vector machine and recursive feature elimination with information complexity:** An innovative approach is proposed by taking advantages from both the variable ranking method and the robust kernel-based method. This new approach is the hybridized support vector machine and recursive feature elimination with information complexity.

2. **Dual variable functional support vector machine:** Data representation for functional structures is one of the key issues in implementing the functional support vector machine. In some cases, a combination of derivatives with different orders may lead to better classification performance. The dual variables functional support vector machine approach that uses both first and second derivatives.
3. **Improved floating search method to optimize the number of variables:** Because dual or multiple data representations leads to a higher-dimension space, the modified floating search finds the optimal variables that have the highest classification power, so as to start with the best variable set in time series data.
4. **Multi-scale vertical energy thresholding wavelet method based on the scale information:** The multi-scale based wavelet transformation can extract useful information in compressed wavelet coefficients and thus can be used to perform noise suppression and pre-processing.
5. **Two-stage scheme for incorporating a wavelet de-noising and reduction method with a support vector machine-based variable selection method:** The use of the concentrated information with selected variables, instead of full variables, for the classification of high-dimensional data, can minimize classification error and improve computation speed significantly.
6. **Perception-decision-cognition methodology for discriminant analysis based on the human decision-making process:** The proposed methodology consists of three basic steps that emulate the thinking process: perception, decision, and cognition. In these steps two concepts known as the support vector machine and information complexity are integrated to evaluate learning models.

1.3 Outlines of the Dissertation

The remainder of this dissertation is organized as follows:

Chapter 2 shows a novel wrapper method based on hybridized support vector machine and recursive feature elimination with information complexity to classify nonlinear high-dimensional data sets and to carry out subset selection of the variables in the original data space.

In Chapter 3, a dual variable functional support vector machine and modified floating search based variable selection are presented. The different pre-processing techniques and the floating search method are explained.

Chapter 4 shows a two-stage classification procedure based on multi-scale vertical energy wavelet thresholding and support vector machine-based gradient recursive feature elimination. A wavelet-based data compression and de-noising technique and a support vector machine-based variable ranking algorithm are presented in detail.

In Chapter 5, a novel methodology based on the human decision-making process for discriminant analysis is presented. The proposed methodology consists of three basic steps emulating the thinking process: perception, decision, and cognition.

In Chapter 6, a summary and conclusions are presented.

Chapter 2 Hybridized Support Vector Machine and Recursive Feature Elimination with Information Complexity

2.1 Introduction

In many classification problems there are very high-dimensional input data sets and finding the best subset of the original input features or variables which mostly contribute to the separation of the classes or groups is a challenge. Therefore, variable selection is a difficult combinatorial problem in machine learning and it has very high practical importance in many applications.

Kernel-based methods have gained popularity for classification, clustering, and regression analysis in machine learning since the introduction of support vector machine (SVM) during the early 1990s. After obtaining support vectors (SVs) to classify a data set, questions such as: “*How do we know which variables are more responsible for, and important to, the classification?*” have often been raised. This is due to the fact that the mapping is not one-to-one and onto in SVM. The application of a kernel function is thus an uninvertible process, and there is no way to go from the feature space back to the original space. Because of this geometry, SVM does not lend itself to automated internal relevant variable selection easily. Hence algorithms for variable selection play an important role in SVM.

In the literature of machine learning, as discussed in Fröhlich (2002) in detail, there are two main approaches to solve the variable selection problem: (a) the filter approach, and (b) the wrapper approach. Both approaches differ in the way they evaluate a given variable subset. The filter method uses some relevance measure, which is independent of the performance of the learning algorithm. On the other hand, in the wrapper method, each variable subset is taken into consideration with the classifier. That is, the variables are evaluated by estimating the generalization performance (i.e. the expected risk) of the learning machine trained.

In this chapter, the wrapper method called SVM-ICOMP$_{\text{PERF}}$-RFE, which combines an information-theoretic measure of complexity (ICOMP) criterion and recursive feature elimination especially designed for SVM based variable selection developed by Guyon *et al*. (2002) is considered and emphasized. In the usual RFE, backward variable elimination is performed to find say, *m*, variables which lead to the largest margin of class separation. This combinatorial problem is solved in a greedy fashion. In the two-class case the RFE algorithm begins with the set of all variables and sequentially evaluates each variable based on sensitivity analysis for an appropriately defined criterion that is a measure of predictive ability (and is inversely proportional to the margin). Then, the RFE algorithm at each step eliminates the variable which keeps this quantity small. Assuming the change of the set of support vectors when removing only one variable is negligible.

An information-theoretic measure of complexity (ICOMP) criterion of Bozdogan (1988a, 1988b, 1990, 1994, 2000) is used in RFE rankings of the variables as an effective measure. ICOMP plays an important role not only in choosing an optimal kernel function from a portfolio of many other kernel functions but also in selecting important subset(s) of variables. It takes into account either the badness of fit or the lack of fit and the model complexity at the same time in one criterion function.

The potential and the flexibility of the proposed method is illustrated on two real data sets, one is ionosphere data which includes radar returns from the ionosphere, and another is aorta data which is used for the early detection of atheroma most commonly resulting heart attack. Also, the proposed method is compared with other RFE based methods (Guyon *et al*. 2002; Youn 2002; Cho *et al*. 2009) using different measures (i.e., weight and gradient) for variable rankings.

2.2 Support Vector Machine

The SVM finds the optimal separable hyperplane that maximizes the margin between the classes (Vapnik 1995). Consider the case of classifying a set of training data into two groups. Assume a set of training data is given by $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i$ is an input vector, $y_i \in (-1,1)$ is a binary class index, and *n* is the size of training data. Then, a decision boundary (i.e. classifier) that partitions the underlying vector space into two classes can be represented by the following hyperplane:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0, \tag{1}$$

where $\mathbf{w}$ is the weight vector and b is the bias. The objective of the SVM is to find the maximum margin(*M*) decision boundary between the two parallel hyperplanes, $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 1$ and $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = -1$. An example of SVM is illustrated in Figure 1. Since the maximum margin is given by $2/\|\mathbf{w}\|$, the corresponding optimization problem can be written as follows:

$$\begin{aligned}&\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{Subject to} \quad \begin{cases} y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \ge 1-\xi_i,\ i=1,2,\ldots,n \\ \xi_i \ge 0,\ i=1,2,\ldots,n \end{cases}\end{aligned} \tag{2}$$

where ξ_i is the positive slack variable and $C\,(>0)$ is a pre-designated regularization coefficient. The linearly-constrained optimization problem can be solved as a dual problem that maximizes the following function:

$$L(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right), \tag{3}$$

subject to the constraint

$$\begin{cases} \sum_{i=1}^{n}\alpha_i\mathbf{x}_i = 0,\ i=1,2,\cdots,n \\ 0 \le \alpha_i \le C,\ i=1,2,\cdots,n \end{cases}. \tag{4}$$

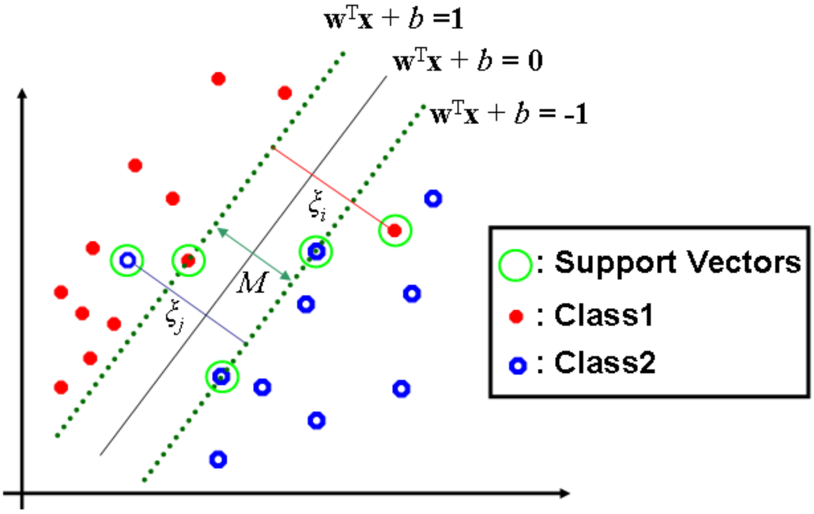

$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 1$
$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$
$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = -1$
ξ_i
M
ξ_j
○: Support Vectors
● : Class1
o : Class2

Figure 1: Illustration of Linear SVM for Nonlinearly Separable Case

Once the optimum values $\left(\alpha^*, b^*\right)$ are obtained, based upon the training set of points, a new point $\mathbf{x}_{new}$ of the test data set is classified by the following decision rule:

$$\begin{cases} \text{Class 1} \quad \text{if } D\left(\mathbf{x}_{new}\right) = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_{new}) + b^* < 0 \\ \text{Class 2} \quad \text{if } D\left(\mathbf{x}_{new}\right) = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_{new}) + b^* > 0 \end{cases} \tag{5}$$

where $D(\bullet)$ is a classifier based upon the training data set. $K(\mathbf{x}_i, \mathbf{x}_{new})$ is the kernel trick proposed by Aizerman *et al*. (1964). The kernel maps input data in the original space with nonlinearly into a high-dimensional feature space with linearity. The Table 1 presents some common kernel functions.

2.3 Information-Theoretic Measure of Complexity

An information-theoretic measure of complexity called ICOMP has been proposed by Bozdogan (1988a, 1988b, 1990, 2000) as a decision rule for model selection

Table 1: Kernel Functions

Function	*K*(**X**,**Y**)	Parameters
Linear	$(\mathbf{X}^{\mathrm{T}}\mathbf{Y}+b)^a$	*a*=1, *b*=0
Polynomial (degree=2)	$(\mathbf{X}^{\mathrm{T}}\mathbf{Y}+b)^a$	*a*=2, *b*=1
Polynomial (degree=3)	$(\mathbf{X}^{\mathrm{T}}\mathbf{Y}+b)^a$	*a*=3, *b*=1
Gaussian	$\exp(-(\frac{1}{a^b}\Vert\mathbf{X}-\mathbf{Y}\Vert^2)^c)$	*a*=2, *b*=*c*=1
Cauchy	$(1+\frac{1}{a}\Vert\mathbf{X}-\mathbf{Y}\Vert^2)^{-1}$	*a*=1
Inverse Multi-Quadratic	$(\Vert\mathbf{X}-\mathbf{Y}\Vert^2+a^2)^{-1/2}$	*a*=1

such as AIC (Akaike, 1973), and BIC (Schwarz, 1978). The development and construction of ICOMP is based on a generalization of the covariance complexity index originally introduced by van Emden (1971). Instead of penalizing the number of free parameters directly, ICOMP penalizes the covariance complexity of the model. It is defined by

$$ICOMP = -2\log L(\hat{\theta}_k) + 2C(\hat{\mathbf{\Sigma}}_{Model}), \tag{6}$$

where $L(\hat{\theta}_k)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector θ_k under the model M_k, and *C* represents a real-valued complexity measure and $\widehat{Cov}(\hat{\theta}_k) = \hat{\mathbf{\Sigma}}_{Model}$ represents the estimated covariance matrix of the parameter vector of the model. ICOMP should not be confused with the stochastic complexity (SC) or the minimum description length (MDL) of Rissanen (1986, 1987, 1989), although they both use the notion of complexity of a model class based on coding theory. The detailed information-theoretic measure of complexity (ICOMP) is recapitulated in the subsections for the benefit of the readers who may not be familiar with ICOMP criterion.

2.3.1 Mutual Information in High Dimensions

For a random vector, the complexity is defined as follows.

Definition: The complexity of a random vector is a measure of the interdependency among its components.

A continuous *p*-variate distribution is used with joint density function $f(\mathbf{x}) = f(x_1,...,x_p)$ and marginal density functions $f_j(x_j), j = 1,..., p.$ Following

Kullback (1997), and Harris (1978), the *information measure of dependence* is defined as follows:

$$\begin{aligned} I(\mathbf{x}) = I(x_1,...,x_p) &= E_f[\log\frac{f(x_1,...,x_p)}{f_1(x_1)\cdots f_p(x_p)}] \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_1,...,x_p)\log\frac{f(x_1,...,x_p)}{f_1(x_1)\cdots f_p(x_p)}dx_1\cdots dx_p \end{aligned} \qquad (7)$$

where $I(\mathbf{x})$ is the Kullback-Leibler information divergence (Kullback and Leibler 1951) against independence. The properties of the Kullback-Leibler information divergence are as follows:

- $I(\mathbf{x}) \equiv I(x_1,...,x_p) \geq 0$ i.e., the expected mutual information is nonnegative.
- $I(\mathbf{x}) \equiv I(x_1,...,x_p) = 0$ if and only if $f(x_1,...,x_p) = f_1(x_1)\cdots f_p(x_p)$ for every *p*-tuple $(x_1,...,x_p)$, i.e., if and only if the random variables $x_1,...,x_p$ are mutually statistically independent.

The KL divergence is related to Shannon's entropy (Shannon 1948) by the important identity

$$I(\mathbf{x}) \equiv I(x_1,...,x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1,...,x_p) \qquad (8)$$

where

- $H(x_j)$ is the marginal entropy, and
- $H(x_1,...,x_p)$ is the global or joint entropy

Watanabe (1985) calls this latter quantity the strength of structure and a measure of inter-dependence.

To define the information-theoretic measure of complexity of a multivariate distribution, let $f(\mathbf{x}) = f(x_1,...,x_p)$ be a multivariate Gaussian density function given by

$$\begin{aligned} f(\mathbf{x}) &= f(x_1,...,x_p) \\ &= (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}, \end{aligned} \tag{9}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)^{\mathrm{T}}, -\infty < \mu_j < \infty, j = 1, 2, ..., p$ and $\mathbf{\Sigma} > 0$ (positive definite)

As a short hand, let

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma}). \tag{10}$$

Then the joint entropy $H(\mathbf{x}) = H(x_1,...,x_p)$ from equation (8) for the case in which $\boldsymbol{\mu} = \mathbf{0}$ is given by

$$\begin{aligned} H(\mathbf{x}) = H(x_1,...,x_p) &= -\int_{\mathbf{R}^p} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbf{R}^p} f(\mathbf{x}) \left[\frac{p}{2}\log(2\pi)|\mathbf{\Sigma}| + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] d\mathbf{x} \\ &= \frac{p}{2}\log(2\pi)|\mathbf{\Sigma}| + \frac{1}{2} tr\left[\int_{\mathbf{R}^p} f(\mathbf{x})\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} d\mathbf{x}\right]. \end{aligned} \tag{11}$$

Then, since $E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}] = \mathbf{\Sigma},$ the joint entropy is

$$\begin{aligned} H(\mathbf{x}) = H(x_1,...,x_p) &= \frac{p}{2}\log(2\pi) + \frac{p}{2} + \frac{1}{2}\log|\mathbf{\Sigma}| \\ &= \frac{p}{2}[\log(2\pi)+1] + \frac{1}{2}\log|\mathbf{\Sigma}|. \end{aligned} \tag{12}$$

From equation (11), the marginal entropy $H(x_j)$ is

$$\begin{aligned} H(x_j) &= -\int_{-\infty}^{+\infty} f(x_j)\log f(x_j) dx_j \\ &= \frac{1}{2}\log(2\pi) + \frac{1}{2} + \frac{1}{2}\log(\sigma_j^2), j = 1, 2, ..., p, \end{aligned} \tag{13}$$

where σ_j^2 is the variance of the jth variable.

2.3.2 Initial Definition of Covariance Complexity

van Emden (1971, p. 61) provides a reasonable initial definition of complexity of a covariance matrix $\mathbf{\Sigma}$ for the multivariate Gaussian distribution. This measure is given by:

$$\begin{aligned} I(x_1,\ldots,x_p) &\equiv C_0(\mathbf{\Sigma}) = \sum_{j=1}^{p} H(x_j) - H(x_1,\ldots,x_p) \\ &= \sum_{j=1}^{p}\left[\frac{1}{2}\log(2\pi)+\frac{1}{2}\log(\sigma_{jj})+\frac{1}{2}\right]-\frac{p}{2}\log(2\pi)-\frac{1}{2}\log|\mathbf{\Sigma}|-\frac{p}{2}. \end{aligned} \tag{14}$$

This reduces to

$$C_0(\mathbf{\Sigma}) = \frac{1}{2}\sum_{j=1}^{p}\log(\sigma_{jj}) - \frac{1}{2}\log|\mathbf{\Sigma}|, \tag{15}$$

where $\sigma_{jj} \equiv \sigma_j^2$, is the variance of the *j*th variable, and is the *j*th diagonal element of $\mathbf{\Sigma}$. The characteristics of covariance complexity C_0 are as follows:

- $C_0(\mathbf{\Sigma}) = 0$ if and only if $\mathbf{\Sigma}$ is a diagonal matrix.
- $C_0(\mathbf{\Sigma}) = \infty$ if and only if $|\mathbf{\Sigma}| = 0$.
- The first term of equation (15) is not invariant under orthonormal transformations.

As pointed out by van Emden (1971), the result in equation (15) is not an effective measure of the amount of complexity in the covariance matrix $\mathbf{\Sigma}$, since:

- $C_0(\mathbf{\Sigma})$ depends on the coordinates of the original random variables $x_1,\ldots,x_p$.
- The first term of $C_0(\mathbf{\Sigma})$ in equation (15) would change under orthonormal transformations.

2.3.3 Definition of Maximal Covariance Complexity

To improve upon $C_0(\mathbf{\Sigma})$ in equation (15), a maximal covariance complexity is proposed as follows.

Proposition: A maximal information theoretic measure of complexity of a covariance matrix $\mathbf{\Sigma}$ of a multivariate Gaussian distribution is defined as follows:

$$\begin{aligned} C_1(\mathbf{\Sigma}) &= \max_T C_0(\mathbf{\Sigma}) = \max_T \{H(x_1)+\cdots+H(x_p)-H(x_1,\ldots,x_p)\} \\ &= \frac{p}{2}\log\left[\frac{tr(\mathbf{\Sigma})}{p}\right]-\frac{1}{2}\log|\mathbf{\Sigma}| \qquad (16) \\ &= \frac{p}{2}\log\frac{\overline{\lambda}_a}{\overline{\lambda}_g}, \end{aligned}$$

where the maximum is taken over the orthonormal similarity transformation, T of the overall coordinate systems $x_1,\ldots,x_p$ and $\overline{\lambda}_a$ and $\overline{\lambda}_g$ are arithmetic and geometric means of the eigenvalues. The properties of maximal information-theoretic measure of complexity are as follows:

- $C_1(\mathbf{\Sigma})$ is the log ratio between the arithmetic and geometric mean of the eigenvalues.
- $C_1(\mathbf{\Sigma})$ incorporates the two most basic scalar measures of multivariate scatter-trace and determinant.
- $C_1(\mathbf{\Sigma}) \to 0$ as $\mathbf{\Sigma} \to \mathbf{I}_p$.
- As interaction between variables increases, so does $C_1(\mathbf{\Sigma})$.

2.3.4 Modified Maximal Covariance Complexity

Following van Emden (1971), the geometric definition of covariance complexity is defined by the Frobenius norm given by

$$C_F(\mathbf{\Sigma}) = \frac{1}{s} \| \mathbf{\Sigma} \|^2 - \left(\frac{tr(\mathbf{\Sigma})}{s} \right)^2, \tag{17}$$

where $\| \mathbf{\Sigma} \|^2 = tr(\mathbf{\Sigma}^{\mathrm{T}} \mathbf{\Sigma})$, the square of the Frobenius norm of $\mathbf{\Sigma}$.

In terms of the eigenvalues (or singular values), $C_F(\mathbf{\Sigma})$ reduces to

$$C_F(\mathbf{\Sigma}) = \frac{1}{s} \sum_{j=1}^{s} (\lambda_j - \overline{\lambda}_a)^2, \tag{18}$$

where s is the rank of $\mathbf{\Sigma}$, λ_j is the j^{th} eigenvalue of $\mathbf{\Sigma} > 0$, j = 1,2,. . .,s and $\overline{\lambda}_a$ is arithmetic mean of the eigenvalues. Note that $C_F(\mathbf{\Sigma}) \geq 0$ with $C_F(\mathbf{\Sigma}) = 0$ only when all $\lambda_j = \overline{\lambda}_a$.

$C_1(\mathbf{\Sigma})$ can be approximated in terms of the eigenvalues λ_j, $j = 1, 2, \ldots, s$ by

$$C_1(\mathbf{\Sigma}) \cong \frac{1}{4} \sum_{j=1}^{s} \left(\frac{\lambda_j - \overline{\lambda}_a}{\overline{\lambda}_a}\right)^2. \tag{19}$$

Since in the feature space orthonormal matrices are dealt with to prevent the C_1 complexity not to go to zero, C_1 and C_F are related as a second order equivalent measure of complexity denoted by C_{1F} . Hence, the modified maximal entropic complexity $C_{1F}(\mathbf{\Sigma})$ is defined as follows:

$$C_{1F}(\mathbf{\Sigma}) = \frac{s}{4}\frac{C_F(\mathbf{\Sigma})}{\left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2} = \frac{s}{4}\frac{\frac{1}{s}\|\mathbf{\Sigma}\|^2 - \left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2}{\left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2}. \tag{20}$$

In terms of the eigenvalues, $C_{1F}(\mathbf{\Sigma})$ is given by

$$\begin{aligned} C_{1F}(\mathbf{\Sigma}) &= \frac{s}{4}\frac{\frac{1}{s}tr(\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}) - \left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2}{\left(\frac{tr(\mathbf{\Sigma})}{s}\right)^2} \\ &= \frac{s}{4}\frac{1}{s\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2 \\ &= \frac{1}{4\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2. \end{aligned} \tag{21}$$

where $s = rank(\mathbf{\Sigma})$. The properties of the modified maximal entropic complexity C_{1F} are as follows:

- $C_{1F}(\mathbf{\Sigma})$ is scale-invariant, and $C_{1F}(\mathbf{\Sigma}) \geq 0$ with $C_{1F}(\mathbf{\Sigma}) = 0$ only when all $\lambda_j = \overline{\lambda}_a$.
- $C_{1F}(\mathbf{\Sigma})$ measures the relative variation in the eigenvalues rather than absolute variation of the eigenvalues.

2.3.5 ICOMP as a Performance Measure: ICOMP$_{\mathbf{PERF}}$

Singularity of the estimated *covariance matrix* is a common problem that has recently attracted many researchers' work. Because of this, many methods have been proposed to make the covariance matrix *well-conditioned*, so that the covariance matrix can be estimated. The usual response to *singular* or *ill-conditioned* covariance matrix estimates is the *“naive” ridge regularization,* $\hat{\mathbf{\Sigma}}^* = [\hat{\mathbf{\Sigma}} + \alpha \mathbf{I}_p]$, which works to counteract the ill-conditioning by adjusting the eigenvalues of $\hat{\mathbf{\Sigma}}$. The ridge parameter, α, is typically chosen to be very small. This, of course, begs the questions

- *How large of a perturbation do we need?*
- *How small a perturbation can we get away with*?

This is a case where simplicity is not necessarily a good thing; it does not solve the problem with many real datasets. Yet another approach that does not seem to work well in practice is to augment $\hat{\mathbf{\Sigma}}$ with a multiple of the *kernel matrix,* as suggested by Mika (2002). After much experimentation with a variety of different methods to improve the condition of the covariance matrix, a stabilization method (Thomaz 2004) is applied to resolve the *ill-conditioning* of a covariance matrix. After the stabilization procedure, the two-*stage stabilization* and *smoothing* process is applied to provide a *well-conditioned* covariance matrix which is both nonsingular and positive definite.

- Stage 1. Stabilization algorithm (Thomaz 2004):

1. Perform spectral decomposition of $\hat{\mathbf{\Sigma}} = \mathbf{V\Lambda V}^{\mathrm{T}}$, where $\mathbf{V}$ is the matrix with eigenvectors and $\mathbf{\Lambda}$ has eigenvalues on the diagonal.

2. Calculate the mean eigenvalue $\bar{\lambda} = (\sum_{\mathrm{i=1}}^{\mathrm{p}} \lambda_i) / p$

3. Form a new matrix of eigenvalues as

$$\mathbf{\Lambda}^* = \begin{bmatrix} \max(\lambda_1, \bar{\lambda}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \max(\lambda_p, \bar{\lambda}) \end{bmatrix}$$

4. Finally, recompose the new stabilized matrix

$$\hat{\mathbf{\Sigma}}_{STA} = \mathbf{V}\mathbf{\Lambda}^{*}\mathbf{V}^{\mathrm{T}}$$

- Stage 2: Compute a Stabilized and Smoothed Convex Sum Covariance Estimator

The second step is to feed the *stabilized* covariance matrix into a *smoothed* convex sum covariance matrix estimator (CSE) was proposed based on the quadratic loss function used by Press (1975) and later by Chen (1976). The stabilized and smoothed convex sum covariance estimator (STA-CSE) is as follows:

$$\hat{\mathbf{\Sigma}}_{STA\_CSE} = \frac{n}{n+m}\hat{\mathbf{\Sigma}}_{STA} + (1-\frac{n}{n+m})\hat{\mathbf{D}}_{STA}, \tag{22}$$

where $\hat{\mathbf{D}}_{STA} = \left(\frac{1}{p}tr(\hat{\mathbf{\Sigma}}_{STA})\right)\mathbf{I}_p$. For $p \geq 2,$ m is chosen to be

$$0 < m < \frac{2[p(1+\beta)-2]}{p-\beta},$$

where

$$\beta = \frac{\left(tr(\hat{\mathbf{\Sigma}}_{STA})\right)^2}{tr(\hat{\mathbf{\Sigma}}_{STA}^2)}.$$

This estimator improves upon $\hat{\mathbf{\Sigma}}_{STA}$ by shrinking all the estimated eigenvalues of $\hat{\mathbf{\Sigma}}_{STA}$ toward their common mean. The motivation of using both stabilization and smoothing of the covariance matrix in the ranking process of RFE subset selection is to extract more information since a reduced rank problem occur in the kernel based methods. To remedy the current existing problems in the usual kernel methods, the use of both stabilization and smoothing the covariance matrix is an attractive approach.

The choice of the best mapping function is not so simple and automatic. In the literature a valid method for selecting the appropriate kernel function does not yet exist.

The goal of SVM is to minimize the probability of misclassification error. Intuitively, then, the penalty term for a poorly-fitting model would be based on the classification error rate. In SVM problems, the error variance σ^2 is estimated by the mean squared difference between actual group labels (y_i) and predicted group labels ($\hat{y}_i$) given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{23}$$

Now following the work of Howe and Bozdogan (2010) the information-theoretic measure of complexity as performance measure of SVM is defined as follows:

$$ICOMP_{PERF} = n\log 2\pi + n\log\hat{\sigma}^2 + n + 2C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE}), \tag{24}$$

where $\hat{\mathbf{\Sigma}}_{STA\_CSE}$ is the *stabilized and smoothed convex sum covariance matrix estimator* (STA-CSE) given by

$$\hat{\mathbf{\Sigma}}_{STA\_CSE} = \frac{n}{n+m}\hat{\mathbf{\Sigma}}_{STA} + (1-\frac{n}{n+m})\hat{\mathbf{D}}_{STA}, \hat{\mathbf{D}}_{STA} = \left(\frac{1}{p}tr(\hat{\mathbf{\Sigma}}_{STA})\right)\mathbf{I}_p,$$

and

$$C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE}) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a)^2.$$

First, the hybrid covariance estimate is calculated, and then the diagonal matrix of the largest singular values as a reduced rank approximation of $\hat{\mathbf{\Sigma}}_{STA\_CSE}$ is computed. By minimizing *ICOMP_{PERF}*, the classification error is minimized under the best fitting model. Also, *ICOMP_{PERF}* is used to choose an optimal kernel function. One of the major motivations of introducing the information measure of complexity (ICOMP) criterion is based on the fact that in SVM-RFE subset selection problems the number of variables is same from one subset to another. In such cases the models in terms of the number of parameters are considered to be equivalent. In equivalent models, AIC, BIC, or MDL type criteria do not have provision of distinguishing one equivalent model from another. Since their penalty terms are fixed, and not varying. In the literature cross-validation-based criteria has been used for variable selection. These types of criteria are too time-consuming due to the high-dimensionality of the feature space. The proposed method shortens the variable selection time.

2.4 Recursive Feature Elimination (RFE)

A variable selection method based on RFE has been developed by Guyon *et al*. (2002) which is called SVM-RFE. SVM-RFE is an application of a recursive feature elimination based on sensitivity analysis using an appropriately defined cost function (**w**: weight). The SVM-Gradient-RFE method (Youn 2002; Cho *et al*. 2009) used the gradient as a cost function. In the proposed method, the used cost function is the $ICOMP_{PERF}$. In the proposed method, the least sensitive variable, which has the minimum value of the $ICOMP_{PERF}$, is eliminated first. This eliminated variable becomes rank p (p: number of variables). Later, the machine is retrained on the remaining p-1 variables and then the variable with the minimum value of $ICOMP_{PERF}$ is eliminated. The process continuous in an iterative fashion until no variable is left in that subset. This means that at the end of this iterative ranking scheme all the variables are ranked according to $ICOMP_{PERF}$ criterion. This is different than the Guyon *et al*. (2002) ranking scheme where only weights have been considered without taking into account the model fit and the complexity of the model.

2.4.1 SVM-RFE Algorithm

Let $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_n)^{\mathrm{T}}$ be a training set with $\mathbf{y} = (y_1,...,y_n)^{\mathrm{T}}$.

1. Construct a training model $\mathbf{X} = \mathbf{X(:,s)}$, where **s** is the subset of variables; **s**=1,2,…,*p*.
2. Until all values of the cost function are obtained with the number of non-ranked variables, compute the cost function for all subsets

$$C(i) = (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}_{(-i)}\boldsymbol{\alpha}, \qquad (25)$$

   where $\mathbf{H} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{H}_{(-i)}$ means a **H** matrix without the i^{th} variable.
3. Find the variable *k* with the smallest cost function value, and add *k* into the ranked subset, **r** and remove *k* from subset, **s**.
4. Repeat 1-3 until subset, **s** is empty.

2.4.2 SVM-Gradient-RFE Algorithm

Let $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_n)^{\mathrm{T}}$ be a training set with $\mathbf{y} = (y_1,...,y_n)^{\mathrm{T}}$.

1. Construct a training model $\mathbf{X} = \mathbf{X(:,s)}$, **s** is the subset of variables; **s**=1,2,…,*p*.
2. Until all values of the average sum of the angles are obtained with the number of non-ranked variables,

   (i) compute the gradient, $\nabla_{(-i)}g(\mathbf{x})$ without i^{th} variable

$$\nabla_{(-i)}g(\mathbf{x}) = \sum_{m\in \mathrm{sv}} \alpha_m y_m \nabla_{(-i)} K(\mathbf{x}_m, \mathbf{x}). \qquad (26)$$

   (ii) compute the sum of angles between $\nabla_{(-i)}g(\mathbf{x})$ and $\mathbf{e}_m$, γ

$$\gamma(i) = \sum_{m\in \mathrm{sv}} \angle(\nabla_{(-i)} g(\mathbf{x}), \mathbf{e}_m), \qquad (27)$$

where (-*i*) means without the *i*th variable, $\mathbf{e}_m$ is unit vectors, and

$$\angle(\nabla_{(-i)} g(\mathbf{x}), \mathbf{e}_m) = \min_{\beta \in \{0,1\}} \left\{ \beta\pi + (-1)^{\beta} \arccos\left(\frac{\langle \nabla_{(-i)} g(\mathbf{x}) \bullet \mathbf{e}_m \rangle}{\| \nabla_{(-i)} g(\mathbf{x}) \|} \right) \right\}.$$

(iii) compute the average sum of the angles $A(i) = 1 - \frac{2}{\pi} \bullet \frac{\gamma(i)}{|\mathrm{SV}|}$.

3. Find the variable *k* with the smallest the average sum of the angle *A*(*i*), add *k* into the ranked subset, **r** and remove *k* from subset, **s**.
4. Repeat 1-3 until subset, **s** is empty.

2.4.3 Proposed SVM-*ICOMP$_{PERF}$*-RFE Algorithm

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^{\mathrm{T}}$ be a training set with $\mathbf{y} = (y_1, ..., y_n)^{\mathrm{T}}$.

1. Construct a training model $\mathbf{X} = \mathbf{X(:,s)}$, where **s** is the subset of variables; **s**=1,2,…,*p*.
2. Until all *ICOMP$_{PERF}$* values are obtained with the number of non-ranked variables, compute *ICOMP$_{PERF}$* based on the error rate obtained from SVM. The *ICOMP$_{PERF}$* is given by

$$ICOMP_{PERF}(i) = n \log 2\pi + n \log \hat{\sigma}^2_{(-i)} + n + 2C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE(-i)}), \qquad (28)$$

where $\hat{\sigma}^2_{(-i)}$ is the estimated error variance without the *i*th variable and $\hat{\mathbf{\Sigma}}_{STA\_CSE(-i)}$ is the stabilized and smoothed convex sum covariance matrix estimator without the *i*th variable in the model.

3. Find the variable *k* with the smallest *ICOMP$_{PERF}$*, add *k* into the ranked subset, **r** and remove *k* from subset, **s**.
4. Repeat 1-3 until subset, **s** is empty.

2.5 Numerical Results

In the data mining literature, data partitioning is an important issue for finding proper models for new datasets. In general one can use different data partitioning to get different results. Most of such data partitioning schemes do not take into account of randomness that may affect the performance of the results which can be different. In the analysis, to avoid partitioning dependency, the data is randomly partitioned into 20% as one set and 80% as another set based on Pareto's principle (Pareto 1909). Two experiments are performed with two different sets; 20%/80% and vice versa as training/test sets. The variable rankings corresponding to kernel functions are determined and reported for those different sets. Also, the smallest value of *ICOMP*$_{PERF}$, and the 95% confidence intervals (CIs) given by $\bar{X}_{error} \pm 1.96\hat{\sigma}_{error}$ for the training and test errors are reported. Ionosphere and aorta datasets are used for these experiments.

2.5.1 Ionosphere Data

The ionosphere data are radar data which was collected by a system in Goose Bay, Labrador (Sigillito *et al.* 1989). The system measures radar returns from the ionosphere. The data consist of 351 observations and 34 variables with binary classes; good and bad returns. Figure 2 shows the scatter plots of the data with groups identified by blue (circle) and red (cross) colors. As shown in Figure 2, the separation in dimension 5 against dimensions 13, 19 and dimensions 18, 29 are quite poor. Tables 2 and 3 show performances of experiments based on *ICOMP*$_{PERF}$. In Table 2, the polynomial kernel with degree 3 on the 20% set shows a narrower confidence interval than the other kernel functions for both training and test sets. As shown in Tables 2 and 3, the smallest $ICOMP_{PERF}$ values are obtained with a polynomial kernel with degree 3 for the 20% set and the 80% set. Tables 4 and 5 show the best subset selection based on the smallest $ICOMP_{PERF}$ values. The training and test errors of the best subsets in both partitioned sets are within the 95% error confidence intervals.

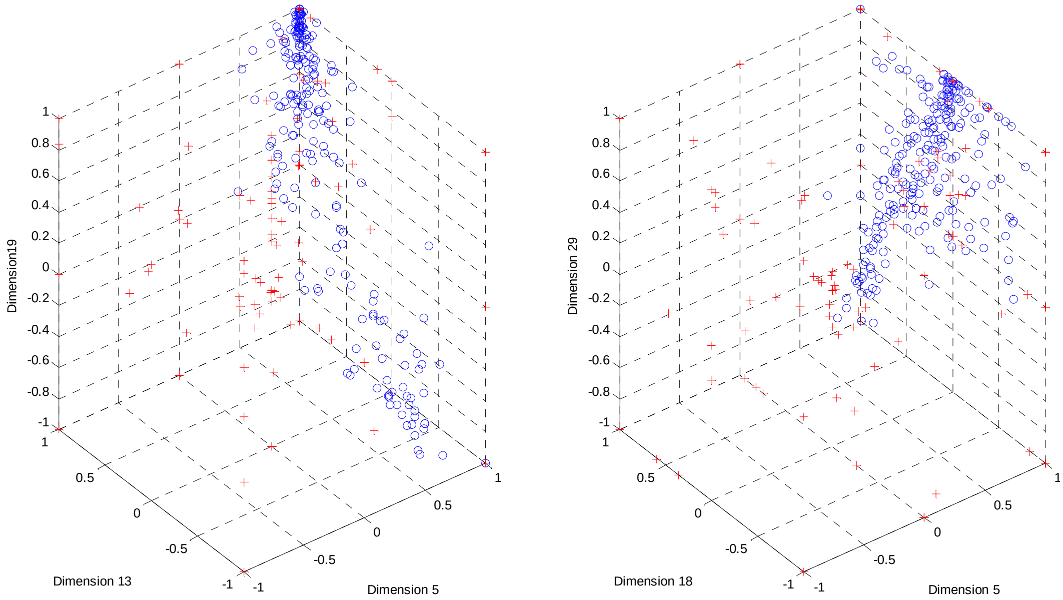


Figure 2: Grouped Scatter Plots for Ionosphere Data

Table 2: Top Subset Variables Selected with 20% Set Using SVM-RFE Ranking

Kernel	Best Subset	Best ICOMP$_{PERF}$	Training Error CI	Testing Error CI
Linear	{27,12}	121.14	[0.03046, 0.33089]	[0.12241, 0.38356]
Ranking	{27,12,24,32,30,31,4,18,20,34,2,26,9,6,8,28,16,14,25,7,5,22,3,29,17,15,21,23,10,1,11,33,19,13}			
Cauchy	{1-9,11-34}	87.61	[0.08101, 0.36773]	[0.25495, 0.42540]
Ranking	{24,5,3,33,26,31,6,9,22,34,18,11,21,19,4,32,23,15,25,12,30,29,13,2,28,20,1,8,16,27,7,14,17,10}			
Polynomial (degree=2)	{2-20,22-30,32-34}	-47953.45	[0, 0.23670]	[0.06150, 0.31593]
Ranking	{30,32,29,12,34,4,2,23,14,26,18,6,20,28,8,33,16,22,7,10,5,24,27,3,17,15,13,19,25,9,11,1,21,31}			
Polynomial (degree=3)	{2,3,8,12-14,18,20,22,24-32}	**-47957.44**	[0, 0.14278]	[0.10669, 0.21464]
Ranking	{3,14,24,26,13,28,2,8,20,30,12,18,27,31,25,29,32,22,6,16,5,4,11,10,34,1,19,33,21,7,23,9,17,15}			

Table 3: Top Subset Variables Selected with 80% Set Using SVM-RFE Ranking

Kernel	Best Subset	Best ICOMP$_{PERF}$	Training Error CI	Testing Error CI
Linear	{7}	606.94	[0.09676, 0.23190]	[0.09271, 0.26610]
Ranking	{7,27,6,31,30,28,32,26,14,8,10,16,2,24,19,4,18,11,3,20,22,34,29,13,25,21,9,33,23,17,1,12,5,15}			
Cauchy	{3}	441.65	[0.02329, 0.20342]	[0, 0.38375]
Ranking	{3,6,4,7,8,5,1,18,14,10,16,12,13,2,24,9,19,15,17,23,21,31,29,25,22,33,34,28,32,30,26,20,11,27}			
Polynomial (degree=2)	{5,14}	454.56	[0, 0.13966]	[0, 0.18645]
Ranking	{5,14,8,16,10,22,32,29,31,3,27,12,34,4,7,20,23,26,25,19,15,9,17,13,33,24,21,28,11,30,6,18,2,1}			
Polynomial (degree=3)	{5}	**441.51**	[0, 0.09553]	[0.02858, 0.02696]
Ranking	{5,4,14,34,33,30,18,22,6,16,31,32,26,25,10,12,8,20,2,29,24,28,21,3,27,7,23,19,13,17,11,15,1,9}			

Table 4: Subset Selection Based on ICOMP$_{PERF}$ with 20% Set (Polynomial: degree=3)

Rank	Variable	ICOMP$_{PERF}$	Training Error	Test Error
1	3	185.9418	0.2	0.19217
2	14	163.37627	0.2143	0.14235
3	24	125.94111	0.1	0.15658
4	26	185.42737	0.1143	0.15302
5	13	190.40902	0.0714	0.15658
6	28	158.21993	0.0571	0.21708
7	2	254.1286	0.1	0.16370
8	8	171.2137	0.0571	0.14591
9	20	123.1558	0.0286	0.14947
10	30	143.0001	0.0143	0.14235
11	12	-47854.7927	0	0.18861
12	18	48313.953	0.0286	0.17082
13	27	136.8903	0.0143	0.10676
14	31	273.9907	0.0429	0.13879
15	25	-47934.01	0	0.11744
16	29	201.188	0	0.13523
17	32	48348.9985	0.0571	0.17794
18	**22**	**-47957.4425**	**0**	**0.17082**
19	6	48366.0982	0.0571	0.17794
20	16	96.5792	0.0143	0.18505
21	5	-47867.1294	0	0.15658
22	4	48260.6735	0.0143	0.14947
23	11	-47852.1665	0	0.22420
24	10	196.314	0	0.15658
25	34	200.772	0	0.13879
26	1	48249.2818	0.0143	0.14591
27	19	-47847.3168	0	0.19573
28	33	185.951	0	0.12100
29	21	205.575	0	0.16370
30	7	204.57	0	0.21352
31	23	208.216	0	0.13879
32	9	188.548	0	0.17438
33	17	48266.37	0.0143	0.13879
34	15	-47870.13	0	0.15658

Table 5: Subset Selection Based on ICOMP$_{\mathrm{PERF}}$ with 80% Set (Polynomial: degree=3)

Rank	Variable	ICOMP$_{\mathrm{PERF}}$	Training Error	Test Error
1	**5**	**441.5118**	**0.1708**	**0.1714**
2	4	541.7953	0.0890	0.1143
3	14	698.4002	0.0676	0.1714
4	34	838.3473	0.0819	0.0857
5	33	717.6374	0.0605	0.1143
6	30	754.7581	0.0534	0.0857
7	18	752.3821	0.0463	0.1286
8	22	769.0320	0.0427	0.0857
9	6	772.8447	0.0391	0.0571
10	16	768.1328	0.0356	0.0857
11	31	697.4870	0.0249	0.0714
12	32	795.0805	0.0249	0.1143
13	26	834.1837	0.0285	0.0857
14	25	603.7533	0.0142	0.1571
15	10	950.3118	0.0249	0.0429
16	12	640.0070	0.0142	0.1429
17	8	717.9700	0.0107	0.1286
18	20	797.0560	0.0107	0.0429
19	2	679.8700	0.0071	0.0714
20	29	801.4970	0.0071	0.1286
21	24	911.7650	0.0107	0.1143
22	28	682.2560	0.0071	0.1143
23	21	907.2940	0.0107	0.0571
24	3	689.8410	0.0071	0.0714
25	27	911.3660	0.0107	0.1000
26	7	501.0110	0.0036	0.1286
27	23	994.9170	0.0071	0.1000
28	19	817.5010	0.0071	0.1143
29	13	612.9350	0.0036	0.0857
30	17	1008.7330	0.0071	0.0429
31	11	808.4890	0.0071	0.0714
32	15	623.0110	0.0036	0.0857
33	1	1001.9020	0.0071	0.0429
34	9	628.2170	0.0036	0.1143

2.5.2 Aorta Data

The aorta data are from medical imaging for a study of heart tissue. Hardening of the arteries is the leading cause of death and debility in the industrial world. Nuclear magnetic resonance (NMR) imaging has a role in diagnosing of arteries for prognosis of heart attack. The NMR aorta data was used by Pearlman (1986). The dataset sampled from 418 patients on 20 different NMR image characteristics. The first group consists of 194 patients who exhibited early atheroma, and the second group consists of 224 patients who were healthy. Figure 3 shows grouped scatter plots for the poor separation of dimension 3 against dimensions 13, 19 and against dimensions 10, 20 (group1: blue, group2: red). Tables 6 and 7 show that the best subset based on $ICOMP_{PERF}$ is obtained at the Cauchy kernel in the 20% set and inverse multi-quadratic kernel in the 80% set. The confidence intervals are obtained based on $ICOMP_{PERF}$. The confidence intervals are significantly narrow intervals in both of the sets. Tables 8 and 9 show the best subset selected based on $ICOMP_{PERF}$.

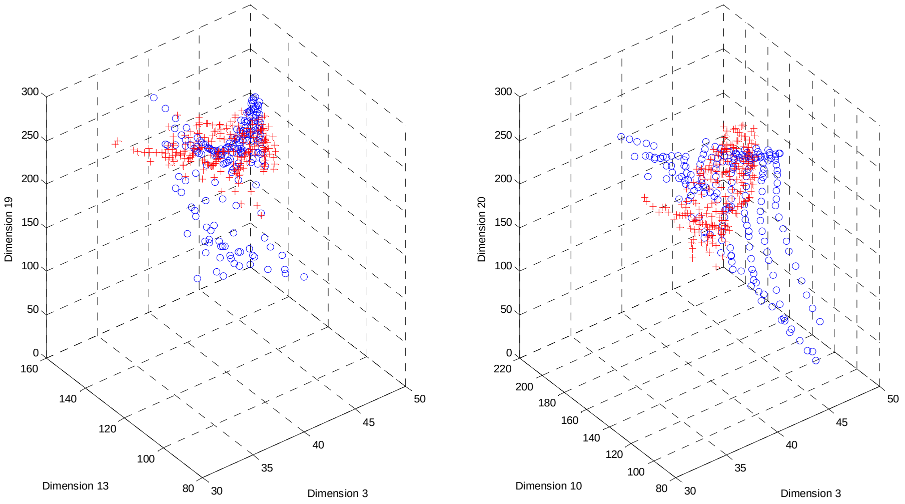


Figure 3: Grouped Scatter Plots for Aorta Data

Table 6: Top Subset Variables Selected with 20% Set Using SVM-RFE Ranking

Kernel	Best Subset	Best ICOMP$_{PERF}$	CI for Training Error	CI for Testing Error
Cauchy	{4}	**-57785.1**	[0, 0]	[0, 0.00767]
Ranking	{4,14,20,5,12,10,11,13,17,9,1,19,18,16,3,6,2,8,15,7}			
Gaussian	{14,13,12,17}	-57071	[0, 0.11714]	[0, 0.28813]
Ranking	{14,13,12,17,10,4,16,20,18,19,11,15,8,9,6,7,5,3,2,1}			
Polynomial (degree=2)	{4}	-57679	[0, 0]	[0, 0.02696]
Ranking	{4,15,10,11,9,6,18,2,8,14,7,1,16,13,12,5,17,19,20,3}			
Inv. Multi Quadratic	{17,7,20,15}	-57414.62	[0, 0.04342]	[0, 0.24672]
Ranking	{17,7,20,15,10,18,16,6,5,14,1,9,2,11,12,3,8,13,4,19}			

Table 7: Top Subset Variables Selected with 80% Set Using SVM-RFE Ranking

Kernel	Best Subset	Best ICOMP$_{PERF}$	CI for Training Error	CI for Testing Error
Cauchy	{20,7,15}	-228526.2	[0, 0.1254]	[0, 0.26101]
Ranking	{20,7,15,11,5,16,6,10,8,4,19,17,13,14,9,3,18,2,1,12}			
Gaussian	{2}	-229734.4	[0, 0]	[0, 0.01033]
Ranking	{2,17,7,10,9,16,6,15,8,20,13,14,1,11,3,4,5,18,12,19}			
Polynomial (degree=2)	{4}	-229608	[0, 0]	[0, 0]
Ranking	{4,16,15,14,11,12,19,18,3,17,1,8,9,10,2,13,5,6,20,7}			
Inv. Multi Quadratic	{4}	**-229759.2**	[0, 0]	[0, 0]
Ranking	{4,7,15,20,16,5,17,10,14,6,8,18,11,13,1,12,9,2,19,3}			

Table 8: Subset Selection Based on ICOMP$_{\text{PERF}}$ with 20% Set (Cauchy)

Rank	Variable	ICOMP$_{\text{PERF}}$	Training Error	Test Error
1	**4**	**-57785.101**	**0**	**0**
2	14	236.839	0	0
3	20	238.263	0	0
4	5	238.381	0	0
5	12	238.382	0	0
6	10	238.382	0	0
7	11	238.381	0	0.006
8	13	238.382	0	0.003
9	17	238.382	0	0.006
10	9	238.381	0	0
11	1	238.382	0	0
12	19	238.382	0	0
13	18	238.381	0	0
14	16	238.382	0	0
15	3	238.382	0	0
16	6	238.381	0	0.003
17	2	238.382	0	0
18	8	238.382	0	0.012
19	15	238.381	0	0
20	7	238.382	0	0

Table 9: Subset Selection Based on ICOMP$_{\text{PERF}}$ with 80% Set (Inv. Multi Quadratic)

Rank	Rank	ICOMP$_{\text{PERF}}$	Training Error	Test Error
1	**4**	**-229759.22**	**0**	**0**
2	7	941.35	0	0
3	15	945.22	0	0
4	20	946.32	0	0
5	16	947.29	0	0
6	5	947.53	0	0
7	17	947.71	0	0
8	10	947.77	0	0
9	14	947.82	0	0
10	6	947.8	0	0
11	8	947.84	0	0
12	18	947.85	0	0
13	11	947.83	0	0
14	13	947.85	0	0
15	1	947.84	0	0
16	12	947.86	0	0
17	9	947.84	0	0
18	2	947.84	0	0
19	19	947.84	0	0
20	3	947.85	0	0

2.6 Comparison with Other RFE Based Methods

To compare three different RFE based methods; SVM-RFE, SVM-Gradient-RFE, SVM-$ICOMP_{PERP}$-RFE, the ionosphere and aorta datasets are used with the same kernel functions that are used in Tables 2, 3, 6, and 7. The datasets are randomly partitioned into two cases; 20%/80% and 80%/20% as training/test sets. Tables 10 and 11 present comparisons of three RFE based methods using the ionosphere data with four different kernel functions in two different cases. The average error rate represents the misclassification error rate for the test set. The SVM-$ICOMP_{PERF}$-RFE is the clear winner for most kernel functions except the linear kernel in the 80%/20% case. The best performance is obtained using the Cauchy kernel in the two cases with 88.12% and 93.28% accuracies. Tables 12 and 13 present comparisons of the three RFE based methods using the aorta data with four different kernel functions in two different cases. As shown in Tables 12 and 13, the SVM-$ICOMP_{PERF}$-RFE is the best method for the polynomial kernel (degree=2) with 99.99% accuracy for the 20%/80% case, the polynomial kernel (degree=2) with 99.88% accuracy for the 80%/20% case, and the inverse multi-quadratic kernel with 100% accuracy for the 80%/20% case. Figure 4 shows line plots of error rates for the test set with the Cauchy kernel function, which gives smallest average error rates using the ionosphere data shown in Tables 10 and 11. Figure 5 shows line plots of error rates for the test set with the polynomial kernel (degree=2) and inverse multi-quadratic kernel functions, which give smallest average error rates using the aorta data shown in Tables 12, and 13. The SVM-$ICOMP_{PERF}$-RFE is competitive with both SVM-RFE and SVM-Gradient-RFE as shown in Figure 4. Also, SVM-$ICOMP_{PERF}$-RFE outperforms SVM-RFE and SVM-Gradient-RFE with few variables as shown in Figure 5.

Table 10: Comparison Using Ionosphere Data with 20%/80%

	SVM-RFE		SVM-Gradient-RFE		SVM-ICOMP$_{\text{PERP}}$-RFE	
	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)
Linear	0.22273	137.98	0.19552	117.39	**0.19510**	877.81
Cauchy	0.16381	422.00	0.16140	155.53	**0.11880**	907.72
Polynomial (degree=2)	0.19992	226.36	0.18903	158.59	**0.17522**	887.28
Polynomial (degree=3)	0.21572	231.63	0.21195	158.78	**0.18830**	970.72

Table 11: Comparison Using Ionosphere Data with 80%/20%

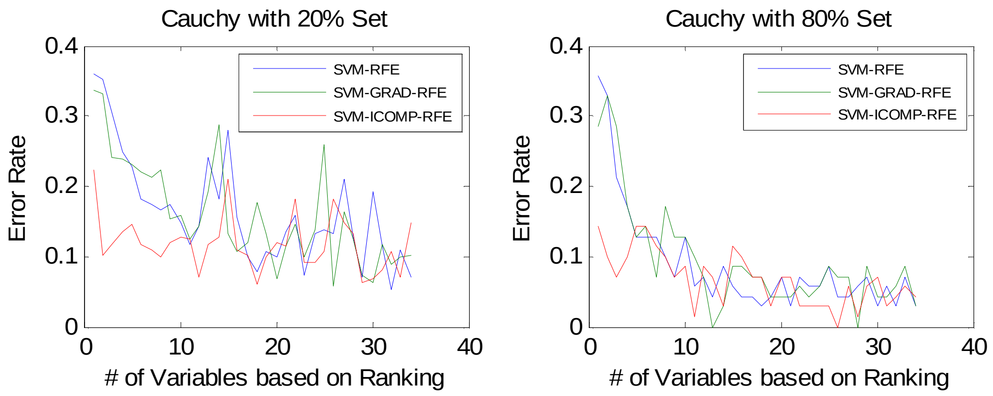
	SVM-RFE		SVM-Gradient-RFE		SVM-ICOMP$_{\text{PERP}}$-RFE	
	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)
Linear	0.15546	1166.61	**0.15420**	1095.47	0.16177	13352.33
Cauchy	0.08908	3557.66	0.09454	1280.13	**0.06723**	17755.05
Polynomial (degree=2)	0.16933	1882.02	0.13445	1192.17	**0.13277**	16334.36
Polynomial (degree=3)	0.17941	1679.30	0.15840	1228.61	**0.13656**	14771.78

Table 12: Comparison Using Aorta Data with 20%/80%

	SVM-RFE		SVM-Gradient-RFE		SVM-ICOMP$_{\text{PERP}}$-RFE	
	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)
Cauchy	**0.00374**	388.38	0.13488	144.78	0.04880	480.70
Gaussian	**0.05749**	337.89	0.13084	143.44	0.10195	534.39
Polynomial (degree=2)	0.05404	126.20	0.11033	114.31	**0.00015**	980.63
Inv. Multi Quadratic	**0.02784**	327.25	0.12590	128.97	0.05434	496.58

Table 13: Comparison Using Aorta Data with 80%/20%

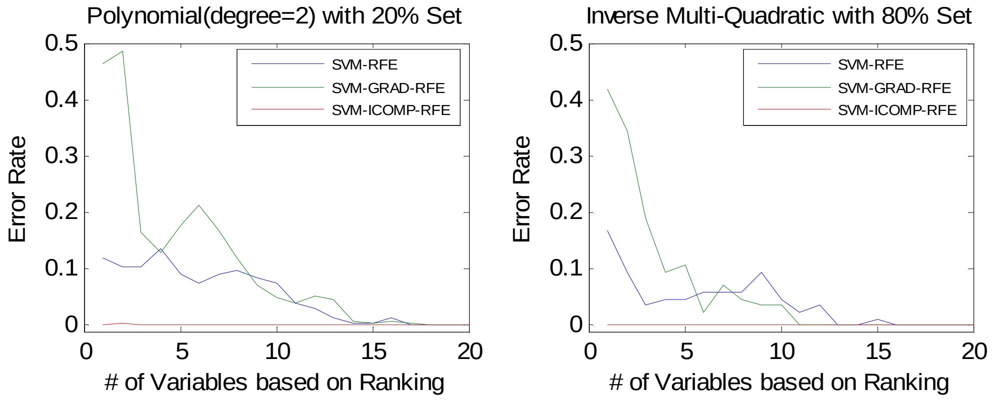
	SVM-RFE		SVM-Gradient-RFE		SVM-ICOMP$_{\text{PERP}}$-RFE	
	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)	Average Error Rate	Time(sec.)
Cauchy	**0.01548**	4159.63	0.07738	1151.69	0.04167	8501.92
Gaussian	**0.02083**	4141.45	0.06310	1361.56	0.03393	12093.58
Polynomial (degree=2)	0.03095	1235.92	0.05000	1086.34	**0.00119**	9572.74
Inv. Multi Quadratic	0.03929	4372.67	0.06845	1432.59	**0**	8743.09



(a) (b)

Figure 4: Best Results of SVM-*ICOMP*$_{PERF}$-RFE Using Ionosphere Data

(a) Cauchy Kernel Function with 20% Set (b) Cauchy Kernel Function with 80% Set



(a) (b)

Figure 5: Best Results of SVM-*ICOMP*$_{PERF}$-RFE Using Aorta Data

(a) Polynomial Kernel (degree=2) Function with 20% Set

(b) Inverse Multi-Quadratic Kernel Function with 80% Set

Chapter 3 Dual Variables Functional Support Vector Machine and Modified Floating Search Based Variable Selection

3.1 Introduction

Secondary batteries have become an essential part of portable multimedia devices such as mobile phones, camcorders, and computers. Among a number of secondary batteries used in the current market, lithium-ion batteries have overcome several weaknesses of traditional nickel cadmium (Ni-Cd) and nickel metal hybrid (Ni-MH) secondary batteries, such as heavy weight and potential for pollution. Further, due to their demonstrated excellent energy density and cycle-life performance, lithium-ion batteries have taken the largest part of commercial markets for powering high-end electronics applications (Broussely and Archdale 2004).

In the mass production stage of secondary batteries, it is crucial to assure product quality within a limited time. Cycle-life, which is directly related to the battery life, is one of the major characteristics to be monitored. Evaluation of the cycle-life requires a lot of charge/discharge cycles, thus it is a very time-consuming task. This has caused a major difficulty for battery manufacturers to reduce product development time. For this reason, a more time-efficient method for assessing the cycle-life of secondary batteries is needed.

In this chapter, a new time-efficient method is proposed to assess the quality of secondary batteries where their cycle-lives are subject to monitoring. For this, a dual-variables functional support vector machine (FSVM) was developed to minimize the errors in discriminating between the conforming and nonconforming batteries.

3.2 Motivating Example

A lithium-ion battery is composed of four basic elements: the cathode (positive electrode), anode (negative electrode), electrolyte solution, and separator. While the battery is being charged, lithium-ions from the cathode leaving it with a net negative charge are forced onto the anode giving it a positive charge. During the discharge, the ions flow in the opposite direction, from the anode to the cathode. Because such reaction is reversible in the secondary battery while impossible in the primal battery, secondary batteries are capable of being recharged and reused up to hundreds of cycles (one cycle represents one charge/discharge). Basic performance of a lithium-ion battery is characterized by its capacity, which is generally defined as the amount of charge available expressed in ampere-hours (*Ah*). Cycle-life is defined as the number of complete charge/discharge cycles before its nominal capacity falls below the pre-specified value of its initial capacity. Although it is desirable that the battery retains the initial capacity as much as possible during usage time, the capacity is subject to decrease through repetitive charge/discharge cycles. Research issues to improve the cycle-life have attracted a lot of attention (Johnson and White 1998; Broussely and Archdale 2004). Figure 6 shows the remaining capacities of 43 battery cells as the cycle proceeds. They were randomly selected from the manufactured lots for several months. In the cycle-life tests for qualification, each sample is subject to check whether or not its capacity reaches to a fixed threshold during a specific number of cycles. The requirements of the threshold level and the number of cycles are generally pre-determined based on the industry standard or customer requirement. For instance, U.S. Advanced Battery Consortium (USABC 1996) defines the threshold value as 80% of its initial capacity during 600-cycles for electrical vehicle batteries (given by 80% during 400-cycles here). Then, the battery cell is classified as either a conforming or a nonconforming cell according to its requirement.

As shown in Figure 6, the capacity degradation has a typical nonlinear trend: capacity degrades sharply during some initial cycles, and then the degradation rate

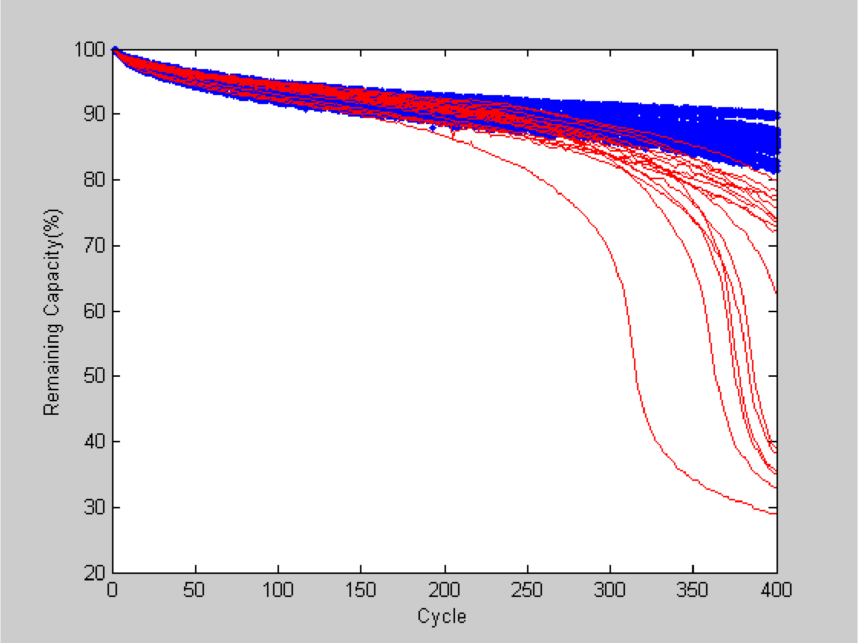


Figure 6: Remaining Capacities of Selected Battery Samples during Cycle-Life Tests

becomes relatively slow. In recent years, much attention has been placed particularly on relating observed phenomenon with state-of-art analysis techniques. It is commonly recognized that such capacity degradation is accompanied by loss of active lithium ions and an increase in the internal impedance of the battery. Both of them presumably are caused by an electrochemical parasite reaction (Bloom *et al*. 2001; Wright and Motloch 2001; Sikha *et al*. 2004; Yoshida *et al*. 2006; Ning *et al*. 2006). Because each cell in Figure 6 has possibly experienced different sources of variation during fabrication, its capacity degradation is slightly different. Nevertheless, some degradation paths show catastrophic drops in the end of cycles, leading to products of poor quality.

It usually takes a long time to finish the whole set of test cycles specified in the requirements; for example, 400 cycles require at least fifty calendar days. Because such long testing spans have been impeding efficient operation of manufacturing lines, battery engineers have struggled to devise various ways to reduce the test duration. However, if one determines the acceptability of a lot with a shorter cycle, more risky decisions may be made. Figure 7 shows separate distributions of the remaining capacity for conforming (26 cells) and nonconforming (17 cells) battery samples at some fixed cycles. Note that the boundaries between both groups are not distinctive at relatively short cycles; even the mean capacity of nonconforming cells is larger than that of conforming cells. It is practically impossible to visually discriminate between the conforming and nonconforming cells with this short testing duration.

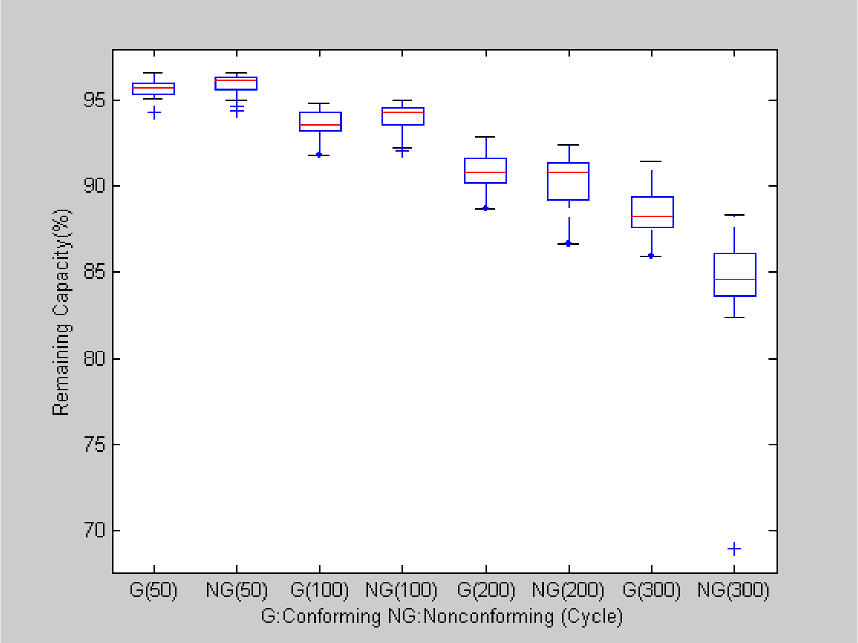


Figure 7: Box-Plots of Remaining Capacity at Some Cycles

This facilitates introduction to a support vector machine (SVM). As the latest classification technique exploited in the data-mining field, the SVM separates a given data set into several groups based upon a certain classification rule. Because of its excellent classification performance and lasting progress from a methodological perspective, the SVM has been found in many useful applications (Burges 1998). Making full use of such a state-of-the-art method, a superior rule can be expected in some sense to discriminate conforming cells from nonconforming cells even with shorter cycle data. In this study, the number of cycle runs is assumed to be a continuous variable.

3.3 Dual Variables Functional Support Vector Machine for Classification of Cycle-Life Curves

Functional SVM (FSVM) is an extension of SVM to functional data generated by a number of individuals repeatedly in a regular sequence, in which each observation reflects a smooth variation in input data. In some sense, the FSVM can be considered as a generalization of SVM with respect to the type of data structures (Jank and Shmueli 2006). Figure 8 shows the procedure of the dual-variables FSVM. Dual-variables FSVM uses both first and second derivatives for data representation. In addition, a modified floating search is proposed to reduce the computational time in the iterative variable selection.

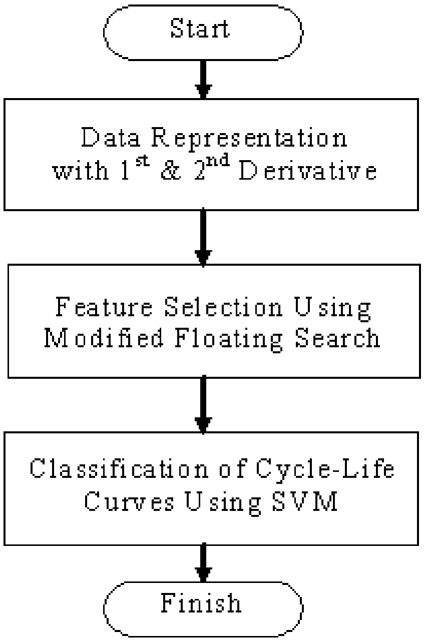

Start
Data Representation with 1st & 2nd Derivative
Feature Selection Using Modified Floating Search
Classification of Cycle-Life Curves Using SVM
Finish

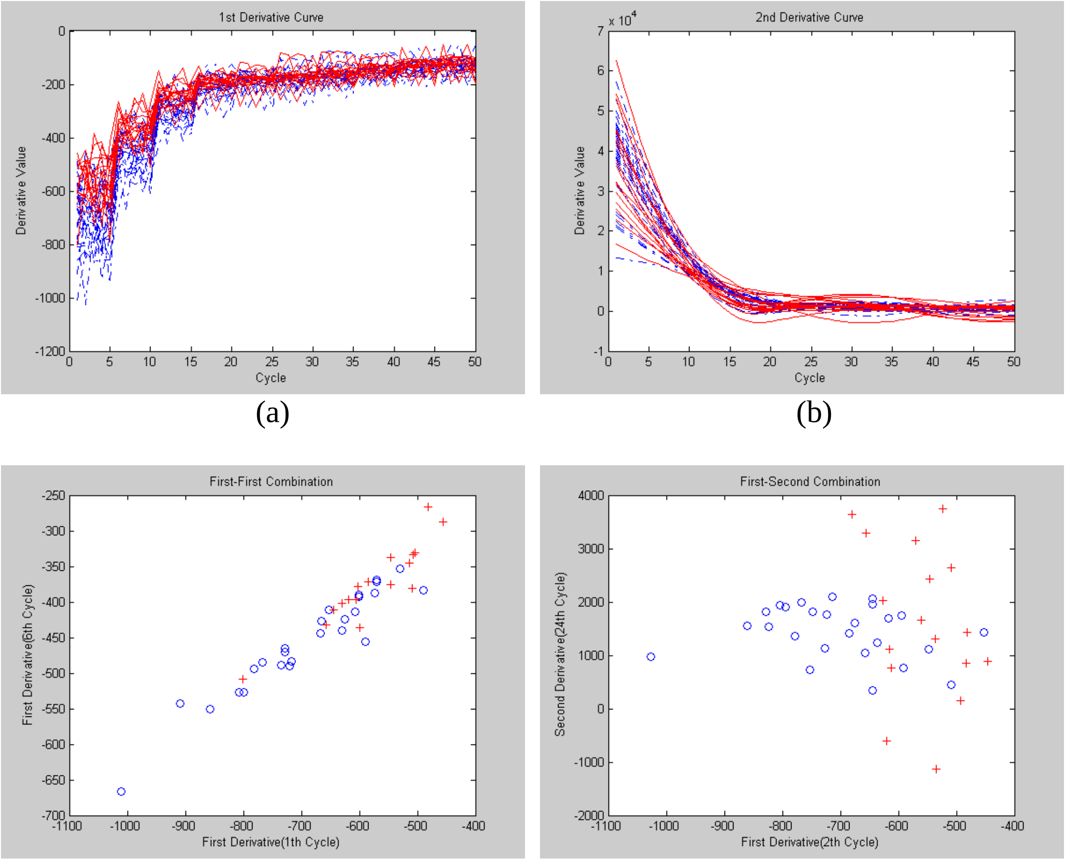
Figure 8: Flowchart of the Dual Variables FSVM

3.3.1 Data Representation with the First and Second Derivatives

Data representation for functional structures is one of the key issues in implementing the FSVM. Rossi and Villa (2006) proposed various types of data representations, such as derivatives, wavelets and Fourier representation. Ramsay and Silverman (2005) claimed that much of the variation between curves can be explained at the level of certain derivatives. For this reason, even with its own simplicity, functional derivative representation showed successful results in many studies (Ferraty and Vieu 2003; Rossi and Conan-Guez 2005). Figures 9a and 9b show the first and second derivatives of degradation curves of Figure 6 over 50 cycles, respectively. In some cases, a combination of the derivatives with different orders may lead to better classification performance. Figures 9c and 9d show the potential for the early discrimination of defective lithium-ion batteries using the dual variables of first and second derivatives. The derivatives can be calculated using a B-spline (de Boor 2001) approximation to avoid numerical stability problem of direct computation. Let $\{B_1, B_2, ..., B_p\}$ be the B-spline basis where p stands for the number of knots. Then, the q^{th} derivative of discretized curve $\mathbf{y}_i = \{y_i(t_1), y_i(t_2), \cdots, y_i(t_o)\}$ for observed number of cycles t_o can be approximated by

$$\hat{y}_i^{(q)}(t) = \sum_{k=1}^{p} \hat{c}_{ik} B_k^{(q)}(t), \tag{29}$$

where $\hat{\mathbf{c}}_i = (\hat{c}_{i1}, \hat{c}_{i2}, \cdots, \hat{c}_{ip}) = \underset{(c_1, \cdots, c_p) \in \Re^p}{\arg\min} \sum_{j=1}^{o} \left\{ y_i(t_j) - \sum_{k=1}^{p} c_{ik} B_k(t_j) \right\}^2$.



(a)

(b)

(c)

(d)

Figure 9: (a) First Derivatives of the Cycle-Life Curves (b) Second Derivatives of the Cycle-Life Curves (c) Classification with a First-First Derivative Combination (d) Classification with a First-Second Derivative Combination

3.3.2 Variable Selection Using Modified Floating Search

Because a dual or multiple data-representations lead to higher-dimensional space, a variable selection technique is needed to reduce the dimension. The main idea of variable selection is to find a proper subset of input variables by eliminating variables with redundant or meaningless information. Heavy computing often follows, due to the iterative process for finding the proper subset. A floating search technique is an excellent method to guarantee a near optimal subset without any exhaustive searching (Pudil *et al*. 1994).

For the given cycle *t*, variable selection is conducted to search for the optimal variables that have the highest separability between the good and defective batteries, and repeat for the next cycle *t*+1. There is a high possibility of redundant iterations because the existing algorithm is repeatedly applied to similar input data set as the observation period prolongs. a modified floating search algorithm is proposed so as to start with the best variables set of cycle $(t-1)$ at cycle t . The detailed procedure for the iterative variable selection is explained in Figure 10. A divergence is calculated and compared to find the best subset of *d* variables from a given set of G variables $(1 \le d \le G)$ in the SFFS process. Divergence is one of the popular criteria for class separability. It takes into account the correlation that exists among selected variables and influences classification capabilities of the selected variables. Assuming *p*-dimensional multivariate normal distribution, the divergence between a class *i* and *j* is given by (Fukunaga 1990)

$$J_{ij}(\mathbf{y}) = \frac{1}{2} trace\left[(\mathbf{\Sigma}_i^{-1} + \mathbf{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \right] + \frac{1}{2} trace\left[(\mathbf{\Sigma}_i^{-1} + \mathbf{\Sigma}_j^{-1})(\mathbf{\Sigma}_i^{-1} - \mathbf{\Sigma}_j^{-1}) \right] \quad (30)$$

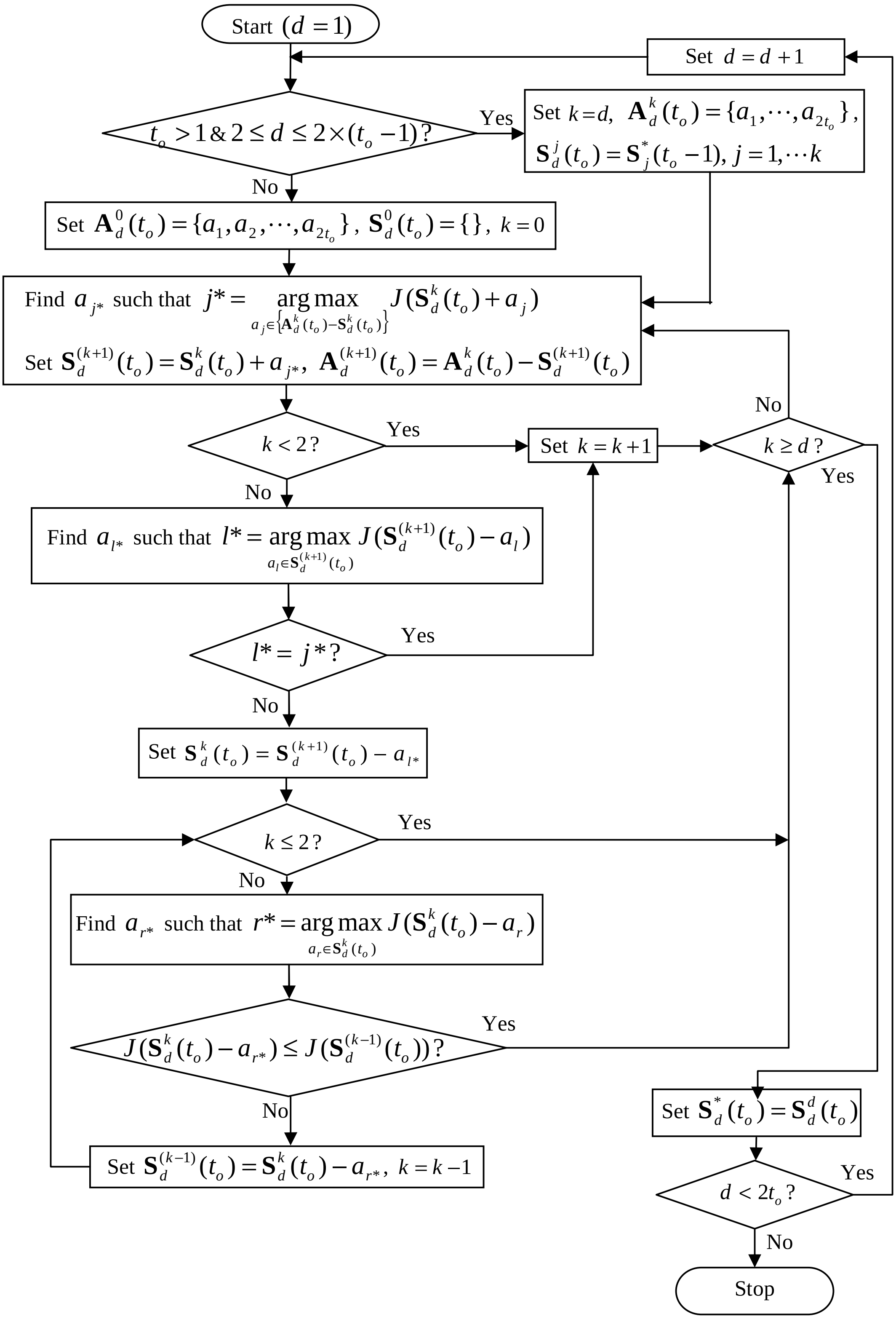

Start $(d=1)$
Set $d=d+1$
$t_o > 1 \,\&\, 2 \le d \le 2\times(t_o-1)$?
Yes
Set $k=d$, $\mathbf{A}_d^k(t_o)=\{a_1,\cdots,a_{2t_o}\}$, $\mathbf{S}_d^j(t_o)=\mathbf{S}_j^*(t_o-1), j=1,\cdots k$
No
Set $\mathbf{A}_d^0(t_o)=\{a_1,a_2,\cdots,a_{2t_o}\}$, $\mathbf{S}_d^0(t_o)=\{\}$, $k=0$
Find a_{j*} such that $j^* = \underset{a_j\in\{\mathbf{A}_d^k(t_o)-\mathbf{S}_d^k(t_o)\}}{\arg\max} J(\mathbf{S}_d^k(t_o)+a_j)$
Set $\mathbf{S}_d^{(k+1)}(t_o)=\mathbf{S}_d^k(t_o)+a_{j*}$, $\mathbf{A}_d^{(k+1)}(t_o)=\mathbf{A}_d^k(t_o)-\mathbf{S}_d^{(k+1)}(t_o)$
$k<2$?
Yes
Set $k=k+1$
$k \ge d$?
No
Yes
No
Find a_{l*} such that $l^* = \underset{a_l\in\mathbf{S}_d^{(k+1)}(t_o)}{\arg\max} J(\mathbf{S}_d^{(k+1)}(t_o)-a_l)$
$l^* = j^*$?
Yes
No
Set $\mathbf{S}_d^k(t_o)=\mathbf{S}_d^{(k+1)}(t_o)-a_{l*}$
$k \le 2$?
Yes
No
Find a_{r*} such that $r^* = \underset{a_r\in\mathbf{S}_d^k(t_o)}{\arg\max} J(\mathbf{S}_d^k(t_o)-a_r)$
$J(\mathbf{S}_d^k(t_o)-a_{r*}) \le J(\mathbf{S}_d^{(k-1)}(t_o))$?
Yes
No
Set $\mathbf{S}_d^{(k-1)}(t_o)=\mathbf{S}_d^k(t_o)-a_{r*}$, $k=k-1$
Set $\mathbf{S}_d^*(t_o)=\mathbf{S}_d^d(t_o)$
$d < 2t_o$?
Yes
No
Stop

Figure 10: Block Diagram of the Modified SFFS

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix of class *i*, respectively. Then, the best subset of *d* variables is chosen to maximize the divergence value.

Suppose that the observed number of cycles and desired number of variables are given by t_o and d, respectively. Further, let $\mathbf{A}_d(t_o)$ and $\mathbf{S}_d(t_o)$ be respective sets of available and selected variables for given t_o and d. In the SFFS process, new variables from $\mathbf{A}_d(t_o)$ are included in the current $\mathbf{S}_d(t_o)$ and successive steps follow to exclude the worst variables in the newly updated $\mathbf{S}_d(t_o)$, provided further improvement can be made to the previous set (Pudil *et al*. 1994).

3.3.1 Dual Variables FSVM-Based Detection of Defective Lithium-Ion Batteries with Degradation Curves

Given a subset of *d* variables, dual-variables FSVM is applied to detect defective lithium-ion batteries based on degradation curves. The SVM finds the optimal separating hyperplane that maximizes the margin between the classes (Vapnik 1995). Consider the case of classifying a set of linearly separating data into two groups. Assume a set of training data is given by $\mathbf{M} = \{(\mathbf{y}_1, z_1), \cdots, (\mathbf{y}_n, z_n)\}$ where $\mathbf{y}_i$ is an input vector, $z_i \in (-1,1)$ is a binary class index, and *n* is the size of training data. Then, a decision boundary (i.e. classifier) that partitions underlying vector space into two classes can be represented by the following hyperplane:

$$\mathbf{w}^{\mathrm{T}}\mathbf{y} + b_s = 0 \qquad (31)$$

where $\mathbf{w}$ is the weight vector and b_s is the bias. The objective of SVM is to find maximum margin decision boundary between two parallel hyperplanes, $\mathbf{w}^{\mathrm{T}}\mathbf{y}+b_s=1$ and $\mathbf{w}^{\mathrm{T}}\mathbf{y}+b_s=-1$. Since the margin is given by $2/\|\mathbf{w}\|$, the corresponding optimization problem can be written as (Vapnik 1995)

$$\begin{aligned} &\text{Minimize } \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } \ \begin{cases} z_i(\mathbf{w}^{\mathrm{T}}\mathbf{y}_i + b_s) \geq 1-\xi_i, \ i=1,2,...,n \\ \qquad\qquad \xi_i \geq 0, \ i=1,2,...,n \end{cases} \end{aligned} \tag{32}$$

where ξ_i is positive slack variable and $C\,(>0)$ is a pre-designated weight. The linearly constrained optimization problem in equation (32) can be solved in a dual problem that maximizes the following Lagrangian function:

$$L_F(\lambda) = \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j z_i z_j(\mathbf{y}_i^{\mathrm{T}}\bullet\mathbf{y}_j) \tag{33}$$

Subject to the constraint

$$\sum_{i=1}^{n}\lambda_i y_i = 0, \quad 0 \leq \lambda_i \leq C, \ i=1,2,\cdots,n \tag{34}$$

The Lagrange multipliers λ_i's can be estimated using a quadratic programming method. Once the optimum values $(\boldsymbol{\lambda}^*, b_s^*)$ are found based upon the training set of points, a new point $\mathbf{y}^0$ of the test data set is classified by the following decision rule:

$$\mathbf{y}^0 \in \begin{cases} \text{Class 1} \quad \text{if } \Psi(\mathbf{y}^0 | \mathbf{M}) = \sum_{i=1}^{n}\lambda_i^* z_i \mathbf{y}_i^{\mathrm{T}}\mathbf{y}^0 + b_s^* < 0 \\ \text{Class 2} \quad \text{if } \Psi(\mathbf{y}^0 | \mathbf{M}) = \sum_{i=1}^{n}\lambda_i^* z_i \mathbf{y}_i^{\mathrm{T}}\mathbf{y}^0 + b_s^* > 0 \end{cases} \tag{35}$$

where $\Psi(\bullet|\mathbf{M})$ is a classifier based upon the training data set $\mathbf{M}$.

Optimal subset size d^* and corresponding variables at a given number of cycles t_o are determined such that classification accuracy of the SVM is maximized. Leave-One-Out Cross Validation (LOOCV) is used to estimate the classification accuracy, which gives proper measure when there are limited samples. At this time, d^* is given by

$$d^* = \arg\max_{1 \le d \le 2t_o} \frac{1}{n} \sum_{i=1}^{n} I\left[\Psi_{s_d^*(t_o)} \left\{ \mathbf{y}_i \middle| \mathbf{M}^{(-i)} \right\},\ z_i \right], \tag{36}$$

where $\mathbf{M}^{(-i)}$ is the original input data with $(\mathbf{y}_i, z_i)$ removed, $\mathbf{S}_d^*(t_o)$ is the best subset of d variables at t_o, $\Psi_{s_d^*(t_o)}\left\{ \bullet \middle| \mathbf{M}^{(-i)} \right\}$ is a classifier obtained from implementing SVM with $\mathbf{S}_d^*(t_o)$ on the training data set $\mathbf{M}^{(-i)}$, $I[a_1, a_2] = 1$ if $a_1 = a_2$ and 0 otherwise.

3.4 Motivating Example Revisited

The proposed method was applied to 43 sample curves; 26 curves of conforming samples and 17 curves of nonconforming samples. In order to see how flexible a combination of the derivatives having different orders was, the changes in optimal variable sets were observed. For example, Table 14 shows each set of selected variables ($d = 10$) at 20 and 50 cycles. There is no second derivative on the list for 20 cycles (No. 1~20: 1st derivative, No. 21~40: 2nd derivative) and, on the other hand, 4 of 10 variables are second derivatives on the list for 50 cycles (No. 1~50: 1st derivative, No. 51~100: 2nd derivative). It implies that the functional classification using the first derivative only may be successful for the data of the initial period, but may not be effective for the wider range of the observation period. The computational times for selecting the best subset of variables ($d = 15$) using the existing and modified SFFS algorithms are summarized in

Figure 11. A computer with Pentium IV 3.6-GHz processor and MATLAB 6.5 as the programming language are used. As expected, the algorithms found the same best variable sets but the existing algorithm required a longer computational time than the modified one, particularly at the longer cycle time. Figure 12 shows error rates produced by applying the dual-variables FSVM to discriminate between the conforming and nonconforming cells. During the classification procedure, if the error rate at the current cycle run is greater than the one at the previous stage, then the current error rate is set to the previous value. Therefore, the error rate has a monotonic non-increasing function of cycles. Further, the error rates are compared with those values of the cases where either 1st or 2nd derivative is solely employed. The proposed method gives better performance than the others. Note that one may have tolerable error rates even with heavily truncated number of cycles (e.g., 20 cycles).

Table 14: Changes in Selected Variables Sets (* Second Derivative)

Cycle Runs	Number of Selected Variables $(d = 10)$
20	7, 2, 12, 10, 15, 9, 4, 14, 17, 6
50	7, 2, 74*, 75*, 12, 73*, 76*, 10, 15, 9

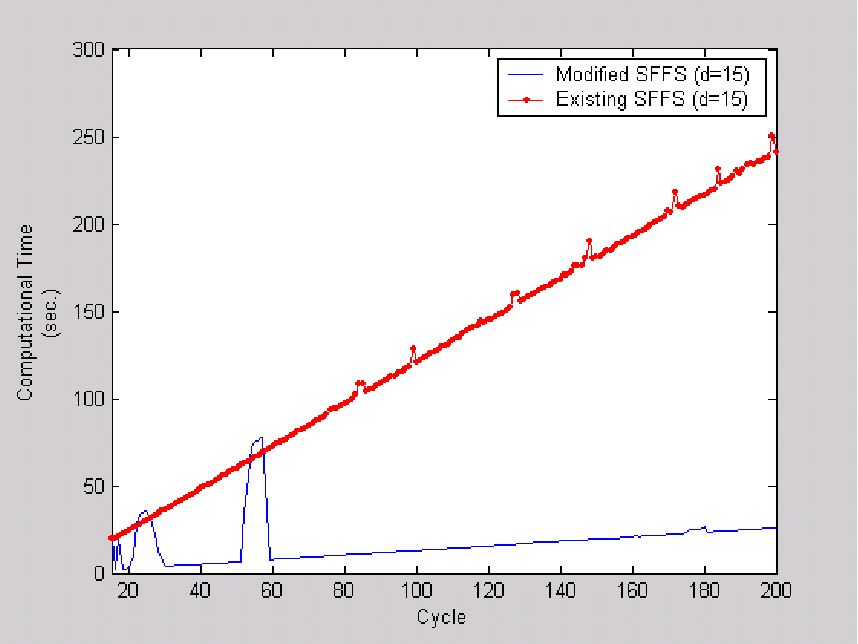


Figure 11: Computational Time for Selecting the Best Variables ($d = 15$)

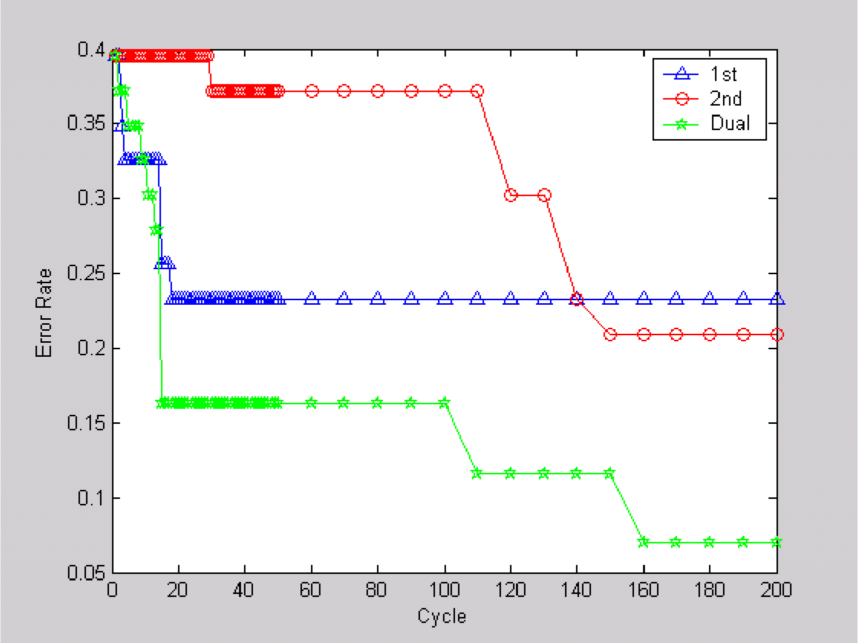


Figure 12: Error Rate versus Cycle

Chapter 4 Two-Stage Classification Procedure Based on Multi-Scale Vertical Energy Wavelet Thresholding and SVM-Based Gradient Recursive Feature Elimination

4.1 Introduction

Near-infrared (NIR) spectroscopy has been used extensively in many areas as a fast, reliable, cost-effective, and non-destructive measurement method (Kalivas 1997). NIR data often consist of several hundred to some thousand variables (wavelengths), where different parts of the spectrum are correlated with each other. Considering both the high-dimensionality and the redundant nature of NIR data, it is necessary to reduce the dimension of the data for the subsequent processing and to select a few wavelengths that better explain the data.

Variable selection is an important problem in machine learning (Bradley *et al*. 1998) and is the process of selecting input variables that are most predictive of a given output. Variable selection identifies a small subset of variables so that the classifier constructed with the selected variables minimizes error and better explains the data (Koller and Sahami 1996). Benefits of variable selection include reducing computation times, providing better discriminating hyperplanes and giving a better understanding of the data.

Wavelets are popular as preprocessing tools for spectral data (Chau *et al*. 1997). Usually variable selection is based directly on high-dimensional wavelet coefficients and this can be computationally expensive (Staszewski 1998; Subramani 2006). This chapter proposes a two-stage scheme for the classification of NIR spectral data. In the first stage, the dimension of high-dimensional spectral data is reduced using a multi-scale vertical energy thresholding (MSVET) procedure. In the second stage, a few important wavelet coefficients is selected using SVM gradient-recursive feature elimination (RFE).

In order to reduce the dimension of spectral data, many thresholding techniques have been used, including the shrinkage method (Donoho and Johnstone 1995), Stein's unbiased risk estimate (SURE) method (Donoho and Johnstone 1994) and the approximation minimum description length (AMDL) method (Saito 1994). However, these techniques were designed for de-noising purposes. Jung *et al*. (2006) proposed a vertical-energy-thresholding (VET) procedure for the data reduction of multiple data curves. The VET procedure does not consider the information scale of wavelets, which includes different types of information for decision-making. A multi-scale vertical energy thresholding (MSVET) procedure is proposed. It determines an optimal threshold for each of the scales by extending the idea of the VET procedure.

Recently, several researchers developed variable selection methods based on support vector machines (SVM) (Rakotomamonjy 2003; Weston *et al*. 2003; Mao 2004). Kernel-based methods including SVM are fast becoming standard tools for solving various problems. Guyon *et al*. (2002) proposed SVM-RFE for the selection of genes in micro-array data. The SVM gradient-RFE procedure is applied to identify a subset of predetermined size of all variables available for inclusion in the support vector classifier.

In the proposed two-stage scheme, the MSVET wavelet analysis performs noise suppression and data reduction of high-dimension spectral data. SVM gradient-RFE variable selection identifies an optimal subset of compressed wavelet coefficients for classification. Performing variable selection in the wavelet domain on reduced-dimension NIR spectral is expected to yield more reliable classification accuracy, with higher computation efficiency, than handling the full sets of noisy data. The performance of the proposed method is demonstrated using four NIR data sets.

4.2 Backgrounds

4.2.1 Wavelet

The wavelet transform can be used for multi-scale analysis of a signal through dilation and translation, so it can extract time-frequency variables of a signal effectively. For orthonormal bases, the scaling and wavelet functions are selected as:

$$\phi_{L,k}(t) = 2^{L/2}\phi(2^L t - k), \quad \mathrm{L}, k \in \mathrm{Z} \tag{37}$$

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad j \geq L \; and \; j, k \in \mathrm{Z} \tag{38}$$

Approximation and detail of a signal can be constructed as an orthonormal basis and then the signal function is:

$$f(t) = \sum_k a_{L,k}\phi_{L,k}(t) + \sum_{j \geq L}\sum_k d_{j,k}\psi_{j,k}(t) \tag{39}$$

where $k \in$ all the possible integer values, $f \in L^2(\Re)$

The coefficients can be obtained by the following equations:

$$
\begin{aligned}
a_{L,k} &= \int_{\Re} f(t)\phi_{L,k}(t)dt \\
d_{j,k} &= \int_{\Re} f(t)\psi_{j,k}(t)dt
\end{aligned}
\qquad (40)
$$

where $a_{L,k}$ is the coarse level coefficient (described as the smoother signal) and $d_{L,k}$ is the finer level coefficient (described as the finer signal).

Let the data set $\mathbf{y} = [y(t_1), y(t_2), \cdots, y(t_N)]^{\mathrm{T}}$ which come from the signal $f(t)$ at time $t_i, i = 1, \cdots, N$. Then, the discrete wavelet transform of $\mathbf{y}$ is described as:

$$\mathbf{d} = \mathbf{Wy} \qquad (41)$$

where $\mathbf{W}$ is the orthonormal DWT matrix $N \times N$ and $\mathbf{d} = (a_L, d_{L,} d_{L+1}, \cdots, d_J)$ are $N \times 1$ wavelet coefficient vector. The wavelet coefficient is described as approximations (a_L) and details (d_L) of signals that are determined to 2^{L-1}.

4.2.2 Support Vector Machine

The main goal of SVM is to determine a hyperplane which minimizes the empirical classification error by maximizing the distance (i.e., margin) between the separating hyperplane and the data (Vapnik 1995). In SVM, input data are first mapped into a high-dimensional feature space where an optimal decision function can be obtained. As shown in Figure 13, an optimally-separating hyperplane is found which maximizes the margin.

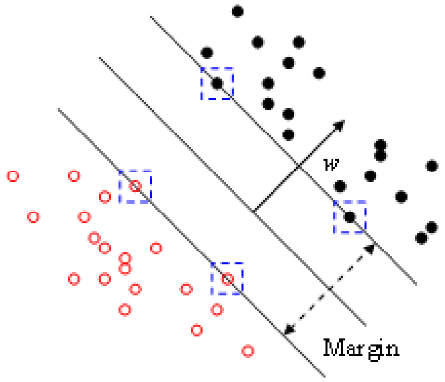


Figure 13: An Illustration of SVM for Two-Class Separation

This decision function satisfies inequality constraints

$$y_i(\mathbf{w}\Phi(\mathbf{x}_i)+b)-1\geq 0\ \forall_i\,. \qquad (42)$$

The optimal decision function is obtained by minimizing $1/2\|\mathbf{w}\|^2$ with constraints (6). Non-separable problems are solved by introducing ξ_i and Lagrangian

$$L=1/2\|\mathbf{w}\|^2+C\sum\xi_i-\sum\alpha_i[y_i(\mathbf{w}\Phi(\mathbf{x}_i)+b)-1+\xi_i]-\sum\mu_i\xi_i\,. \qquad (43)$$

Instead of this quadratic programming problem, a corresponding dual problem is preferred because it is easier to solve, which is given by

$$L_d=\sum\alpha_i-1/2\sum\alpha_i\alpha_j y_i y_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)\,. \qquad (44)$$

The solution is obtained as $\mathbf{w}=\sum\alpha_i y_i\Phi(\mathbf{x}_i)$, where this calculation is executed for support vectors with $\alpha_i>0$. Here dot products can be replaced with a kernel function (called a kernel trick) $K(\mathbf{x}_i,\mathbf{x}_j)=\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$ (Müller *et al*. 2001). Training SVM is to find α_i, *b*, and support vectors with given kernel function parameters and *C*. The use of a kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$ allows the computation of dot products in a nonlinear feature space *F*, without the use of nonlinear mappings. By replacing canonical (Euclidean) dot products in F by a kernel function, the execution of the nonlinear mappings and the dot products in F becomes unnecessary. Commonly used kernel functions include a radial basis function (RBF) $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \sigma)$ and polynomial $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$ functions.

4.2.3 Support Vector Machine Recursive Feature Elimination

The variable selection method of SVM-RFE (Guyon *et al.* 2002) is an application of a recursive feature elimination based on sensitivity analysis for an appropriately defined cost function. In the linear kernel case, define a cost function $J = (1/2)\|\mathbf{w}\|$. Then the least sensitive variable, which has the minimum magnitude of the weight, is eliminated first. This eliminated variable becomes ranking *n*. The machine is retrained without the eliminated variable and removes the variable with the minimum magnitude of weights. This eliminated variable becomes ranking *n*-1. By doing this process repeatedly until no variable is left, I can rank the variables.

Given training instances $\mathbf{X}_{all} = [\mathbf{x}_1, \ldots, \mathbf{x}_l]'$ with class labels $\mathbf{y} = [y_1, \ldots, y_l]'$, initialize the subset of variables $\mathbf{s} = [1, 2, \ldots, n]$ and $\mathbf{r} = []$. For the linear kernel case, repeat (i) through (v) until **s** becomes an empty array:

(i) Construct new training instances $\mathbf{X} = \mathbf{X}_{all}(:, \mathbf{s})$

(ii) Train SVM(**X**, **y**) to obtain *g*(**x**)

(iii) Compute the gradient $\mathbf{w} = \nabla g(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i$

(iv) Find the variable *f* with the smallest $\mathbf{w}$, $f = \arg\min(|\mathbf{w}|)$

(v) Update **r** and eliminate the variable *f* from **s**: $\mathbf{r} = [\mathbf{s}(f), \mathbf{r}], \mathbf{s} = \mathbf{s} - \{\mathbf{s}(f)\}$.

For general kernel cases, let us define a cost function

$$J = (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e} \qquad (45)$$

where $\boldsymbol{\alpha}$ is a vector with Lagrange multipliers, $\mathbf{H}_{hk} = y_h y_k K(\mathbf{x}_h, \mathbf{x}_k)$, and **e** is an l dimensional vector of ones. To compute the change in *J* caused by the removal of the variable *i*, it is assumed that there is no change in **α**. Thus

$$\mathbf{H}(-i)_{hk} = y_h y_k K(\mathbf{x}_h(-i), \mathbf{x}_k(-i)) \qquad (46)$$

where (-*i*) indicates that the variable *i* has been removed. As a result, the sensitivity function is given by

$$\begin{aligned} DJ(i) &= J - J(-i) \\ &= (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}(-i)\boldsymbol{\alpha} \end{aligned} \qquad (47)$$

The SVM-RFE algorithm for general kernels is to repeat (i) through (v) until **s** becomes an empty array:

(i) Construct new training instances $\mathbf{X} = \mathbf{X}_{all}(:, \mathbf{s})$

(ii) Train SVM(**X**, **y**) to obtain α

(iii) Compute the ranking criterion $DJ(i) = (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}(-i)\boldsymbol{\alpha}$

(iv) Find the variable *f* such that $f = \arg\min_i DJ(i)$

(v) Update **r** and eliminate the variable *f* from **s**: $\mathbf{r} = [\mathbf{s}(f), \mathbf{r}], \mathbf{s} = \mathbf{s} - \{\mathbf{s}(f)\}$.

4.3 A Two-Stage Classification Procedure for Spectral Data

This section proposes a two-stage scheme for the classification of spectral data. The proposed method incorporates wavelet-based preprocessing with SVM gradient-based variable selection. Figure 14 shows a schematic of the proposed two-stage framework. As shown in Figure 14, it can be decomposed into two stages, namely, a multi-scale vertical-energy-thresholding (MSVET)-based data reduction and SVM gradient-recursive feature elimination (Gradient-RFE)-based variable selection.

The proposed MSVET wavelet transformation can extract useful information in compressed wavelet coefficients and thus can be used to perform noise suppression and pre-processing of spectral data effectively. A schematic diagram of the proposed MSVET method is shown in figure 15. The proposed SVM gradient-RFE variable selection enables us to identify an optimal subset of compressed wavelet coefficients so that the classifier constructed with the selected wavelet coefficients minimizes the classification error. In addition, the use of this concentrated information, instead of the full spectra, for the classification of high-dimensional spectral data is expected to improve computational speed significantly.

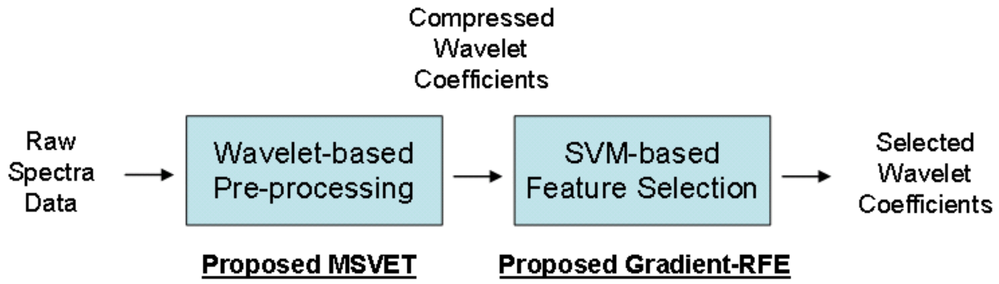


Figure 14: A Schematic of the Proposed Two-Stage Framework

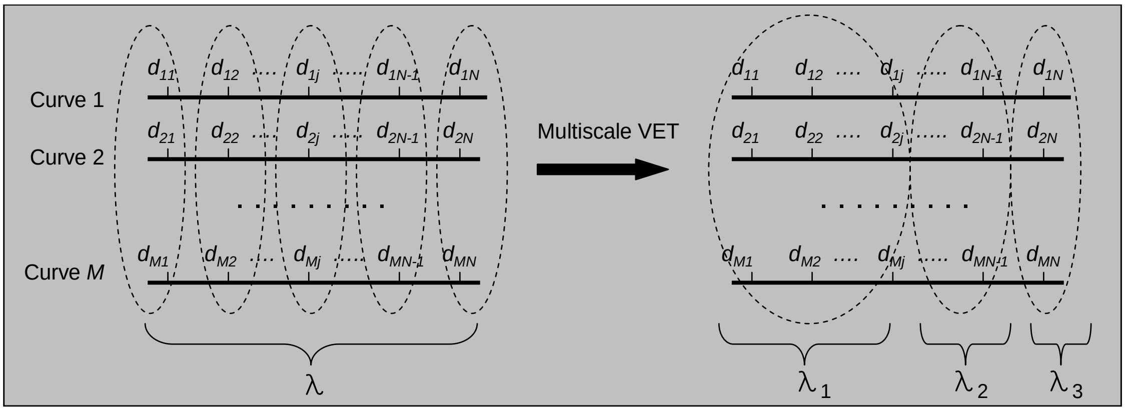

Curve 1
Curve 2
Curve M
d11 d12 .... d1j d1N-1 d1N
d21 d22 .... d2j d2N-1 d2N
dM1 dM2 .... dMj ..... dMN-1 dMN
λ
Multiscale VET
λ1
λ2
λ3

Figure 15: A Schematic Diagram of the Proposed MSVET Method

4.3.1 Multi-Scale VET-Based Wavelet

Wavelet thresholding methods are powerful tools for de-noising (Donoho 1995). The objective of these methods is to estimate a wide class of functions in smoothness spaces from noisy data. The wavelet technique is effective because the energy of a smooth function is often concentrated on few coefficients while the energy of noise is still spread over all coefficients in the wavelet domain.

For the given *M* spectral data, the vertical energy of each wavelet coefficient can be defined by

$$\left\|\mathbf{d}_{vm}\right\|^2 = d_{1m}^2 + d_{2m}^2 + \ . \ . \ . \ . \ . \ + d_{Mm}^2 \,. \qquad (48)$$

The original VET method of Jung *et al*. (2006) minimizes the overall relative reconstruction error (ORRE) to determine the threshold value λ

$$ORRE(\lambda) = \frac{\sum_{m=1}^{N} E[\|\mathbf{d}_{vm}(1 - I(\|\mathbf{d}_{vm}\|^2 > \lambda))\|^2}{\sum_{m=1}^{N} E[\|\mathbf{d}_{vm}\|^2]} + \frac{\sum_{m=1}^{N} E[\| I(\|\mathbf{d}_{vm}\|^2 > \lambda)\|^2}{N} \qquad (49)$$

However, the VET procedure does not consider the scale information of wavelets and a fixed threshold value is obtained for each scale even though each scale may include different types of information for further decision-making. A multi-scale vertical energy thresholding (MSVET) procedure is proposed. It has a different optimal thresholding value for each scale by extending the idea of the VET procedure. In the MSVET procedure, the multi-scale overall relative reconstruction error (MSORRE) is defined as follows:

$$MSORRE(\lambda_L,\ \lambda_{L+1}.\ .\ .\ ,\ \lambda_J) = \frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\left\|\mathbf{d}_{ij}(1-I(\left\|\mathbf{d}_{ij}\right\|^2 > \lambda_i))\right\|^2\right\}}{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\left\|\mathbf{d}_{ij}\right\|^2\right\}} + \xi\frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{I(\left\|\mathbf{d}_{ij}\right\|^2 > \lambda_i)\right\}}{N-2^{J-L}} \tag{50}$$

Here $\mathbf{d}_{ij} = (d_{ij1}, d_{ij2}, ..., d_{ijM})$ and $\left\|\mathbf{d}_{ij}\right\|^2 = d_{ij1}^2 + d_{ij2}^2 + \ldots + d_{ijM}^2$, where d_{ijM} represents the wavelet coefficient at the j^{th} wavelet position of the i^{th} scale for the M^{th} curve. The following Lemma 1 shows that an optimal threshold level for each scale depends on the vertical energy of each scale of the signal. The MSVET is compared with existing wavelet thresholding methods in Appendix A1, and the robustness of MSVET is shown in Appendix A2.

Lemma 1

The objective function $MSORRE(\lambda_i)$, $i = L, \ldots, J$, is minimized uniquely at $\lambda_i = \lambda_i^*$ where

$$\lambda_i^* = \frac{\sum_{j=1}^{2^{i+J-2L}} E\left\{\left\|\mathbf{d}_{ij}\right\|^2\right\}}{2^{i+J-2L}} \tag{51}$$

Proof of Lemma 1

Denote

$$\Lambda_{i\,j}(\lambda_i) = E\left\{\left\| I(\|\mathbf{d}_{ij}\|^2 < \lambda_i)\mathbf{d}_{ij}\right\|^2\right\} = E\left\{I(\|\mathbf{d}_{ij}\|^2 < \lambda_i)\|\mathbf{d}_{ij}\|^2\right\} = E\left\{I(Y_{ij} < \lambda_i)y_{ij}\right\} = \int_0^{\lambda_i} y_{ij}\, f_{ij}(y_{ij})dy_{ij}$$

$$\boldsymbol{\Psi}_{ij}(\lambda_i) = E\left\{I(\|\mathbf{d}_{ij}\|^2 > \lambda_i)\right\} = \Pr\left\{\|\mathbf{d}_{ij}\|^2 > \lambda_i\right\} = \Pr\left\{Y_{ij} > \lambda_i\right\} = 1 - \int_0^{\lambda_i} f_{ij}(y_{ij})dy_{ij}$$

where, $f_{ij}(y_{ij})$ is a non-central chi-square density of Y_{ij}. Then, the first term in the formula of MSORRE can be represented as

$$\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\left\|\mathbf{d}_{ij}(1 - I(\|\mathbf{d}_{ij}\|^2 > \lambda_i))\right\|^2\right\} = \sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\|\mathbf{d}_{ij}\|^2\, I(\|\mathbf{d}_{ij}\|^2 < \lambda_i)\right\} = \sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} \Lambda_{ij}(\lambda_i)$$

And the MSORRE can be rewritten as follows:

$$MSORRE(\lambda_L,\ \lambda_{L+1}.\ .\ .\ ,\ \lambda_J) = \frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\left\|\mathbf{d}_{ij}(1 - I(\|\mathbf{d}_{ij}\|^2 > \lambda_i))\right\|^2\right\}}{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\|\mathbf{d}_{ij}\|^2\right\}} + \xi\frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{I(\|\mathbf{d}_{ij}\|^2 > \lambda_i)\right\}}{N - 2^{J-L}}$$

$$= \frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} \Lambda_{ij}(\lambda_i)}{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} E\left\{\|\mathbf{d}_{ij}\|^2\right\}} + \xi\frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L} \boldsymbol{\Psi}_{ij}(\lambda_i)}{N - 2^{J-L}}$$

Because

$$\frac{\partial\boldsymbol{\Psi}ij(\lambda_i)}{\partial\lambda_i} = \frac{\partial\left(1 - \int_0^{\lambda_i} f_{ij}(y_{ij})dy_{ij}\right)}{\partial\lambda_i} = -f_{ij}(\partial\lambda_i) < 0$$

and

$$\frac{\partial\Lambda_{ij}(\lambda_i)}{\partial\lambda_i} = \frac{\partial\left(\int_0^{\lambda_i} y_{ij}\, f_{ij}(y_{ij})dy_{ij}\right)}{\partial\lambda_i} = \lambda_i\, f_{ij}(\lambda_i) = -\lambda_i\,\frac{\partial\boldsymbol{\Psi}_{ij}(\lambda_i)}{\partial\lambda_i}.$$

The thresholding value can be obtained from the following property give by

$$\frac{\partial MSORRE(\lambda_L,\ \lambda_{L+1}.\ .\ .\ ,\ \lambda_J)}{\partial \lambda_i}$$

$$= -\lambda_i \left(\sum_{i=L}^{J} \sum_{j=1}^{i+J-2L} \frac{\partial \Psi_{ij}(\lambda_i)}{\partial \lambda_i} \right) \left(\frac{1}{\sum_{i=L}^{J} \sum_{j=1}^{i+J-2L} E\left\{ \left\| \mathbf{d}_{ij} \right\|^2 \right\} \frac{\partial \Psi_{ij}(\lambda_i)}{\partial \lambda_i}} \right)$$

$$+ \frac{1}{N-2^{J-L}} \left(\sum_{i=L}^{J} \sum_{j=1}^{i+J-2L} \frac{\partial \Psi_{ij}(\lambda_i)}{\partial \lambda_i} \right)$$

$$= \left(-\frac{\lambda_i}{\sum_{i=L}^{J} \sum_{j=1}^{i+J-2L} E\left\{ \left\| \mathbf{d}_{ij} \right\|^2 \right\}} + \frac{1}{N-2^{J-L}} \right) \left(\sum_{i=L}^{J} \sum_{j=1}^{i+J-2L} \frac{\partial \Psi_{ij}(\lambda_i)}{\partial \lambda_i} \right) = 0$$

if and only if $\lambda_i = \frac{1}{N-2^{J-L}} \sum_{i=L}^{J} \sum_{j=1}^{i+J-2L} E\left\{ \left\| \mathbf{d}_{ij} \right\|^2 \right\}.$

4.3.2 SVM Gradient-RFE Variable Selection

The SVM Gradient-RFE combines two existing variable selection methods: SVM-RFE and SVM Gradient (Guyon *et al.* 2002; Hermes and Buhmann 2000). The new method has the merits of these two methods so it should be competitive to SVM-RFE in terms of prediction accuracy while maintaining speedy computation. The SVM Gradient-RFE uses the gradient for variable selection criteria, but in order to give a ranking for all the variables, the machines are trained using all the variables, and then the variable with a minimum angle is eliminated. The ranking of this eliminated variable then becomes *n*. The machine is then trained without the eliminated variable, and the variable with the minimum selection criterion is eliminated. This eliminated variable becomes ranking n-1. By recursively eliminating all the variables, one ranks the variables.

The time complexity of the SVM gradient-RFE algorithm can be analyzed as follows. The training time complexity of SVM is known to be $O(\max(l,n)\ \min(l,n)^2)$, where l is the number of samples and n is the number of variables (Chapelle 2007). The computation in each iteration of the algorithm is dominated by the step (ii) SVM training, which has $O(\max(l,n)\ \min(l,n)^2)$ time complexity. Suppose n $<l$. Then the total time complexity of the algorithm is $ln^2 + l(n\text{-}1)^2 + ... + l2^2 + l1^2 = O(ln^3)$. Similarly, I have $O(l^2n^2)$ for the other case. One can combine these two cases using the min operator and this leads to the time complexity of $O(ln^2 \min(l,n))$ for the SVM gradient-RFE algorithm. The SVM gradient-RFE algorithm can be summarized as follows. Given training instances $\mathbf{X}_{all} = [\mathbf{x}_1, \ . \ . \ . \ , \ \mathbf{x}_l]^{\mathrm{T}}$ with class labels $\mathbf{y} = [y_1, \ . \ . \ . \ , \ y_l]^{\mathrm{T}}$, initialize the subset of variables $\mathbf{s} = [1, 2, \ldots, n]$ and $\mathbf{r}$=[]. For a given kernel function, repeat (i) through (vii) until $\mathbf{s}$ becomes an empty array:

(i) Encode training instances as $\mathbf{X} = \mathbf{X}_{all}(:, \mathbf{s})$

(ii) Train SVM($\mathbf{X}$,$\mathbf{y}$) to obtain $g(\mathbf{x})$

(iii) Compute the gradient $\nabla g(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i \nabla_{\mathbf{x}} K(\mathbf{x}_i, \mathbf{x}), \ \ \forall \mathbf{x} \in SV$

(iv) Compute the sum of angles between $\nabla g(\mathbf{x})$ and $\mathbf{e}_j$, γ_j, j=1, …, $|\mathbf{s}|$

$$\gamma_j = \sum_{\mathbf{x} \in SV} \angle(\nabla g(\mathbf{x}), \ \ \mathbf{e}_j)$$

where $\angle(\nabla g(\mathbf{x}), \mathbf{e}_j) = \min_{\beta \in \{0,1\}} \left\{ \beta\pi + (-1)^{\beta} \arccos\left(\frac{\langle \nabla g(\mathbf{x}) \cdot \mathbf{e}_j \rangle}{\|\nabla g(\mathbf{x})\|} \right) \right\}$

(v) Compute the averages of the sum of the angles

$$c_j = 1 - \frac{2}{\pi} \cdot \frac{\gamma_j}{|SV|}$$

(vi) Find the variable *f* with the smallest c_j, j=1, … , $|\mathbf{s}|$: $f = \arg\min(c_j)$

Update **r** and eliminate the variable *f* from **s**: $\mathbf{r} = [\mathbf{s}(f), \mathbf{r}], \mathbf{s} = \mathbf{s} - \{\mathbf{s}(f)\}$.

4.4 Results

In this section the proposed framework is demonstrated using four NIR data sets. The four datasets are chosen to evaluate the classification accuracy and computational efficiency of the proposed method. For this purpose, the four datasets are divided into two groups: two popular public datasets with relatively small number of variables and the high-dimensional datasets obtained from real problem. The classification performance of the proposed two-stage classification method is compared with those of four traditional one-stage methods including an operations research-based (OR-based) variable selection (Fung and Mangasarian 2004). SVM-RFE, gradient-RFE with RBF kernels, linear kernel-based method and the OR-based method are used for comparison. These methods are different from the proposed two-stage framework in that all the variables available are used for classification without multi-scale VET-based preprocessing. Specifically, the idea of the OR-based variable selection is to suppress input space variables using a fast Newton method and linear programming formulation. Variables are ranked by the magnitudes of the coefficients of the linear decision function obtained from the method.

In addition, computational efficiency is evaluated based on run time needed for each of the methods.

4.4.1 NIR Data and Implementation

The real NIR data were obtained from two wood products (referred to as Example 1 and Example 2) and were collected to determine whether each of the two wood products (Douglas-fir and Spruce) were treated by a specific, proprietary preservative. This is a two-class classification problem where the preservative of interest should be distinguished from other competitors with similar ingredients (Taylor and Lloyd 2007). The public data of Example 3 measure absorbance of finely chopped meat samples (Ferraty and Vieu 2006). The NIR data of Example 3 are divided into two classes based on a fat content of a meat sample smaller or larger than 20%. Example 4 represents wheat NIR data, which are divided into two classes, namely, low (<14.5%) and high (>14.5%) moisture content (Kalivas 1997).

Multi-scale VET wavelet was implemented using MATLAB (The MathWorks Inc., Natick, MA) and WaveLab version 8.02. For the implementation of the proposed SVM-based gradient-RFE method a SVM MATLAB toolbox was used, which is available at http://www.isis.ecs.soton.ac.uk/resources/svminfo/. All computations for this study were done on an IBM compatible PC with an Intel Pentium IV CPU running at 3.6 GHz with 1GB RAM. The family of Symmlet-8 was used for all NIR spectra.

4.4.2 Results of Two Real Data Sets

The Douglas-fir NIR data consisted of 360 instances with 2,151 wavelengths. The data is randomly divided into training (288 observations) and test (72 observations) data sets. The data sets were treated by padding the original signal into the nearest dyadic length, which are done by zero or linear padding depending on the problem of interest. Thus prior to performing multi-scale VET-based wavelet analysis, a zero padding was applied to the Douglas-fir data so that 2,151 wavelengths were reduced to 2,048 (=2^{11}).

The Douglas-fir NIR data were preprocessed by applying the proposed multi-scale VET-based wavelet for a compression and de-noising of the high-dimension data. Figure 16 shows the comparisons of the original Douglas-fir NIR data and the reconstructed ones using the multi-scale VET procedure that uses only 410 wavelet coefficients among 2048 wavelet coefficients. As can be seen in this figure, the reconstruction seems to be quite successful because the reconstructed data can capture most of important patterns such as peaks and valleys of spectra.

Table 15 shows classification accuracy for the Douglas-fir data using the five different methods. Here the bold numbers underlined represent the maximum accuracy for each column. Overall, the best classification accuracy is obtained from the proposed method: 99% classification accuracy using only 37 variables. This implies that the set of 37 variables is enough for the classification and the other non-selected variables are either irrelevant or redundant. On the other hand, the other methods to be compared have lower classification accuracy even though they used more variables. For example, the gradient-RFE method achieved 96% classification accuracy using 139 variables, whilst the SVM-RFE method achieved 93% accuracy with 156 variables.

The overall pattern of the five methods in classification accuracy is shown in Figure 17. The proposed two-stage method outperformed the other four methods no matter the number of variables used. When the same number of variables as the proposed method, i.e., 37, is selected, the other methods yield lower classification accuracy. The proposed method also is more efficient computationally than the other methods: the computing time of the proposed method was 83.36 seconds, whilst gradient-RFE required 140.09 seconds and SVM-RFE 2.78 hours.

The Spruce NIR data of Example 2 consisted of 240 instances with 2,151 wavelengths, which were randomly split into 180×2,151 for training and 60×2,151 for test. As in Example 1, prior to performing multi-scale VET-based wavelet analysis, the 2,151 wavelengths were reduced to 2,048 ($=2^{11}$) by a zero padding. By applying the proposed multi-scale VET-based wavelet, a total of 560 wavelet coefficients were selected out of 2,048.

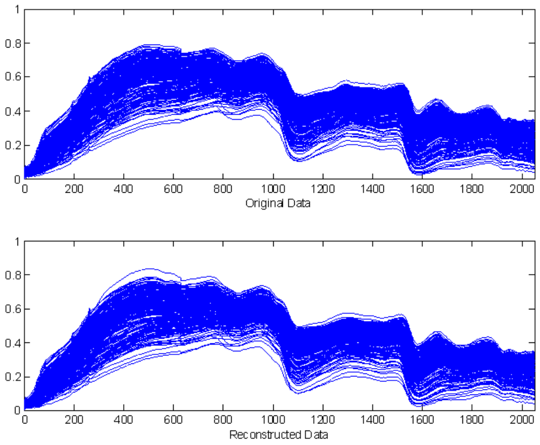


Figure 16: Original versus Reconstructed Data for Example 1

Table 15: Classification Accuracy for Example 1

Number of Variables	Linear Kernel	SVM-RFE	Gradient-RFE	OR-Based	Proposed Method
1	0.50	0.67	0.64	0.72	0.68
2	0.50	0.69	0.71	0.71	0.81
3	0.50	0.82	0.76	0.72	0.86
4	0.50	0.82	0.79	0.72	0.85
37	0.85	0.88	0.93	0.69	**0.99**
50	0.83	0.88	0.93	0.69	0.97
100	0.86	0.90	0.92	0.93	0.97
111	**0.89**	0.90	0.92	0.93	0.97
115	0.88	0.90	0.92	0.92	0.97
139	0.89	0.90	**0.96**	**0.96**	0.97
156	0.88	**0.93**	0.94	0.93	0.97
200	0.88	0.89	0.92	0.92	0.97

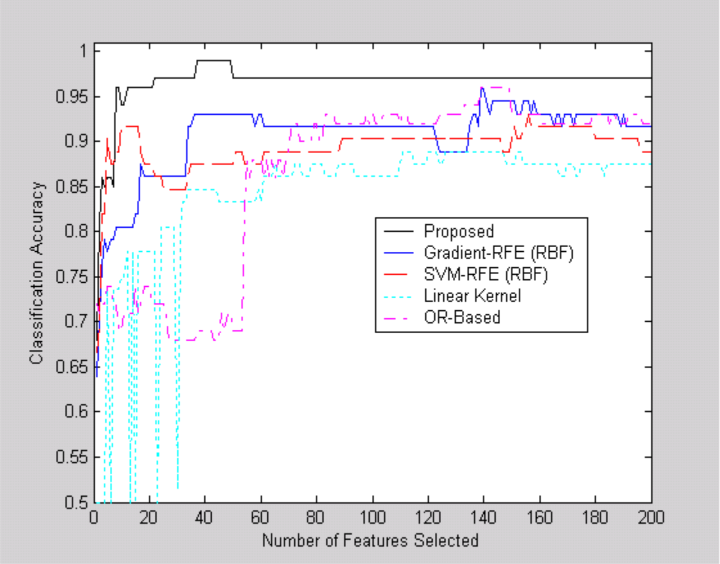


Figure 17: A Plot for Classification Accuracy for Example 1

The performance comparison in terms of classification accuracy for the test data of Example 2 is shown in Figure 18 and Table 16. The best classification accuracy was obtained from the proposed method: 100% classification accuracy with 14 variables. The other methods provided performance comparable to the proposed method. The linear kernel and Gradient-RFE achieved 97% classification accuracy using 12 and 15 variables, respectively. In case of SVM-RFE, however, 100 variables were required to obtain similar performance. A comparison of computational time showed that the proposed method (30.67 seconds) was more efficient that the others, especially the SVM-RFE method that required 0.38 hours. Figure 19 shows the comparisons of the original NIR data and the reconstructed ones. As shown in Figure 19, the reconstruction from the multi-scale VET method seems to approximate the original Spruce NIR spectra quite well.

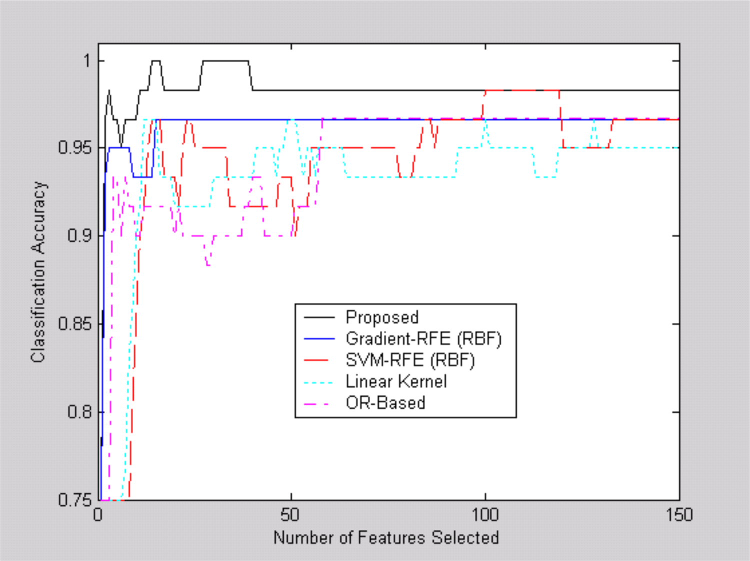


Figure 18: A Plot for Classification Accuracy for Example 2

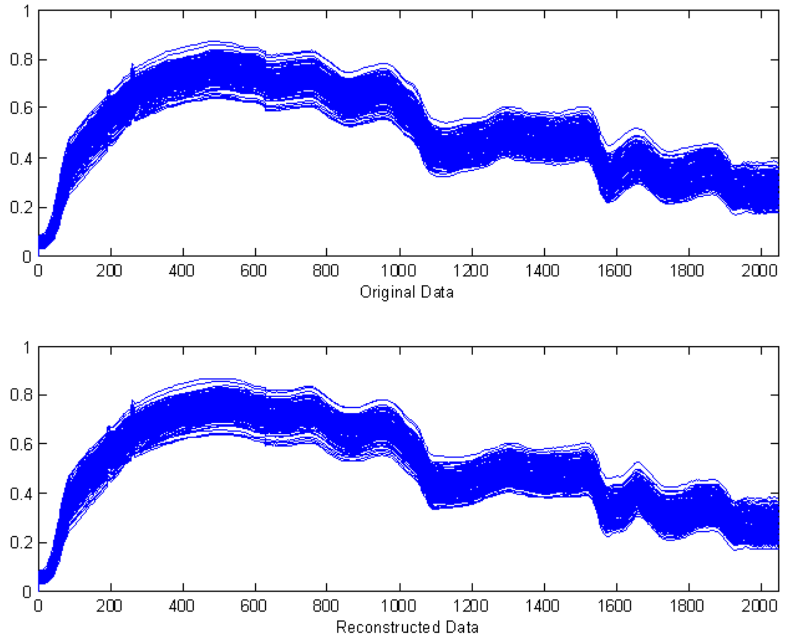


Figure 19: Original versus Reconstructed Data for Example 2

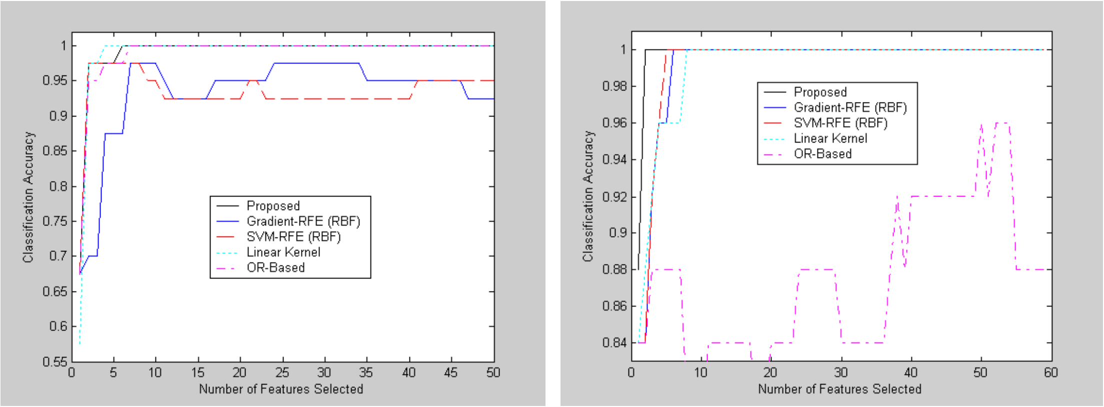
Table 16: Classification Accuracy for Example 2

Number of Variables	Linear Kernel	SVM-RFE	Gradient-RFE	OR-Based	Proposed Method
1	0.75	0.75	0.75	0.75	0.75
2	0.75	0.75	0.93	0.75	0.97
3	0.75	0.75	0.95	0.75	0.98
4	0.75	0.75	0.95	0.93	0.97
12	**0.97**	0.92	0.93	0.92	0.98
14	0.97	0.97	0.93	0.92	**1.00**
15	0.97	0.97	**0.97**	0.92	1.00
50	0.97	0.93	0.97	0.90	0.98
58	0.95	0.95	0.97	**0.97**	0.98
100	0.97	**0.98**	0.97	0.97	0.98
150	0.95	0.97	0.97	0.97	0.98

4.4.3 Results from the Two Public Data Sets

The fat NIR data of Example 3 had 215 instances with 100 wavelengths, which were divided randomly into 175 for training and 40 for test. In this case, a linear padding was applied to the original signal so that their wavelengths are increased to the nearest dyadic length (i.e., $128=2^7$). The wheat NIR data of Example 4 had 100 instances with 701 wavelengths and is randomly divided into 75 for training and 25 for test. A zero padding was applied to the original spectra data so that 701 wavelengths were reduced to 512 ($=2^9$). For the two public data sets, the proposed multi-scale VET-based wavelet was performed, selecting a total of 63 and 59 wavelet coefficients in Example 3 and Example 4, respectively.

The classification performance of the five methods is shown in Table 17 and Figure 20. In case of Example 3, the best classification accuracy (i.e., 100%) was achieved by three methods: linear kernel using 4 variables, OR-based using 7 variables and the proposed method using 6 variables. The SVM-RFE and Gradient-RFE gave lower classification accuracy. The good performance of the linear kernel method comparable to that of the proposed method can be explained by investigating the original spectral curves of Example 3. As shown in Figure 21, there were few peaks and valleys in these curves, in contrast to the other examples. It should be noted that the proposed method is preferred because its computational time (16.75 seconds) is much less than the linear kernel method (276.59 seconds). Similar results were obtained in Example 4; the proposed method achieved the best performance with fewer variables and less computational time. The improved performance of the proposed two-stage method can be explained by comparing plots of original vs. reconstructed data shown in Figure 21. Only a small number of wavelet coefficients were required to capture most of the patterns in the NIR data sets.



(a) (b)

Figure 20: Classification Accuracy Plots for (a) Example 3 (b) Example 4

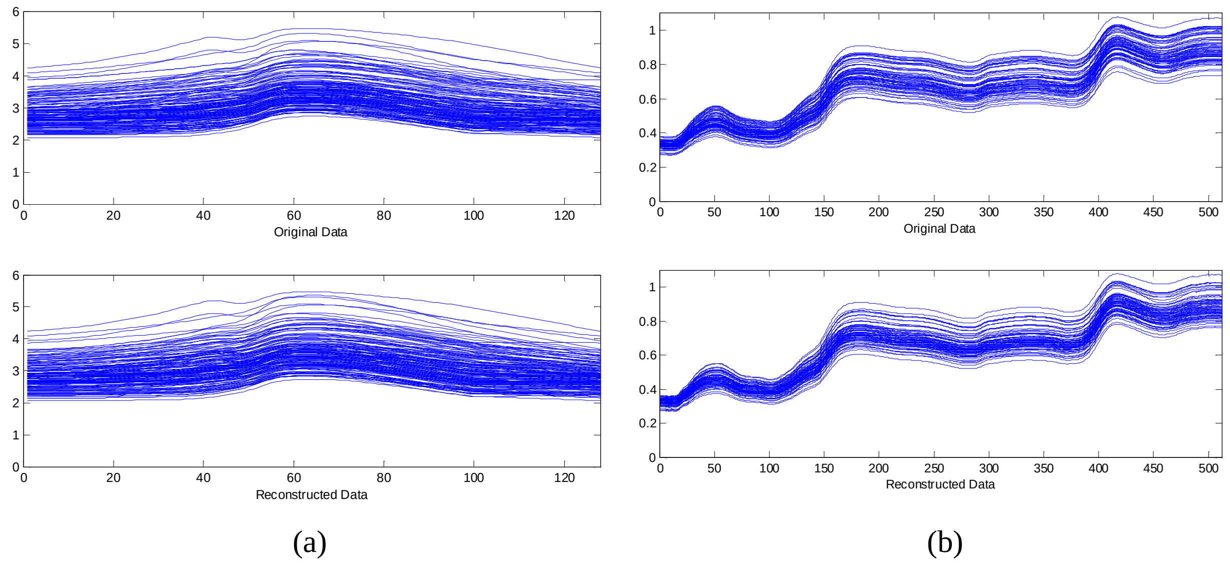


Figure 21: Original versus Reconstructed Data for (a) Example 3 (b) Example 4

Table 17: Classification Accuracy for Two Public Data Sets.

Number of Variables		Linear Kernel	SVM-RFE	Gradient-RFE	OR-Based	Proposed Method
Example 3	1	0.58	0.68	0.68	0.68	0.68
	2	0.98	**0.98**	0.70	0.95	0.98
	3	0.98	0.98	0.70	0.95	0.98
	4	**1.00**	0.98	0.88	0.98	0.98
	5	1.00	0.98	0.88	0.98	0.98
	6	1.00	0.98	0.88	0.98	**1.00**
	7	1.00	0.98	**0.98**	**1.00**	1.00
	8	1.00	0.98	0.98	1.00	1.00
Example 4	1	0.84	0.84	0.84	0.84	0.88
	2	0.88	0.84	0.84	0.84	**1.00**
	3	0.92	0.92	0.92	0.88	1.00
	4	0.96	0.96	0.96	0.88	1.00
	5	0.96	**1.00**	0.96	0.88	1.00
	6	0.96	1.00	**1.00**	0.88	1.00
	7	0.96	1.00	1.00	0.88	1.00
	8	**1.00**	1.00	1.00	0.88	1.00
	50	1.00	1.00	1.00	**0.96**	1.00

Chapter 5 Perception-Decision-Cognition Methodology for Discriminant Analysis Based on Human Decision-Making Process

5.1 Introduction

Data mining procedures are based on statistical principles and machine learning theory, creatively integrated to effect and facilitate the identification of significant informative patterns for a given database. Recurrent strategies used in data mining include preprocessing, data partitioning, machine learning (modeling), and validation. The ultimate goal of these procedures is the disclosure of unknown and valuable information. Hand *et al*. (2000) have discussed several models and patterns.

As indicated by Meisel and Mattfeld (2007), operations research and data mining are complementary and supportive due to three facts: (i) operations research techniques expedite the efficiency of data mining; (ii) data mining methodologies enlarge the scope of operations research applications; and (iii) integration of both data mining and operations research boost systems performance. Furthermore, the key element that allows effective fusion of both areas is the use of optimization algorithms (with particular emphasis on search procedures) to find an accurate model and to develop metaheuristics. An example of such procedures is the search algorithm by Olafsson *et al*. (2008) to find the best variable subset.

Discriminant analysis methods, based on several types of algorithm, have been proposed to find successful models for complicated data in an extensive range of application domains. The objective of discriminant analysis is to identify groups of observations, based on the input variables, which minimize the within-group variability and maximize the between-group variability. Recently, not only the discriminant analysis area but also other supervised or unsupervised learning areas have faced two challenging issues: (i) dimensionality; and (ii) nonlinearity. Several researchers developed new discriminant analysis techniques for preventing problems of high-dimensionality: spectral regression discriminant analysis (Cai *et al*. 2008), automatic non-parameter uncorrelated discriminant analysis (Yang *et al*. 2007), high-dimensional discriminant analysis (Bouveyron *et al*. 2007), and, for avoiding problems of nonlinearity: adaptive nonlinear discriminant analysis (Kim *et al*. 2006), kernel Fisher discriminant analysis (Mika 2002) and support vector machines for classification and regression (Vapnik 1995).

Variable selection is an important area of research in machine learning, pattern recognition, statistics, and related fields. The key idea of variable selection is to find input variables which have predictive information and to eliminate non-informative variables. The use of variable selection techniques is motivated by three reasons: (i) to improve discriminant power; (ii) to find fast and cost-effective variables; and (iii) to reach a better understanding of the application process (Guyon and Elisseeff 2003). In the case of high-dimension data, variable selection plays a crucial role because of four challenges (Theodoridis and Koutroumbas 2006): (i) a large set of variables; (ii) existence of irrelevant variables; (iii) presence of redundant variables; and (iv) data noise.

This chapter proposes a novel methodology based on the human decision-making process to perform three steps known as perception, decision, and cognition. For this reason, the proposed procedure will be referred to as a *Perception-Decision-Cognition Methodology* (PDCM). The main idea of this methodology is to emulate a biological thinking process by integrating both optimal search and data mining procedures. The perception step includes five different dimension reduction methods, based on wavelets, to transform the original data into a representation form that exhibits orthogonality and low noise. The decision step uses information complexity to find informative variables which can be used to identify groups based on prior modeling information. The cognition step recognizes the best model based on the support vector machines for classification, a well-known kernel-based statistical data mining approach. Three numerical experiments were run to compare PDCM to other often-used procedures. The results from the experiments show that the proposed method outperforms all the other procedures tested.

5.2 Wavelet-Based Dimension Reduction Techniques

Dimension reduction is a preferred strategy in the area of machine learning. As anticipated, there are several approaches to perform dimensional reduction. The following methods are among the most popular: principal component analysis (Jolliffe 2002), rotational linear discriminant analysis technique (Sharma and Paliwal 2008), independent component analysis (Stone 2004), semi-definite embedding (Weinberger and Saul 2006), multifactor dimensionality reduction (Ritchie and Motsinger 2005), factor analysis (Basilevsky 1994), and wavelet-based dimension reduction (Donoho and Jonston 1994; Chang and Vidakovic 2002; Jung *et al*. 2006; Cho *et al*. 2009).

The dimension reduction strategy has important benefits that can be measured not only in terms of computational time savings, but also in accuracy improvement. In the new PDCM, the wavelet-based dimension reduction is applied in Step 1. The wavelets approach was selected because of several attributes, among which the following two are most relevant: (a) wavelets adapt effectively to spatial variables of a function such as discontinuities and varying frequency behavior; (b) wavelets have efficient $O(n)$ algorithms to do transformations (Mallat 1999). Wavelet-based techniques are applied to obtain a well-fitted reduced-dimension representation of the original data.

The fitness of the representations can be observed in Figure 22. In this figure, the first curve corresponds to the original data and the remaining five to the following wavelet-based techniques: VisuShrinkUnion, VisuShrinkIntersect, VertiShrink, VET (Vertical Energy Thresholding), and MSVET (Multi-Scale Vertical Energy Thresholding). The wavelet-based techniques are compared in Appendix A1 and checked the robustness in Appendix A2.

Discrete Wavelet Transformation (DWT) is often used for dimension reduction (also known as shrinkage or threshold). Let $\mathbf{y}_m = [y_{m1}, y_{m2}, \cdots, y_{mN}]^{\mathrm{T}}$ is an m^{th} observed sample. For a single sample, the DWT procedure uses the orthonormal matrix $\mathbf{W}$ of dimension $N \times N$ to find the wavelet coefficient

$$\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \cdots, \mathbf{d}_J) \tag{52}$$

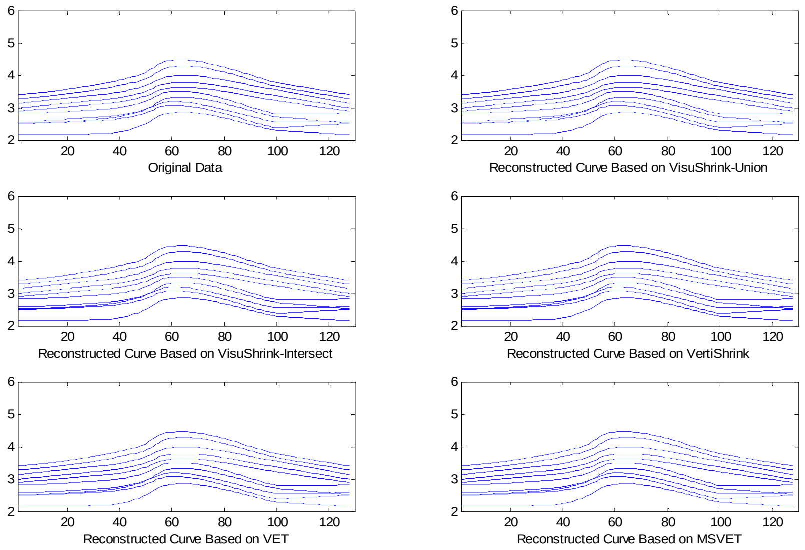


Figure 22: Original and Reconstructed Data Curves

where

$$\mathbf{c}_L = (c_{L0}, \ldots, c_{L2^L-1}), \mathbf{d}_L = (d_{L0}, \ldots, d_{L2^L-1}), \ldots, \mathbf{d}_J = (d_{J0}, \ldots, d_{J2^J-1})$$

through the transformation

$$\mathbf{d} = \mathbf{W}\mathbf{y}.$$

For multiple samples, let vector $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M]$ be the data set with M observed samples. The wavelet coefficient vector is obtained from the transformation

$$\mathbf{D} = \mathbf{W}\mathbf{Y} \tag{53}$$

where $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M]$, and $\mathbf{d}_m = (\mathbf{c}_{mL}, \mathbf{d}_{mL}, \mathbf{d}_{mL+1}, \cdots, \mathbf{d}_{mJ})$, $m = 1, 2, \ldots, M$.

Small absolute values of wavelet coefficients are undesirable since they may be influenced more by noise than by information. On the other hand, large absolute values are more influenced by information than noise. This observation motivates the development of threshold methods. There are two threshold rules usually referred to as *soft* and *hard* thresholds. The soft rule is a continuous function of the data that shrinks each observation, while the hard rule retains unchanged only large observations (Donoho and Johnston 1994). The hard and soft threshold methods are defined as following:

$$D(U,\lambda) = \mathrm{sgn}(U)\max(0,|U|-\lambda) \quad \text{(Soft)}$$
$$D(U,\lambda) = \begin{Bmatrix} U, & \text{all } |U| > \lambda \\ 0, & \text{otherwise} \end{Bmatrix} \quad \text{(Hard)} \tag{54}$$

where λ is the threshold value. The threshold method can be used not only for data reduction but also for de-noising.

5.2.1 VisuShrink (VS)

VisuShrink is a soft thresholding technique that applies a universal threshold proposed by Donoho and Jonstone (1994). The VisuShrink threshold is given by $\sigma\sqrt{2\log N}$, where N is the number of wavelet coefficients, and σ is the standard deviation of the wavelet coefficients (or noise standard deviation). When ε_i is a white noise sequence, independent and identically distributed as $N(0,1)$, then as $N \to \infty$, $P\{\max|\varepsilon_i| > \sqrt{2\log N}\} \to 0$. That is, the maximum of the N values will most likely be smaller than the universal threshold. The VisuShrink guarantees a noise free reconstruction. However, when setting the threshold large, the degree of data fitting may be unsatisfactory. For multiple curves or samples, the VS procedure uses the union (VisuShrinkUnion, VSU) or intersection (VisuShrinkIntersection, VSI) of data sets in the selection of wavelet coefficients (Jung *et al*. 2006).

5.2.2 VertiShrink (VERTI)

Chang and Vidakovic (2002) developed a Stein-type shrinkage method, known as VertiShrink, to maximize the predictive density under appropriate model assumptions regarding wavelet coefficients. The main goal of VertShrink is the estimation of the baseline curve by using the average of block vertical coefficients. The estimated wavelet coefficients are given by:

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{M\sigma^2}{\mathbf{d}^{\mathrm{T}}\mathbf{d}}\right) + \mathbf{d} \tag{55}$$

where **d** is the wavelet coefficient, $\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \cdots, \mathbf{d}_J)$, M is the number of curves and σ is the standard deviation of the wavelet coefficients.

5.2.3 Vertical-Energy-Thresholding (VET)

VET was proposed by Jung *et al.* (2006). The procedure is based on the concept of *energy of a function* with some smoothness, since it is often concentrated on few coefficients, while the *energy of noise* is still spread over all coefficients in the wavelet domain. The *vertical* energy of wavelet coefficients is defined by

$$\| \mathbf{d}_{vj} \|^2 = d_{1j}^2 + d_{2j}^2 + ... + d_{Mj}^2 \tag{56}$$

where d_{mj} is the wavelet coefficient at the *j*th wavelet position for the *m*th data curve, $m = 1, 2, ..., M$.

The VET method minimizes the overall relative reconstruction error (*ORRE*), formulated below, to determine a threshold value, namely λ:

$$ORRE(\lambda)=\frac{\sum_{j=1}^{N}E\left[\|\mathbf{d}_{vj}(1-I(\|\mathbf{d}_{vj}\|^2>\lambda))\|^2\right]}{\sum_{j=1}^{N}E\left[\|\mathbf{d}_{vj}\|^2\right]}+\frac{\sum_{j=1}^{N}E\left[\|I(\|\mathbf{d}_{vj}\|^2>\lambda)\|^2\right]}{N} \tag{57}$$

5.2.4 MultiScale-Vertical-Energy-Thresholding (MSVET)

Since the VET procedure does not consider the scale information of wavelets, an improved procedure proposed by Cho *et al*. (2009) and known as multi-scale vertical energy thresholding (MSVET) obtains a different optimal thresholding value for each scale by extending the idea of the VET procedure. In the MSVET procedure, the multi-scale overall relative reconstruction error (*MSORRE*) is defined as follows to determine the threshold values, λ_i:

$$\begin{aligned}MSORRE(\lambda_L,\lambda_{L+1},\ldots,\lambda_J) &= \frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L}E\left[\|\mathbf{d}_{vji}(1-I(\|\mathbf{d}_{vji}\|^2>\lambda_i))\|^2\right]}{\sum_{j=1}^{N}E\left[\|\mathbf{d}_{vji}\|^2\right]}\\ &+\frac{\sum_{i=L}^{J}\sum_{j=1}^{i+J-2L}E\left[I(\|\mathbf{d}_{vji}\|^2>\lambda_i))\|^2\right]}{N-2^{J-L}}\end{aligned} \tag{58}$$

where $\mathbf{d}_{vji}=(d_{1ji},d_{2ji},\ldots,d_{Mji})$, $\|\mathbf{d}_{vji}\|^2=d_{1ji}^2+d_{2ji}^2+\ldots+d_{Mji}^2$; d_{mji} represents the wavelet coefficient at the j^{th} wavelet position of the i^{th} scale for the m^{th} curve, m = 1,2,…, M.

5.3 Variable Selection Based on Information Complexity and Recursive Feature Elimination

Once the reduced sample space is determined in Step 1, the decision regarding which of the remaining variables should be selected for ranking is made on the basis of minimal information complexity values, following the Information Complexity Performance Testing with Recursive Feature Elimination (*ICOMP$_{PERF}$-RFE*) procedure. This procedure essentially generates a smoothed covariance estimator to calculate the information complexity measure, and, finally performs ranking using recursive elimination on the remaining variables.

The development of information complexity for the discriminant analysis is evaluated using the modified maximal entropic complexity C_{1F}

$$C_{1F}(\hat{\mathbf{\Sigma}}) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a^2), \tag{59}$$

where $s = rank(\hat{\mathbf{\Sigma}})$, λ_j is the j^{th} eigenvalue of $\mathbf{\Sigma} > 0$, $j = 1,2,. . .,s$ and $\bar{\lambda}_a$ is arithmetic mean of the eigenvalues.

ICOMP$_{PERF}$ can be evaluated as indicated below:

$$ICOMP_{PERF} = n\log 2\pi + n\log(\hat{\sigma}^2) + n + 2C_{1F}(\hat{\mathbf{\Sigma}}_{STA\_CSE}) \tag{60}$$

where lack of fit is assessed by means of the first three terms and complexity by the fourth one. In the above expression, $\hat{\sigma}^2$ is the estimated mean squared error given by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and $\hat{\mathbf{\Sigma}}_{STA\_CSE}$ is the stabilized and smoothed *convex sum covariance matrix estimator* (Press 1975; Chen 1976) given by

$$\hat{\mathbf{\Sigma}}_{STA\_CSE} = \frac{n}{n+k}\hat{\mathbf{\Sigma}}_{STA} + \left(1 - \frac{n}{n+k}\right)\left[\frac{trace(\hat{\mathbf{\Sigma}}_{STA})}{h}\right]\mathbf{I}_h, \tag{61}$$

where $\hat{\mathbf{\Sigma}}_{STA}$ is the stabilized covariance matrix proposed by Thomaz (2004), *h* is the number of variables, $\mathbf{I}_h$ is *h*×*h* identity matrix, and *k* is chosen such that

$$0 < k < \frac{2[h(1+\beta)-2]}{h-\beta},$$

and

$$\beta = \frac{\left(tr(\hat{\mathbf{\Sigma}}_{STA})\right)^2}{tr(\hat{\mathbf{\Sigma}}_{STA}^2)}.$$

Specific details on this procedure are provided by Chapter 2.

5.4 Cognition Accuracy of Selected Models

When the ranking decision is finished in Step 2, the corresponding accuracies are determined using the corresponding cognition sets and the support vector machines (SVM) for classification described below. Once the accuracies are calculated for the selected models the most-accurate one is chosen.

The SVM finds an optimal separating hyperplane that maximizes the margin between the classes (Vapnik 1995). Consider the case of classifying a set of linearly separating data into two groups. Assume a set of training data is given by $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)]$, where $\mathbf{x}_i \in \Re^n$ is an input vector, $y_i \in \{-1,1\}$ is a binary class index, and *n* is the size of the training data set. Then, a decision boundary that partitions the underlying vector space into two classes can be represented by the hyperplane

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}+b=0 \tag{62}$$

where $\mathbf{w}$ is the weight vector and b is the bias. The objective of the SVM is to find a maximum margin decision boundary between the two parallel hyperplanes, $\mathbf{w}^{\mathrm{T}}\mathbf{x}+b=1$ and $\mathbf{w}^{\mathrm{T}}\mathbf{x}+b=-1$. The dual model with Lagrange multipliers of the corresponding primal model can be formulated as

$$\text{Max } \mathbf{Q}(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{63}$$

subject to

$$\sum_{i=1}^{n}\alpha_i\mathbf{x}_i=0,\ 0\le\alpha_i\le C,\ i=1,2,\ldots,n \tag{64}$$

where $K(\mathbf{x}_i,\mathbf{x}_j)$ is the kernel function and C is a predefined coefficient. Kernel functions used in the numerical experiments are described in Table 18.

The point $\mathbf{x}^o$ with coordinates corresponding to new data can be classified as indicated below:

$$\text{Class 1: } \sum_{i=1}^{n}\alpha_i^{ov}y_iK(\mathbf{x}_i,\mathbf{x}^o)+b^{ov}<0 \tag{65}$$

and

$$\text{Class 2: } \sum_{i=1}^{n}\alpha_i^{ov}y_iK(\mathbf{x}_i,\mathbf{x}^o)+b^{ov}>0 \tag{66}$$

where α^{ov} and b^{ov} are optimal values found based on the training data. A classification example based on the PDCM is illustrated in Figure 23.

Table 18: Used Kernel Functions

Kernel Function	$K(\mathbf{x}_i, \mathbf{x}_j)$	Parameters
Gaussian	$\exp\left[-\left(\frac{1}{a^b}\|\mathbf{x}_i-\mathbf{x}_j\|^2\right)^c\right]$	a=2, b=c=1
Cauchy	$\left(1+\frac{1}{a}\|\mathbf{x}_i-\mathbf{x}_j\|^2\right)^{-1}$	a=1
Inverse Multi-Quadratic	$\left(\|\mathbf{x}_i-\mathbf{x}_j\|^2+a^2\right)^{-1/2}$	a=1

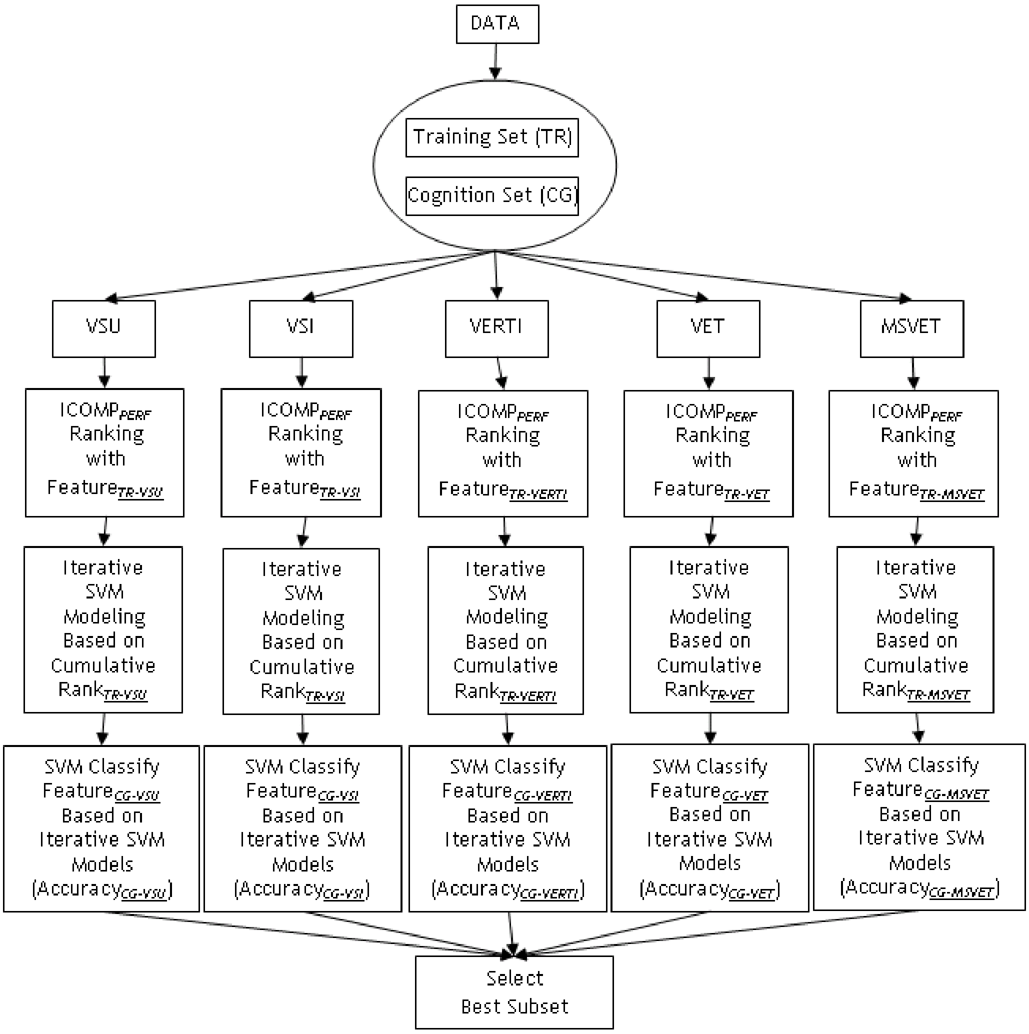

DATA
Training Set (TR)
Cognition Set (CG)
VSU
VSI
VERTI
VET
MSVET
ICOMP$_{PERF}$ Ranking with Feature$_{TR-VSU}$
ICOMP$_{PERF}$ Ranking with Feature$_{TR-VSI}$
ICOMP$_{PERF}$ Ranking with Feature$_{TR-VERTI}$
ICOMP$_{PERF}$ Ranking with Feature$_{TR-VET}$
ICOMP$_{PERF}$ Ranking with Feature$_{TR-MSVET}$
Iterative SVM Modeling Based on Cumulative Rank$_{TR-VSU}$
Iterative SVM Modeling Based on Cumulative Rank$_{TR-VSI}$
Iterative SVM Modeling Based on Cumulative Rank$_{TR-VERTI}$
Iterative SVM Modeling Based on Cumulative Rank$_{TR-VET}$
Iterative SVM Modeling Based on Cumulative Rank$_{TR-MSVET}$
SVM Classify Feature$_{CG-VSU}$ Based on Iterative SVM Models (Accuracy$_{CG-VSU}$)
SVM Classify Feature$_{CG-VSI}$ Based on Iterative SVM Models (Accuracy$_{CG-VSI}$)
SVM Classify Feature$_{CG-VERTI}$ Based on Iterative SVM Models (Accuracy$_{CG-VERTI}$)
SVM Classify Feature$_{CG-VET}$ Based on Iterative SVM Models (Accuracy$_{CG-VET}$)
SVM Classify Feature$_{CG-MSVET}$ Based on Iterative SVM Models (Accuracy$_{CG-MSVET}$)
Select Best Subset

Figure 23: Classification Example Based on PDCM with SVM

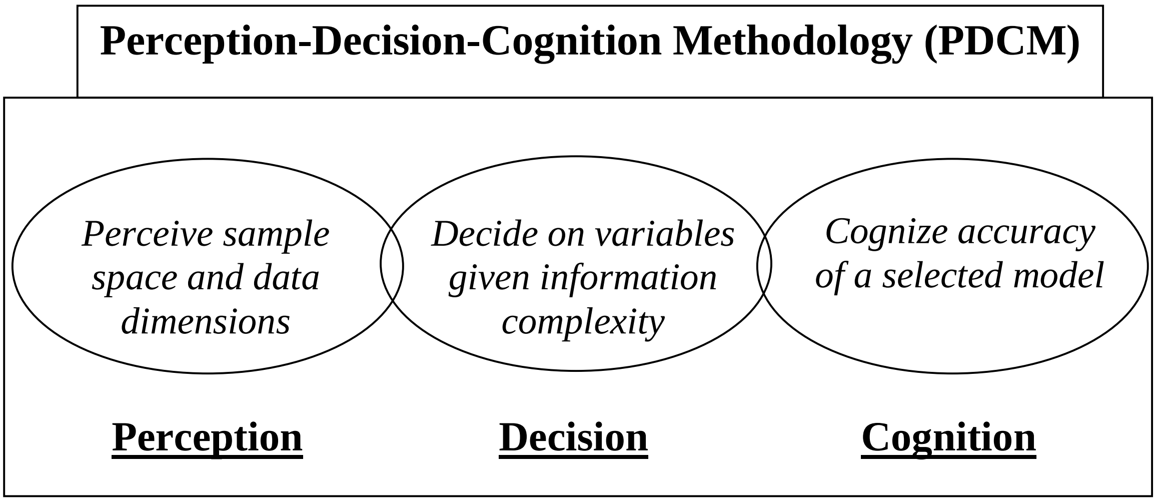

Perception-Decision-Cognition Methodology (PDCM)
Perceive sample space and data dimensions
Decide on variables given information complexity
Cognize accuracy of a selected model
Perception
Decision
Cognition

Figure 24: Conceptual View of the Perception-Decision-Cognition Methodology

5.5 Perception-Decision-Cognition Methodology (PDCM)

The proposed Perception-Decision-Cognition Methodology (PDCM) for discriminant analysis is conceptually represented in Figure 24. As indicated in this figure, it is analogous to a biological thinking process, which consists of three steps:

1. Perceive environmental information.
2. Decide on response (actions).
3. Recognize (evaluate) the accuracy of results to adjust the response.

The algorithm used by the PDCM consists of three steps conceptually described below, after assuming that all data have been classified according to three sets: training set, cognition set, and test set.

Step 1: Perceive Sample Space and Data Dimensions

Let the sample data be $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q)$, and the corresponding response be $\mathbf{y} = (y_1, y_2, ..., y_n)^{\mathrm{T}}$, where q is the dimension of **X** and n is the number of samples. Now apply all available dimension reduction techniques: VisuShrinkUnion, VisuShrinkIntersect, VertiShrink, VET, and MSVET.

For each dimension reduction technique, generate a new training set $\mathbf{X} = (\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_p)$, where the *reduced* dimension p is the number of coefficients *perceived* by the reduction techniques ($p \leq q$).

Step 2: Decide on Variables given Information Complexity

The procedure can be described as follows. Remove each of the *p* variables one at a time, and evaluate the corresponding information complexity measure, $ICOMP_{PERF}$. Once the *p-1* removal procedures are completed, the removed variable resulting in minimum value of $ICOMP_{PERF}$ is identified and assigned the lowest rank (i.e. *p*). This procedure is repeated for the remaining *p-1* variables for which there is no rank yet. As a result of this, a variable receives rank equal to *p-1*. This procedure is repeated until the *p* variables have been arranged according to their ranks.

Step 3: Recognize (evaluate) the Accuracy of Selected Models

Compute the accuracy value of each cognition data set using the SVM for all possible subsets of the ranked variables selected in Step 2. Specifically, first consider the variable with the highest rank (i.e, rank=1), and calculate the cognition accuracy value. After this, the two variables with rank=1 and rank=2 are considered, and a new cognition accuracy value is calculated. This procedure is repeated until all ranked variables are considered. Finally, the subset of variables resulting in the highest accuracy value is chosen as the best model.

5.6 Analysis of Proposed PDCM

In order to emphasize the effectiveness of the PDCM, it will be applied to three different data sets (heart imaging, fat content in meat, and handwritten) used for experiments with low-dimension/high sample size, and high-dimension/low sample size. For the wavelet transformation of the three data sets, the linear padding suggested by (Strang and Nguyen 1997) is applied. This article documents the comparison of the PDCM to the following procedures:

(a) SVM recursive feature elimination (SVM-RFE) (Guyon *et al*. 2002).

(b) Two-stage method (Cho *et al*. 2009).

(c) Several different ranking criteria with SVM: Kullback-Leibler distance (Theodoridis and Koutroumbas 2006); accuracies of K-nearest neighbor (KNN) classifier with k=1 (Hastie *et al*. 2001); absolute value of the u-statistic of a two-sample unpaired Wilcoxon test (Liao *et al*. 2007); absolute value two-sample *t*-test with pooled variance estimate (Zhu *et al*. 2003); Mahalanobis distance (Theodoridis and Koutroumbas 2006); Euclidean distance (Theodoridis and Koutroumbas 2006); and Bhattacharyya distance (Theodoridis and Koutroumbas 2006).

5.6.1 Heart Data (44 Variables)

The data set includes 267 samples and 44 variables on cardiac single proton emission computed tomography (SPECT) images with two categories, i.e., normal and abnormal (Cios and Kurgan 2001). The data set is divided into 80 samples as a training set, 13 samples as a cognition set, and 174 samples as a test set. Table 19 and Table 20 show comparison results in terms of the variables selected, the cognition accuracy, and the test accuracy. Cauchy and Inverse Multi-Quadratic kernel functions are used in Table 19 and Table 20, respectively.

Table 19: PDCM versus Various Ranking Based Method Using Cauchy

Methods	Selected Variables (# of Variables)	Cognition Accuracy	Test Accuracy
PDCM	21, 30, 33, 34, 36,…,51 (20)	84.62%	79.31%
SVM-RFE	8, 9, 14, 22, 26, 29, 30, 32, 35, 36 (10)	92.31%	77.01%
Two-Stage	27, 38, 50 (3)	84.62%	64.94%
Entropy	16, 26, 30, 40, 41, 42, 43, 44 (8)	92.31%	77.59%
KNN	4, 9, 10, 12, 26, 27, 30, 38, 41 (9)	84.62%	74.14%
Wilcoxon test	30, 40, 43 (3)	92.31%	72.41%
t-test	30, 40 (2)	84.62%	71.84%
Mahalanobis	30, 40 (2)	84.62%	71.84%
Euclidean	26, 30 (2)	92.31%	74.14%
Bhattacharyya	30, 40 (2)	84.62%	71.84%

Table 20: PDCM versus Various Ranking Based Method Using Inverse Multi-Quadratic

Methods	Selected Variables (# of Variables)	Cognition Accuracy	Test Accuracy
PDCM	1,…,6, 9, 10, 11, 13,…,20, 27, 31, 33, 35, 36, 39, 41, 43,…,47, 50 (31)	84.62%	74.71%
SVM-RFE	4, 6,…,9, 14, 32, 33, 36 (9)	100%	71.26%
Two-Stage	1, 2, 3, 5, 6, 13, 14, 15, 25,…,29, 31,…,44, 46, 48,…,52 (33)	76.92%	73.56%
Entropy	40 (1)	76.92%	61.49%
KNN	4, 9, 10, 12, 18, 26, 27, 30, 38, 39, 41, 42 (12)	84.62%	74.71%
Wilcoxon test	40 (1)	76.92%	61.49%
t-test	40 (1)	76.92%	61.49%
Mahalanobis	40 (1)	76.92%	61.49%
Euclidean	26, 30 (2)	84.62%	71.26%
Bhattacharyya	40 (1)	76.92%	61.49%

As observed in Table 19, PDCM achieves the same cognition accuracy of other methods, but it yields more accurate results in some cases. Also, as shown in Table 20, PDCM and KNN both reach the highest test accuracy, although KNN requires fewer variables.

5.6.2 Near Infrared Spectroscopy Data (100 Variables)

These data were collected by a Tecator infratec food and feed analyzer to predict the fat content of a meat sample based on near infrared (NIR) spectroscopy. The data set was divided into two classes defined on the basis of fat content; one class corresponded to 20% or less, and another class to more than this level (Rossi and Villa 2006). The entire data set consists of 215 samples with measured values for each of 100 predictive variables (wavelengths). These samples were divided randomly to configure a training set consisting of 108 samples; a cognition set consisting of 11 samples for cognition; and a test set consisting of the remaining 96 samples.

Table 21 and Table 22 show the results in terms of the variables selected, the cognition accuracy, and the test accuracy. Cauchy and Gaussian kernel functions are used in Table 21 and Table 22, respectively. Although both PDCM and Two-Stage reach the 100% accuracy level for the cognition set in Table 21, the test accuracy of PDCM is higher than that of Two-Stage and other methods. Additionally, Table 22 shows that the cognition accuracy of PDCM is 100% and the test accuracy is higher than that of other methods. Furthermore, PDCM uses only 3 and 5 variables to reach the accuracy levels previously mentioned.

Table 21: PDCM versus Various Ranking Based Method Using Cauchy

Methods	Selected Variables (# of Variables)	Cognition Accuracy	Test Accuracy
PDCM	1, 5, 9 (3)	100%	88.54%
SVM-RFE	1,…,17, 54, 55, 56 (20)	81.82%	62.5%
Two-Stage	1,…,45 (45)	100%	83.33%
Entropy	33,…,48, 54,…,100 (63)	81.82%	77.08%
KNN	1,…,12, 17, 18, 25, 26, 40,…,43, 53, 56,…,64, 72,…,75, 78,…,81 (38)	90.91%	87.5%
Wilcoxon test	24,…,100 (77)	90.91%	83.33%
t-test	24,…,100 (77)	90.91%	83.33%
Mahalanobis	24,…,100 (77)	90.91%	83.33%
Euclidean	24,…,100 (77)	90.91%	83.33%
Bhattacharyya	24,…,100 (77)	90.91%	83.33%

Table 22: PDCM versus Various Ranking Based Method Using Gaussian

Methods	Selected Variables (# of Variables)	Cognition Accuracy	Test Accuracy
PDCM	5, 11, 19, 20, 31 (5)	100%	89.58%
SVM-RFE	1,…,34, 46,…,100 (89)	90.91%	79.17%
Two-Stage	1, 2, 3, 7,…,45 (42)	90.91%	77.08%
Entropy	18,…,100 (83)	90.91%	83.33%
KNN	1, 2, 25, 41, 43, 57, 58, 62, 73, 80, 81 (11)	90.91%	85.42%
Wilcoxon test	1, 2, 4, 5, 6, 21,…,100 (85)	90.91%	83.33%
t-test	18,…,100 (83)	90.91%	83.33%
Mahalanobis	18,…,100 (83)	90.91%	83.33%
Euclidean	18,…,100 (83)	90.91%	83.33%
Bhattacharyya	18,…,100 (83)	90.91%	83.33%

5.6.3 Handwritten Data (240 Variables)

This data set has variables of handwritten numerals from 0 to 9 extracted from a collection of Dutch utility maps. The entire set consists of 200 samples digitized in binary images per class and six different variable sets (van Breukelen *et al*. 1998). One of six variable sets is used for the experiment; pixel averages in 2×3 windows. Two classes (0 and 1 in handwritten numerals) out of 10 classes are selected to verify the proposed method. Each class has 200 samples and only 100 out of 200 samples per class are included in the experimental data set. 20 samples are used as training set, 9 samples are used as cognition set and 171 samples are used as test set. Table 23 and Table 24 show comparison results in terms of the selected variables, the cognition accuracy, and the test accuracy.

Cauchy and Inverse Multi Quadratic kernel functions are used in Table 23 and Table 24, respectively. As seen in the tables, PDCM reaches a 100% cognition accuracy level as did the other methods, except SVM-RFE. Furthermore, PDCM achieves a higher accuracy level than the other methods for the test set.

Table 23: PDCM versus Various Ranking Based Method Using Cauchy

Methods	Selected Variables (# of Variables)	Cognition Accuracy	Test Accuracy
PDCM	19, 32, 33, 43 (4)	100%	99.42%
SVM-RFE	1,…,4, 8, 9, 14,…,17, 21,…,26, 29,…,32, 35, 36, 40,…,43, 45, 46, 50, 51, 55,…,59, 61, 64, 65, 66, 68, 69, 70, 73, 74, 76,…,80, 82,…,85, 87, 88, 89, 91,…,95, 97,…,100, 102, 103, 104, 106,…,109, 112,…,115, 117, 118, 119, 121,…,130, 132,…,149, 151,…,164, 167,…,185, 187, 188, 191, 192, 193, 195, 196, 198,…,201, 205,…,208, 210, 211, 213,…,222, 224,…,227, 229,…,240 (183)	88.89%	88.89%
Two-Stage	3, 18, 20, 27, 31, 32, 49,…,53, 63, 81, 82, 84,…,87, 89,…,93, 95, 97,…,101, 103, 104, 107, 108, 111, 113, 114, 117,…,120, 122, 123, 124, 126,…,132 (50)	100%	75.44%
Entropy	83 (1)	100%	95.91%
KNN	68, 82, 83 (3)	100%	94.15%
Wilcoxon test	67, 68, 82, 83 (4)	100%	88.89%
t-test	83 (1)	100%	95.91%
Mahalanobis	83 (1)	100%	95.91%
Euclidean	83 (1)	100%	95.91%
Bhattacharyya	83 (1)	100%	95.91%

Table 24: PDCM versus Various Ranking Based Method Using Inverse Multi-Quadratic

Methods	Selected Variables (# of Variables)	Cognition Accuracy	Test Accuracy
PDCM	9, 22, 33, 43, 56 (5)	100%	98.83%
SVM-RFE	162, 163, 164, 167,…,170, 172, 173, 174, 177, 178, 183, 184, 185, 191, 192, 193, 198, 199, 200, 205, 206, 207, 211, 214, 217, 218, 220, 221, 225, 226, 229, 234, 240 (35)	88.89%	95.32%
Two-Stage	18, 97, 120 (3)	100%	67.25%
Entropy	83 (1)	100%	96.49%
KNN	68, 82, 83 (3)	100%	95.32%
Wilcoxon test	67, 68, 82, 83 (4)	100%	92.4%
t-test	83 (1)	100%	96.49%
Mahalanobis	83 (1)	100%	96.49%
Euclidean	83 (1)	100%	96.49%
Bhattacharyya	83 (1)	100%	96.49%

Chapter 6 Summary and Conclusion

A novel SVM-*ICOMP*$_{PERF}$-RFE method is proposed using an information complexity (*ICOMP*$_{PERF}$) criterion in Chapter 2. SVM-RFE is used in conjunction with *ICOMP*$_{PERF}$ not only to choose an optimal kernel function from a portfolio of many other kernel functions, but also to select important subset(s) of variables. The numerical examples on two benchmark datasets show that the proposed hybridized method exhibits a promising performance for the variable subsetting and the optimal kernel selection. This method provides a unification of both *ICOMP*$_{PERF}$ as the variable selection criterion and RFE as the search algorithm. In this frame work, *ICOMP*$_{PERF}$ is a key cost function. Furthermore, the hybridized covariance matrix known as the stabilized and smoothed convex sum covariance estimator (STA-CSE) is used to avoid the singularity in the kernel based methods. In the literature related to recursive feature elimination such stabilization issues have not been addressed before. As shown in Tables 10, 11, 12, and 13, the comparisons of variable ranking methods demonstrate that SVM-*ICOMP*$_{PERF}$-RFE is a promising way to obtain the best subset of variables.

A new framework is proposed for assessing the cycle-lives of rechargeable batteries within shorter test times in Chapter 3. In such a framework, the dual variables FSVM is proposed and proved to give excellent performance in screening for the purpose of qualification, even with a sizable reduction in the test time. Also, a boosting algorithm can be applied to improve the performance of the proposed algorithm under the environment of small sample sizes.

A novel two-stage classification scheme for high-dimensional spectral data that combines MSVET-based wavelet preprocessing and SVM gradient-based variable selection is proposed in Chapter 4. It is demonstrated using four NIR data sets that the proposed two-stage method has higher computational efficiency due to its effective pre-processing of spectral data. In addition, the proposed two-stage method produced significantly better classification performance than SVM-RFE, Gradient-RFE, and OR-based methods. This is tested by paired t-test for each of the datasets with the results from SVM-RFE (p-value=0.048), Gradient-RFE (p-value=0.011), linear kernel method (p-value=0.061), and OR-based (p-value=0.047). It is attributed to the fact that the proposed method incorporates wavelet-based preprocessing with SVM gradient-based variable selection. The proposed method would also be beneficial to other spectral signals such as mid infrared (MIR) and nuclear magnetic resonance (NMR), to compress high-dimension data and select useful variables in wavelet domain.

The development and application of a new Perception-Decision-Cognition Methodology (PDCM) for discriminant analysis, based on the human decision-making process is documented in Chapter 5. Five different wavelet-based dimension reduction techniques are applied in the perception step. It is shown that the procedure yields a good representation of the original data, using only reduced variables. The decision step is performed using a rank-based variable selection approach, using the information complexity criterion. The information complexity-based variable selection approach shows a good ability to achieve reasonable variable ranks, which in turn can affect decision making. In the cognition step, the number of variables and accuracy are recognized for further discrimination. As supported by the numerical experiments documented in Chapter 5, the PDCM outperforms the currently available data mining approaches, and, furthermore, appears to be applicable to various areas, such as bioinformatics, chemometrics, pattern recognition, and other data mining fields. The PDCM has three advantages:

(i) Dimension simplification.

(ii) Multiple model choices based on simplified dimension.

(iii) Analogous to the biological process of human decision making.

LIST OF REFERENCES

LIST OF REFERENCES

Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821-837.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrov & B.F. Csaki (Eds.), Second International Symposium on Information Theory, (pp. 267-281). Academiai Kiado: Budapest.

Basilevsky, A. (1994). Statistical Factor Analysis and Related Methods: Theory and Application. Wiley, New York.

Bloehdorn, S., and Sure, Y. (2008). Kernel methods for mining instance data in ontologies. *Lecture Notes in Computer Science*, 4825:58-71.

Bloom, I., Cole, B.W., Sohn, J.J., Jones, S.A., Polzin, E.G., Battaglia, V.S., Henriksen, G.L., Motloch, C., Richardson, R., Unkelhaeuser, T., Ingersoll, U., and Case, H.L. (2001). An accelerated calendar and cycle life study of li-ion cells. *Journal of Power Sources*, 101:238-247.

Boser, B. E., Guyon, I. M., and Vapnik, V. (1992). A training algorithm for optimum margin classifiers. Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, ACM.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics-Theory and Methods*, 36:2607-2623.

Bozdogan, H. (1988a). ICOMP: A New Model-Selection Criterion. In Hans H. Bock (Ed.), Classification and Related Methods of Data Analysis. Amsterdam: Elsevier Science (North-Holland), 599-608.

Bozdogan, H. (1988b). The theory and applications of information-theoretic measure of complexity (ICOMP) as a new model selection criterion. Unpublished research report, the Institute of Statistical Mathematics, Tokyo, Japan, and the Department of Mathematics, University of Virginia, Charlottesville, VA, March 1988.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics Theory and Methods*, 19:221-278.

Bozdogan, H. (1994). Mixture-Model Cluster Analysis Using Model Selection Criteria and A New Informational Measure of Complexity. In H. Bozdogan (Ed.), Multivariate Statistical Modeling (Vol. 2), 69-113, Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach, Dordrecht: Kluwer Academic.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44:62-91.

Bradley, P., Mangasarian, O., and Street, W. (1998). Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10(2):209-217.

Broussely, M., and Archdale, G. (2004). Li-ion batteries and portable power source prospects for the next 5-10 years. *Journal of Power Sources*, 136:386-394.

Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston.

Cai, D., He, X., and Han, J. (2008). An efficient algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):1-12.

Chang, W., and Vidakovic, B. (2002). Wavelet estimation a baseline signal from repeated noisy measurements by vertical block shrinkage. *Computational Statistics and Data Analysis*, 40:317-328.

Chapelle, O. (2007). Training a support vector machines in the primal. *Neural Computation*, 19:1155-1178.

Chau, F. T., Gao, J. B., Shin, T. M., and Wang J. (1997). Compression of infrared spectral data using the fast wavelet transform method. *Applied Spectroscopy*, 51(5):649-659.

Chen, M. (1976). Estimation of covariance matrices under a quadratic loss function. Research Report S-46, Department of Mathematics, SUNY at Albany, Albany, N.Y., 1-33.

Cho, H., Baek, S. H., Youn, E., Jeong, M. K., and Taylor, A. (2009). A two-stage classification procedure for near-infrared spectra based on multi-scale vertical energy wavelet thresholding and SVM-based gradient-recursive feature elimination. *Journal of the Operational Research Society*, 60:1107-1115.

Cios, K.J., and Kurgan, L. (2001). Hybrid Inductive Machine Learning: An Overview of CLIP Algorithms. In: Jain L.C., and Kacprzyk J. (Eds). New Learning Paradigms in Soft Computing, Physica-Verlag, Springer.

Cristianini, N., and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, New York.

de Boor, C. (2001). A Practical Guide to Splines, Springer-Verlag, New York.

Donoho, D.L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613-627.

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

Donoho, D.L., and Johnstone, I.M. (1995). Adapting to unknown smoothness in wavelet shrinkage. *Journal of The American Statistical Association,* 90:1200-1224.

Ferraty, F., and Vieu, P. (2003). Curves discriminations: a nonparametric functional approach. *Computational Statistics and Data Analysis,* 44(1-2):161-173.

Ferraty, F., and Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.

Fröhlich, H. (2002). Feature selection for support vector machines by means of genetic algorithms. Diploma Thesis in Computer Science, University of Tübingen, Germany.

Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. Academic, New York.

Fung, G., and Mangasarian, O. (2004). A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications,* 28:185-202.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research,* 3(1):1157-1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1/3):389-422.

Hand, D., Blunt, G., and Kelly, M. (2000). Data mining for fun and profit. *Statistical Science*, 15(2):111-126.

Harris, C. J. (1978). An Information Theoretic Approach to Estimation. In: M. J. Gregson, Recent Theoretical Developments in Control, Academic Press, London, 563-590.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Herbrich, R. (2002). Learning Kernel Classifier: Theory and Algorithms. MIT Press, Cambridge.

Hermes, L. and Buhmann, J. (2000). Feature selection for support vector machines. Proceedings of 15th International Conference on Pattern Recognitio*n*, 2:716-719.

Howe, J.A. and Bozdogan, H. (2010). Regularized SVM Classification with Information Complexity and the Genetic Algorithm. In H. Bozdogan (Ed.), Multivariate Statistical Modeling in High-Dimensions. to appear.

Jank W., and Shmueli, G., (2006). Functional data analysis in electronic commerce research. *Statistical Science*, 21(2):155-166.

Jin, J., and Shi, J. (1999). Feature preserving data compression of stamping tonnage information using Wavelets. *Technometrics*, 41(4):327-339.

Jin, J., and Shi, J. (2001). Automatic feature extraction of waveform signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing*, 12:257-268.

Johnson, B.A., and White, R.E. (1998). Characterization of commercially available lithium-ion batteries. *Journal of Power Sources*, 70:48-54.

Jolliffe, I.T. (2002). Principal Component Analysis. Springer, New York.

Jung, U., Jeong, M. K., and Lu, J. C. (2006). A vertical-energy-thresholding procedure for data reduction with multiple complex curves. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 36(5):1128-1138.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37:255-259.

Kim, H., Drake, B.L., Park, H. (2006). Adaptive nonlinear discriminant analysis by regularized minimum squared errors. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):603-612.

Koller, D., and Sahami, M. (1996). Toward optimal feature selection. Proceedings of the 13th International Conference on Machine Learnin*g*, Bari, Italy, 284-292.

Kullback, S. (1997). Information Theory and Statistics. Dover, New York.

Kullback, S., and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79-86.

Liao, C., Li, S., and Luo, Z. (2007). Gene selection using Wilcoxon rank sum test and support vector machine for cancer classification. *Lecture Notes in Computer Science*, 4456:57-66.

Maklad, M.S., and Nichols, T. (1980). A New Approach to Model Structure Discrimination. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-10(2):78-84.

Mallat, S. (1999). A Wavelet Tour of Signal Processing. Academic Press, San Diego.

Mao, K.Z. (2004). Feature subset selection for support vector machines through discriminative function pruning analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 34:60-67.

Meisel, S., and Mattfeld, D.C. (2007). Synergies of data mining and operations research. Proceedings of the 40th Annual Hawaii International Conference on System Sciences, IEEE Computer Society, Washington D.C., 56.

Mika, S. (2002). Kernel fisher discriminants. Ph.D. dissertation, Technical University of Berlin, Berlin, Germany.

Mika, S., Ratsch, G., Weston, J., Schölkopf, B., and Muller, K. (1999). Fisher discriminant analysis with kernels. In Hu, Y., Larsen, J., Wilson, E., and Douglas, S., editors, *IEEE Neural Networks for Signal Processing LX*, 41-48.

Müller, K, Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12:181-201.

Ning, G., White, R.E., and Popov, B.N. (2006). A generalized cycle life model of rechargeable li-ion batteries. *Electrochimica Acta*, 51:2012-2022.

Olafsson, S., Li, X., and Wu, S. (2008). Operations research and data mining. *European Journal of Operational Research*, 187(3):1429-1448.

Pareto, V. (1909). *Manual of Political Economy*. New York: Kelley, 1971. English translation of the 1909 French edition of the 1906 Italian *Manuale d'economia politica con una introduzione alla scienza sociale*, Milan: Società Editrice Libraria, 1906.

Pearlman, J. (1986). Nuclear magnetic resonance spectral signatures of liquid crystals in human atheroma as basis for multi-dimensional digital imaging of atherosclerosis. Ph.D. dissertation, University of Virginia.

Press, S. (1975). Estimation of a normal covariance matrix, P-5436, Santa Monica, CA: The Rand Corporation.

Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(12):1119-1125.

Rakotomamonjy, A. (2003). Variable selection using SVM based criteria. *Journal of Machine Learning Research*, 3:1357–1370.

Ramsay, J., and Silverman, B. (2005). Functional Data Analysis. Springer-Verlag, New York.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, 14:1080-1100.

Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of Royal Statistical Society, Series B*, 49(3):223-239 and 252-265.

Rissanen, J. (1989). Stochastic Complexity in Statistical Inquiry. Teaneck, New Jersey: World Scientific Publishing Company.

Ritchie, M.D., and Motsinger, A.A. (2005). Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. *Pharmacogenomics*, 6(8):823-834.

Rossi, F., and Conan-Guez, B. (2005). Functional multi-layer perceptron: a nonlinear tool for functional data analysis. *Neural Networks*, 18(1):45-60.

Rossi, F., and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7-9):730-742.

Schölkopf, B., and Smola, A.J. (2002). Learning with Kernel. MIT Press, Cambridge.

Schölkopf, B., Smola A., and Müller K.R. (1999). Kernel Principal Component Analysis. Advances in Kernel Methods - Support Vector Learning, 327-352. (Eds.) Schölkopf, B., Burges, C.J.C., Smola, A.J., MIT Press, Cambridge, MA.

Saito, N. (1994). Simultaneous Noise Suppression and Signal Compression Using A Library of Orthonormal Bases and The Minimum Description Length Criterion in Wavelets in Geophysics. E. Foufoula-Georgiou and P. Kumar, Eds., Academic, New York, 299-324.

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

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Systems Technology Journal*, 27:379-423.

Sharma, A., and Paliwal, K.K. (2008). Rotational linear discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(10):1336-1347.

Shurygin, A. M. (1983). The linear combination of the simplest discriminantor and fisher's one. In Nauka (Ed.), *Applied Statistics*, Moscow, 114-158 (in Russian)

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10:262-266.

Sikha, G., Popov, B.N., and White, R.E. (2004). Effect of porosity on the capacity fade of a lithium-ion battery. *Journal of the Electrochemical Society*, 151(7):1104-1114.

Staszewski, W. J. (1998). Wavelet based compression and feature selection for vibration analysis. *Journal of Sound and Vibration*, 211(5):735-760.

Stone, J. (2004). Independent Component Analysis: A Tutorial Introduction. MIT press, London, England.

Strang, G., and Nguyen, T. (1997). Wavelets and Filter Banks. Wellesley-Cambridge Press.

Subramani, P., Sahu, R., and Verma, S. (2006). Feature selection using Haar wavelet power spectrum. *BMC Bioinformatics*, 7:432.

Taylor, A., and Lloyd, J. (2007). Potential of near infrared spectroscopy to quantify boron retention in treated wood. *Forest Products Journal*. 57(1/2):116-117.

Theodoridis, S., and Koutroumbas, K. (2006). Pattern Recognition. Academic Press, San Diego.

Thomaz, C. (2004). Maximum entropy covariance estimate for statistical pattern recognition, Ph.D. dissertation, University of London and for the Diploma of the Imperial College (D.I.C), London, UK.

USABC (1996). Electric vehicle battery test procedures manual, revision 2. Technical Report (US DOE/ID-10479).

van Breukelen, M., Duin, R.P.W., Tax, D.M.J., and den Hartog, J.E. (1998). Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381-386.

van Emden, M. (1971). An analysis of complexity, Number 35 in Mathematical Centre Tracts, Mathematisch Centrum, Amsterdam.

Vapnik, V. (1995). The Nature of Statistical Learning Theory, Springer-Verlag: New York.

Watanabe, S. (1985). Pattern Recognition: Human and Mechanical. Wiley, New York.

Weinberger, K.Q., and Saul, L.K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77-90.

Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research,* 3:1439–1461.

Wright, R.B., and Motloch, C.G. (2001). Cycle-life studies of advanced technology development program gen 1 lithium ion batteries. Technical Report (US DOE/ID-10845).

Yoshida, T., Takahashi, M., Morikawa, S., Ihara, C., Katsukawa, H., Shiratsuchi, T., and Yamaki, J. (2006). Degradation mechanism and life prediction of lithium-ion batteries. *Journal of the Electrochemical Society*, 153(3):576-582.

Youn, E. S. (2002). Feature selection in support vector machines, M.S. Thesis, University of Florida.

Zhou, W. (1998). Structured wavelet antenna signal modeling and random scale generalized linear model. PhD dissertation, North Carolina State University.

Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., and Kovach, J.S. (2003). Detection of cancer-specific markers amid massive mass spectral data. Proceedings of National Academy of Science, 100(25):14666-14671.

APPENDICES

APPENDICES

A1. Comparison of Wavelet Based Dimension Reduction Methods

Antenna data curves, tonnage signals and Mallat piecewise signals are used for comparison. The antenna data curves were collected at Nortel's production facility located in the Research Triangle Park, North Carolina. The testing equipment receives antenna signals at different degrees of azimuth and elevation. The antenna data set consist of 18 curves (Zhou 1998). The tonnage signals were collected from sheet-metal stamping processes which are known as very complicated and sensitive manufacturing processes. Recently, stamping tonnage sensors have been widely used to measure the stamping force for each stamped part in order to monitor the health of stamping processes (Jin and Shi 1999 & 2000). Mallat's piecewise signals characterize the combined pattern of transient signals with sharp changes and smooth signals at some parts (Mallat 1999). The six different measures are used for comparison of five wavelet based dimension reduction methods (VisuShrinkUnion, VisuShrinkIntersection, VertiShrink, VET, and MSVET). The comparison measures are as follows: (1) Relative Error: RE=$\sum_{i=1}^{M} || \mathbf{f}_i - \hat{\mathbf{f}}_\mathrm{i} || / \sum_{i=1}^{M} || \mathbf{f}_i ||$; (2) Reduction Ratio: RR= (1-*k*/*N*), where *k* is the number of selected positions; (3) ORRE (Overall Relative Reconstruction Error); (4) Approximate Minimum Description Length: AMDL(k)=$1.5kM \log_2 NM + 0.5NM \log_2 \sum_{i=1}^{M} || \mathbf{f}_i - \hat{\mathbf{f}}_i ||$ which is close to the Akaike information criterion used in model selection for the regression problem (Antoniadis *et* *al*. 1997); (5) L_2 Error= $\sum_{i=1}^{M} \| \mathbf{f}_i - \hat{\mathbf{f}}_i \|$ which is the root mean square error; (6) L_∞ Error= $\max_{i,j} | f_{i,j} - \hat{f}_{i,j} |$ which is the maximum error where, $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$.

Figure 25 shows the original antenna curves and the reconstructed ones using the wavelet based methods. As shown in figure, the reconstruction curves of MSVET are very similar to original ones. The reconstructed curves capture the patterns in peaks and valleys reasonably. Table 25 presents results of wavelet based methods with different comparison measures. The data reduction ratio of MSVET is 61.72% and the reconstructed curves of MSVET in the Figure 25 are reasonably reconstructed as similar as original curves in terms of capturing the patterns in peaks and valleys, although the RE, RR and ORRE of VET are smaller than the MSVET. Moreover, the MSVET has the smallest AMDL. Figure 26 shows 24 tonnage curves under the normal conditions. We applied the 5 data reduction procedures to the tonnage signals. Table 26 presents results of wavelet based methods with different comparison measures. The relative error and the L_2 error of MSVET is very small, comparing to VET. Also, the reduction ratio of MSVET is 66% which is reasonable for reconstruction. The VET method has the largest reduction ratio and the smallest overall relative reconstruction error.

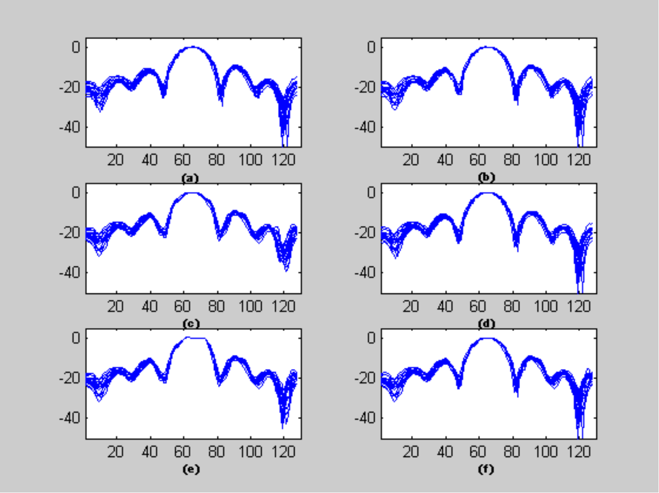


Figure 25: Antenna Data Curves

(a) Original (b) VisuShrink-Union (c) VisuShrink-Intersection
(d) VertiShrink (e) VET (f) MSVET

Table 25: Results for Antenna Curves (N = 128)

Methods	RE	RR	ORRE	AMDL	L_2 error	L_∞ error
VisuShrink-Union	0.0005566	0.64063	0.35993	2.3729e+004	19.3948	2.6610
VisuShrink-Intersection	0.0044363	0.75	0.25444	2.2956e+004	54.7550	16.5182
VertiShrink	5.3168e-005	0.34375	0.6563	3.1286e+004	5.9943	0.9324
VET	0.0035937	0.77344	0.23016	2.1701e+004	49.2814	10.3224
MSVET	0.00056168	0.61719	0.39572	1.4998e+004	19.4831	3.2759

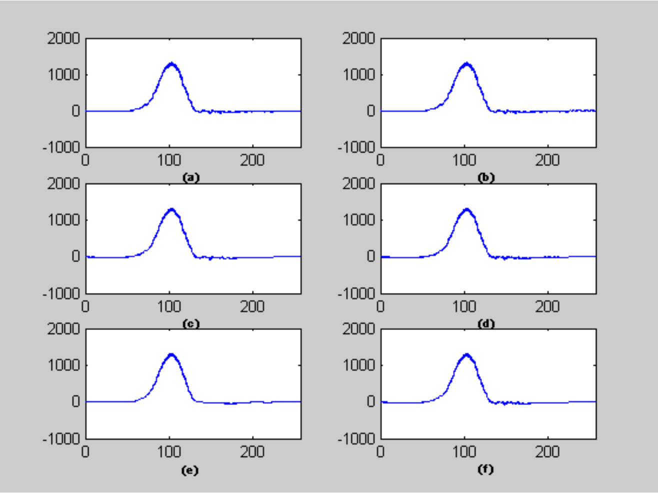


Figure 26: Tonnage Signals

(a) Original (b) VisuShrink-Union (c) VisuShrink-Intersection
(d) VertiShrink (e) VET (f) MSVET

Table 26: Results for Tonnage Signals (*N* = 256)

Methods	RE	RR	ORRE	AMDL	L_2 error	L_∞ error
VisuShrink-Union	6.9778e-007	0.12891	0.87109	1.3009e+005	26.5281	3.2870
VisuShrink-Intersection	0.00016532	0.77344	0.22673	7.9568e+004	408.3273	43.3440
VertiShrink	7.1448e-006	0.38672	0.61329	1.1050e+005	84.8870	7.8056
VET	0.0011886	0.93359	0.067595	6.9735e+004	1.0949e+003	69.0460
MSVET	8.9914e-005	0.66016	0.3509	7.5510e+004	301.1338	27.4553

Figure 27 shows the Mallat's piecewise signals with the combined pattern of transient signals with sharp and smooth changes and the reconstructed curves. The reconstructed curves are quite reasonable using 5 other methods. Table 27 presents results of wavelet based methods with different comparison measures. The relative error, L_2 error and L_∞ error of MSVET are smaller than the results of VET. In this case, the reduction ratio of MSVET is smaller than other methods: VisuShrinkUnion, VisuShrinkIntersection, and VET except VertiShrink. Since the Mallat signals have the sharp changes, it may affect the performance of the MSVET. When we compare the results of Table 25, 26, 27, MSVET has consistent reduction ratio (around 60%). In other words, MSVET may give promising results for complicated datasets which have high-dimensions and many peaks or shapes.

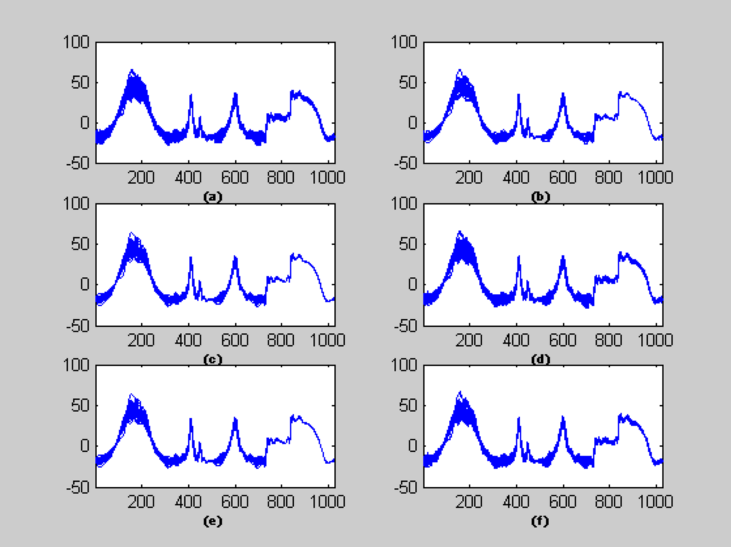


Figure 27: Mallat's Piecewise Signals

(a) Original (b) VisuShrink-Union (c) VisuShrink-Intersection
(d) VertiShrink (e) VET (f) MSVET

Table 27: Results for Mallat Piecewise Signals (*N* = 1024)

Methods	RE	RR	ORRE	AMDL	L_2 error	L_∞ error
VisuShrink-Union	0.0022	0.9141	0.0881	3.6018e+005	186.8799	3.6834
VisuShrink-Intersection	0.0023	0.9326	0.0697	3.4575e+005	192.4471	7.2698
VertiShrink	0.0009	0.4717	0.5293	7.1909e+005	121.6239	3.4212
VET	0.0026	0.9346	0.0681	3.4747e+005	204.8175	12.2439
MSVET	0.0013	0.5938	0.4140	5.6902e+005	144.6165	5.5863

A2. Robustness of Reduction Methods against Random Noises

Dimension reduction methods against random noises are tested for robustness in this section. For the experimental study, three noises with random normal are added to the signals and compared. Signal-to-noise-ratio (*SNR*) is defined as $\hat{\sigma}(\mathbf{f})\backslash\sigma$ where, $\hat{\sigma}(\mathbf{f})$ is the standard deviation of each signal points, and σ is the standard deviation of noise. Figure 28, 29, and 30 shows one original curve and 3 noise added curves. Table 28, 29, 30, and 31 provide relative error for model fitting and reduction ratio for dimension reduction using five wavelet based methods in the cases of *SNR* = 3, 15, and 30. Smaller *SNR* means that signals includes more noise. That is, noise level (σ) is large and a few of wavelet coefficients should be selected. As shown in tables, for all methods, the relative error with less noise (*SNR* = 30) is much smaller than one with more noise (*SNR* = 3). It means that it is difficult to find suitable model using complicated dataset with more noise. VET has better reduction ratio than MSVET, but relative error of MSVET is smaller than one of VET. Moreover, when curves or signals has more noise (*SNR*=3), MSVET mostly has smaller relative error than VisuShrinkUnion, VisuShrinkIntersection and VET. Even though, the signals have much noise, the reduction ratio of MSVET is similar to one of less noise. Consequently, MSVET has robustness for noise dataset.

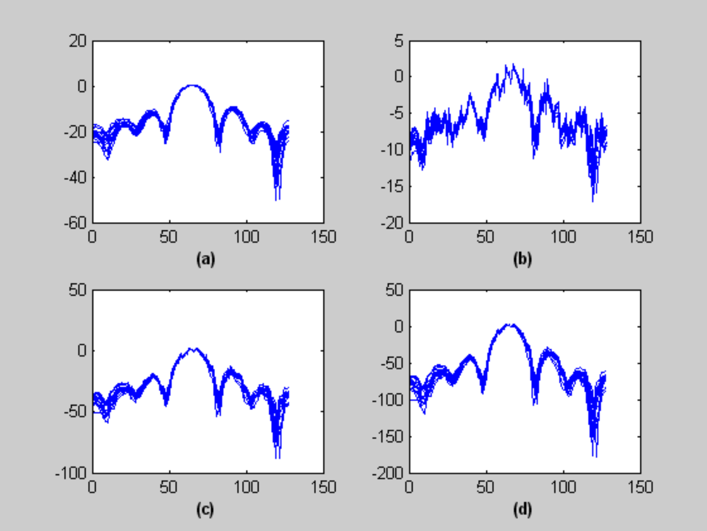


Figure 28: Antenna Data Curves

(a) Original (b) *SNR*=3 (c) *SNR*=15 (d) *SNR*=30

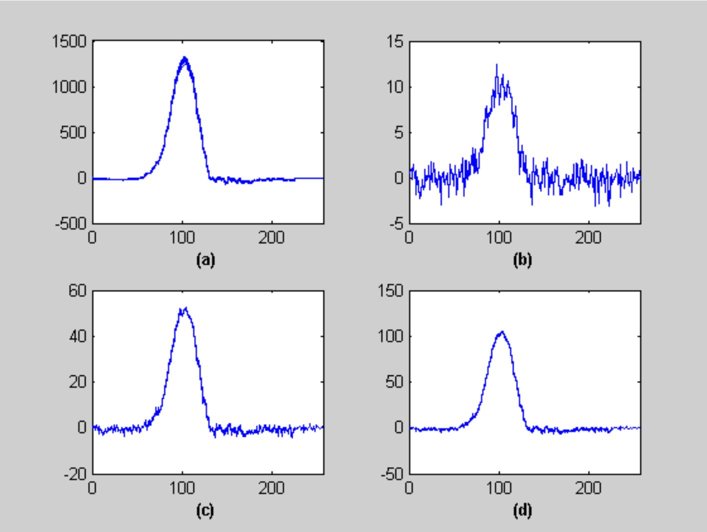


Figure 29: Tonnage Signals.

(a) Original (b) *SNR*=3 (c) *SNR*=15 (d) *SNR*=30

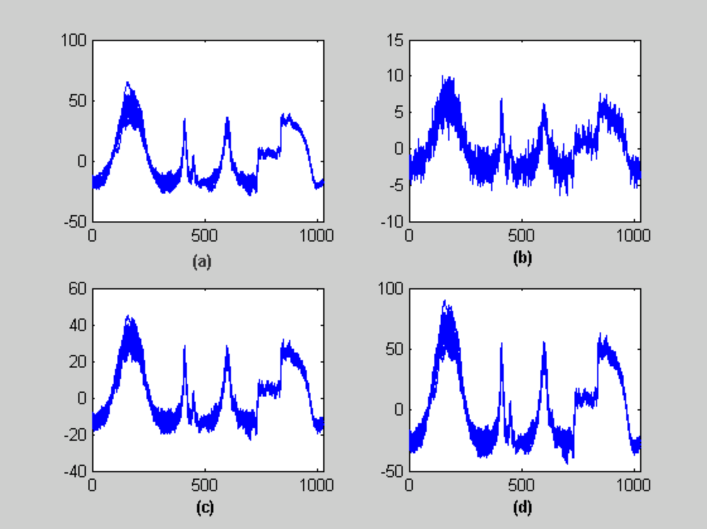


Figure 30: Mallat's Piecewise Signals

(a) Original (b) *SNR*=3 (c) *SNR*=15 (d) *SNR*=30

Table 28: *SNR* Results for Antenna Curves

Methods	*SNR*=3		*SNR*=15		*SNR*=30	
	RE	RR	RE	RR	RE	RR
VisuShrink-Union	0.0176	0.7109	0.0007	0.5391	0.0001	0.3750
VisuShrink-Intersection	0.0233	0.7500	0.0042	0.7344	0.0038	0.7344
VertiShrink	0.0043	0.4844	0.0002	0.3984	0.0001	0.3594
VET	0.0286	0.7969	0.0042	0.7734	0.0036	0.7734
MSVET	0.0052	0.7578	0.0008	0.7813	0.0006	0.8516

Table 29: *SNR* Results for Tonnage Signals

Methods	*SNR*=3		*SNR*=15		*SNR*=30	
	RE	RR	RE	RR	RE	RR
VisuShrink-Union	0.0850	0.8750	0.0039	0.8594	0.0009	0.8008
VisuShrink-Intersection	0.0850	0.8750	0.0041	0.8672	0.0013	0.8594
VertiShrink	0.0181	0.6367	0.0008	0.5898	0.0002	0.5469
VET	0.1019	0.9766	0.0065	0.9609	0.0031	0.9570
MSVET	0.0206	0.7461	0.0011	0.7461	0.0004	0.7461

Table 30: *SNR* Results for Mallat Piecewise Signals (*N*=1024)

Methods	*SNR*=3		*SNR*=15		*SNR*=30	
	RE	RR	RE	RR	RE	RR
VisuShrink-Union	0.0975	0.9600	0.0056	0.9170	0.0012	0.4648
VisuShrink-Intersection	0.1006	0.9619	0.0068	0.9580	0.0037	0.9580
VertiShrink	0.0185	0.6396	0.0022	0.6240	0.0015	0.5938
VET	0.1129	0.9688	0.0077	0.9600	0.0037	0.9590
MSVET	0.0240	0.7100	0.0031	0.7227	0.0022	0.7080

Table 31: *SNR* Results for Mallat Piecewise Signals (*N*=8192)

Methods	*SNR*=3		*SNR*=15		*SNR*=30	
	RE	RR	RE	RR	RE	RR
VisuShrink-Union	0.1027	0.9673	0.0063	0.9418	0.0019	0.6694
VisuShrink-Intersection	0.1253	0.9785	0.0170	0.9692	0.0113	0.9679
VertiShrink	0.0217	0.6638	0.0024	0.6492	0.0015	0.6051
VET	0.2139	0.9894	0.1123	0.9883	0.1047	0.9880
MSVET	0.1129	0.6969	0.0945	0.7255	0.0938	0.7059

VITA

Seung Hyun Baek was born in Suburi Ungcheoneup, Boryeong city, Republic of Korea in 1974. In 1993, Seung Hyun matriculated to Myongji University, Republic of Korea. After 2 years, he joined Korean Army as a field artillery soldier from 1995 to 1997. In 2000, he obtained his Bachelor's degree in Industrial Engineering. After completing his Bachelor's degree, he joined Georgia Institute of Technology to pursue a Master's degree in 2001 and completed his Master's degree in Industrial and Systems Engineering in 2002.

His research interests are Data Mining and Knowledge Discovery, Machine Learning, Statistical Data Modeling, Kernel-Based Methods, Model Selection, Data Preprocessing, Pattern Recognition, Spectrum Data Analysis (near infrared (NIR) spectroscopy, optical emission spectroscopy (OES), nuclear magnetic resonance (NMR) spectroscopy, fourier transform infrared (FTIR) spectroscopy), Functional Data Analysis (curve, surfaces, or anything else varying over a continuum), Quality and Reliability Engineering (quality and process improvement, control charts, reliability analysis of complex systems), Chemometrics, Bioinformatics, Spatial Data Mining, Transportation Modeling, Climate Modeling.

He published several papers in peer-reviewed journals such as Wood and Fiber Science, Wood Science and Technology, Journal of the Operational Research Society, IEEE Transactions on Systems, Man, and Cybernetics, Part C, Annals of Operations Research, and Expert Systems with Applications.