DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1136
February 21, 2007

Statistical Learning in Medical Data Analysis

Grace Wahba[1]

Prepared for the Handbook of Statistics: Epidemiology and Medical Statistics
C. R.Rao, J. Philip Miller and D. C Rao, Eds. Elsevier.

[1]Grace Wahba is the IJSchoenberg-Hilldale Professor of Statistics, Professor of Biostatistics and Medical Informatics, and Professor of Computer Sciences. Research supported in part by NIH Grant EY09946, NSF Grants DMS-0505636, DMS-0604572 and ONR Grant N0014-06-0095

Contents

Abstract

This article provides a tour of statistical learning regularization methods that have found application in a variety of medical data analysis problems. The uniting feature of these methods is that they involve an optimization problem which balances fidelity to the data with complexity of the model. The two settings for the optimization problems considered here are Reproducing Kernel Hilbert Spaces (a brief tutorial is included), and ℓ_1 penalties, which involve constraints on absolute values of model coefficients. The tour begins with thin plate splines, smoothing spline ANOVA models, multicategory penalized likelihood estimates and models for correlated Bernoulli data for regression, in these two settings. Leaving regression, the tour proceeds to the the Support Vector Machine, a modern and very popular tool for classification. Then classification based on dissimilarity information rather than direct attribute information is considered. All of the learning models discussed require dealing with the so-called bias-variance tradeoff, which means choosing the right balance between fidelity and complexity. Tuning methods for choosing the parameters governing this tradeoff are noted. The chapter ends with remarks relating Empirical Bayes and Gaussian Process Priors to the regularization methods

1 Introduction

In this article we will primarily describe regularization methods for statistical learning. In this class of methods a flexible, or nonparametric, statistical learning model is built as the solution to an optimization problem which typically has a term (or group of terms) that measure closeness of the model to the observations, balanced against another term or group of terms which penalize complexity of the model. This class of methods encompass the so called "kernel methods" in the machine learning literature which are associated with Support Vector Machines (SVMs) — SVMs are of primary importance for nonparametric classification and learning in biomedical data analysis. The classic penalized likelihood methods are also regularization/kernel methods methods, and between SVMs, penalized likelihood methods and other regularization methods, a substantial part of statistical learning methodology is covered.

The general learning problem may be described as follows: We are given a labeled (or partly labeled) training set: $\{y_i, x(i), i = 1, \ldots, n\}$ where $x(i)$ is an attribute vector of the ith subject and y_i is a response associated with it. We have $x \in \mathcal{X}$, $y \in \mathcal{Y}$, but we are deliberately not specifying the nature of either $\mathcal{X}$ or $\mathcal{Y}$ — they may be very simple or be highly complex sets. The statistical learning problem is to obtain a map $f(x) \to y$ for $x \in \mathcal{X}$, so that, given a new subject with attribute vector $x_* \in \mathcal{X}$, $f(x)$ generalizes well. That is, $f(x_*)$ predicts $\hat{y}_* \in \mathcal{Y}$, such that, if y_* associated with x_* were observable, then $\hat{y}_*$ would be a good estimate of it. More generally, one may want to estimate a conditional probability distribution for $y|x$. The use to which the model f is put may simply be to classify, but in many interesting examples, x is initially a large vector, and it is of scientific interest to know how f or some functionals of f depend on components or groups of components of x — the sensitivity, interaction, or variable selection problem. A typical problem in demographic medical studies goes as follows: Sets of $\{y_i, x(i)\}$ are collected in a defined population, where the attribute vectors are vectors of relevant medical variables such as age, gender, blood pressure, cholesterol, body mass index, smoking behavior, lifestyle factors, diet, and other variables of interest. A simple response might be whether or not the person exhibits a particular disease of interest ($y \in \{yes, no\}$). A major goal of evidence-based medicine is to be able to predict the likelihood of the disease for a new subject, based on its attribute vector. Frequently the nature of f (for example, which variables/patterns of variables most influence f) is to be used to understand disease processes and suggest directions for further

biological research.

The statistical learning problem may be discussed from different points of view, which we will call "hard" and "soft" (Wahba 2002). For hard classification, we would like to definitively assign an object with attribute x_* to one of two or more classes. For example given microarray data it is desired to classify leukemia patients into one of four possible classes (Brown, Grundy, Lin, Cristianini, Sugnet, Furey, Ares & Haussler 2000) (Lee, Lin & Wahba 2004). In the examples in (Lee et al. 2004) classification can be carried out nearly perfectly with a multicategory SVM (for other methods, see the references there). The difficulty comes about when the attribute vector is extremely large, the sample size is small, and the relationship between x and y is complex. The task is to mine the data for those important components or functionals of the entire attribute vector which can be used for the classification. Soft classification as used here is just a synonym for risk factor estimation where one desires to form an estimate of a probability measure on a set of outcomes — in typical demographic studies, if the outcome is to be a member of one of several classes, the classes are generally not separable by attribute vector, since two people with the same attribute vector may well have different responses. It's just that the probability distribution of the responses is sensitive to the attribute vector. The granddaddy of penalized likelihood estimation for this problem (O'Sullivan, Yandell & Raynor 1986) estimated the 19 year risk of a heart attack, given blood pressure and cholesterol at the start of the study. Classes of people who do and do not get heart attacks are generally far from separable on the basis of their risk factors - people with high blood pressure and high cholesterol can live a long life, but as a group their life expectancy is less than people without those risk factors. In both hard and soft classification, frequently one of the major issues is to understand which attributes are important, and how changes in them affect the risk. For example, the results can be used by doctors to decide when to persuade patients to lower their cholesterol, or for epidemiologists to estimate disease rates and design public health strategies in the general population. In other problems, particularly involving genetic data, it is of particular interest to determine which components of the genome may be associated with a particular response, or phenotype.

In Section 2 we review soft classification, where the emphasis is on obtaining a variety of flexible, nonparametric models for risk factor estimation. Vector-valued observations of various types are considered. A brief review of Reproducing Kernel Hilbert Spaces (RKHS) is included here. Section 3 describes recent developments in soft classification where individual

variable selection and variable pattern selection is important. Section 4 goes on to classification with SVMs, including multiple categories and variable selection. Section 5 discusses data that is given as dissimilarities between pairs of subjects or objects, and Section 6 closes this article with an overview of some of the tuning methods for the models discussed.

2 Risk factor estimation: penalized likelihood estimates

2.1 Thin plate splines

The Western Electric Health Study followed 1,665 men for 19 years and obtained data including men who were alive at the end of the followup period and those who had died from heart disease. Participants dying from other causes were excluded. Penalized likelihood estimation for members of the exponential family (McCullagh & Nelder 1989), which includes Bernoulli data (that is, zero-one, alive or dead, etc.) was first proposed in (O'Sullivan et al. 1986). The authors used a penalized likelihood estimate with a thin plate spline (tps) penalty to to get a flexible estimate of the 19 year risk of death by heart attack as a function of diastolic blood pressure and cholesterol. Figure 1 from (O'Sullivan et al. 1986) gives a parametric (linear) and nonparametric tps fit to the estimated log odds ratio after transformation back to probability.

It can be seen that the nonparametric fit has a plateau, which cannot be captured by the parametric fit. We now describe penalized likelihood estimation for Bernoulli data, and how the tps is used in the estimate in (O'Sullivan et al. 1986). Let x be a vector of attributes, and $y = 1$ if a subject with attribute x has the outcome of interest and 0 if they do not. Let the log odds ratio $f(x) = logp(x)/(1 - p(x))$ where $p(x)$ is the probability that $y = 1$ given x. Then $p(x) = e^{f(x)}/(1 + e^{f(x)})$. f is the so called canonical link for Bernoulli data (McCullagh & Nelder 1989). Given data $\{y_i, x(i), i = 1, \cdots, n\}$, the likelihood function is $\prod_{i=1}^{n} p(x(i))^{y_i}(1 - p(x(i))^{1-y_i}$, and the negative log likelihood can be expressed as a function of f:

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} -y_i f(x(i)) + log(1 + e^{f(x(i)}). \tag{1}$$

Linear (parametric) logistic regression would assume that $f(x) = \sum_\ell c_\ell B_\ell(x)$, where the B_ℓ are a small, fixed number of basis functions appropriate to the problem, generally linear or low degree polynomials in the components of x.

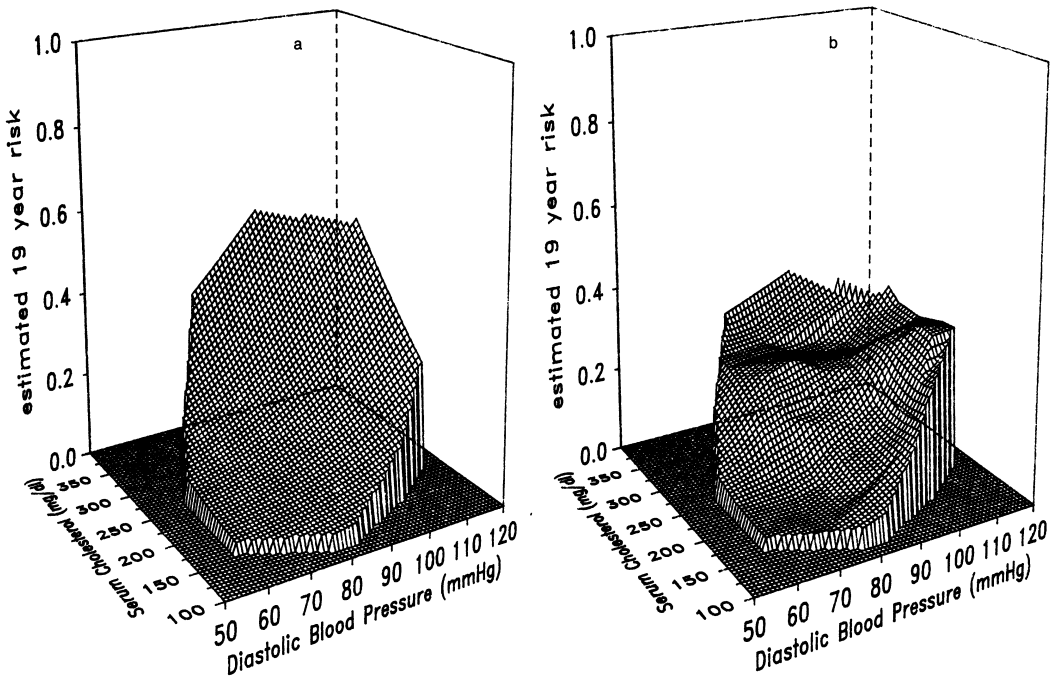

Figure 1: 19 year risk of a heart attack given serum cholesterol and diastolic blood pressure. Left: linear model in the log odds ratio. Right: tps estimate.©JASA

The penalized likelihood estimate of f is a solution to an optimization problem of the form: find f in $\mathcal{H}$ to minimize

$$\mathcal{L}(y, f) + \lambda J(f). \tag{2}$$

Here $\mathcal{H}$ is a special kind of RKHS (Gu & Wahba 1993a). For the Western Electric study, $J(f)$ was chosen so that f is a tps. See (Duchon 1977) (Meinguet 1979) (O'Sullivan et al. 1986) (Wahba 1990) (Wahba & Wendelberger 1980) for technical details concerning the tps. For the Western Electric study, the attribute vector $x = (x_1, x_2) =$ (cholesterol, diastolic blood pressure) was of dimension $d = 2$, and the two dimensional tps penalty functional of order 2 (involving second derivatives) is

$$J(f) = J_{2,2}(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2 dx_1 dx_2, \tag{3}$$

where the subscript $(2, 2)$ stands for (dimension, order). In this case f is known to have a representation

$$f(x) = d_0 + d_1x_1 + d_2x_2 + \sum_{i=1}^{n} c_i E(x, x(i)) \tag{4}$$

where

$$E(x, x(i)) = \|x - x(i)\|^2 log\|x - x(i)\|, \tag{5}$$

where $\|\cdot\|$ is the Euclidean norm. There is no penalty on linear functions of the components (x_1, x_2) of the attribute vector (the "null space" of $J_{2,2}$). It is known that the c_i for the solution satisfy $\sum_{i=1}^n c_i = 0$, $\sum_{i=1}^n c_i x_1(i) = 0$ and $\sum_{i=1}^n c_i x_2(i) = 0$, and furthermore,

$$J(f) = \sum_{i,j=1,\cdots,n} c_i c_j E(x(i), x(j)). \tag{6}$$

Numerically, the problem is to minimize (2) under the stated conditions and using (6) to obtain $d_0, d_1, d_2, c = (c_1, \cdots, c_n)$.

We have described the penalty functional for the thin plate spline and something about what it looks like for the $d = 2, m = 2$ case in (3). However, the tps is available for general d and for any m with $2m - d > 0$. The general tps penalty functional in d dimensions and m derivatives is

$$J_{d,m} = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right)^2 \prod_j dx_j. \tag{7}$$

See (Wahba 1990). Note that there is no penalty on polynomials of degree less than m, so that the tps with d greater than 3 or 4 is rarely attempted because of the very high dimensional null space of $J_{d,m}$.

The choice of the tuning parameter λ here governs the tradeoff between the goodness of fit to the data, as measured by the likelihood, and the complexity, or wiggliness of the fit. Note that second derivative penalty functions limit curvature, and tend to agree with human perceptions of smoothness, or lack of wiggliness. When the data are Gaussian (as in (Wahba & Wendelberger 1980)) rather than Bernoulli, the tuning (smoothing) parameter λ can be chosen by the GCV (Generalized Cross Validation) method; a related alternative method is GML (Generalized Maximum Likelihood), also known as REML (Restricted Maximum Likelihood). The order m of the tps may be chosen by minimizing with respect to λ for each value of m and then choosing m with the smallest minimum. See (Craven & Wahba 1979) (Golub, Heath & Wahba 1979) (Gu & Wahba 1991). As λ tends to infinity, the solution tends to its best fit in the unpenalized space, and as λ tends to 0, the solution attempts to interpolate the data. In the case of biomedical data it is sometimes the case that a simple parametric model (low degree polynomials, for example) is adequate to describe the data. The experimenter can design such a model to be in the null space of the penalty functional, and then a sufficiently large λ will produce the parametric model. Detailed discussion of tuning parameters for Bernoulli data is in Section 6.

A number of commercial as well as public codes exist for computing the thin plate spline, with the GCV or GML method of choosing the tuning parameters. Public codes in R (`http://cran.r-project.org/` include `assist, fields, gss, mgcv`. The original fortran tps code is found in netlib (`www.netlib.org/gcv`). Further details on the thin plate spline can be found in the historical papers (Duchon 1977) (Meinguet 1979) (Wahba 1990) (Wahba & Wendelberger 1980), in the documentation for the `fields` code in R, and elsewhere. Thin plate splines are used in the "morphing" of medical images (Bookstein 1997), and have been used to fit smooth surfaces to data that has been aggregated over irregular geometrical shapes such as counties (Wahba 1981).

2.2 Positive definite functions and Reproducing Kernel Hilbert Spaces

We will give a brief introduction to positive definite functions and RKHSs here, because all of the so-called "kernel methods" which we will be discussing have their foundation as optimization problems in these spaces. The reader who wishes to avoid this technicality may skip this subsection. Let $\mathcal{T}$ be *some* domain, emphasizing the generality of the domain. For

concreteness you may think of $\mathcal{T}$ as Euclidean d-space. $K(\cdot,\cdot)$ is said to be positive definite if, for every n and any $t(1),\cdots,t(n)\in\mathcal{T}$ and $c_1,\cdots,c_n$

$$\sum_{i,j=1}^{n} c_i c_j K(t(i),t(j)) \geq 0. \tag{8}$$

In this article we denote the inner product in an RKHS by $<\cdot,\cdot>$. To every positive definite function $K(\cdot,\cdot)$ there is associated an RKHS $\mathcal{H}_K$ (Aronszajn 1950) (Wahba 1990) which can be constructed as a collection of all functions of the form

$$f_L^a(t) = \sum_{\ell=1}^{L} a_\ell K(t,t(\ell)) \tag{9}$$

with the inner product

$$< f_L^a, f_M^b >= \sum_{\ell,m} a_\ell b_m K(t(\ell),t(m)) \tag{10}$$

and all functions that can be constructed as the limits of all Cauchy sequences in the norm induced by this inner product; these sequences can be shown to converge pointwise. What makes these spaces so useful is that in an RKHS $\mathcal{H}_K$ we can always write for any $f\in\mathcal{H}_K$

$$f(t_*) =< K_{t_*}, f > \tag{11}$$

where $K_{t_*}(\cdot)$ is the function of t given by $K(t_*,t)$ with t_* considered fixed. A trivial example is $\mathcal{T}$ is the integers $1,\cdots,n$. There K is an $n\times n$ matrix, the elements of $\mathcal{H}_K$ are n-vectors, and the inner product is $<f,g>=f'K^{-1}g$. Kernels with penalty functionals that involve derivatives are popular in applications. A simple example of a kernel whose square norm involves derivatives is the kernel K associated with the space of periodic functions on $[0,1]$ which integrate to 0 and which have square integrable second derivative. It is $K(s,t)=B_2(s)B_2(t)/(2!)^2 - B_4(|s-t|)/4!$, where $s,t\in[0,1]$, and B_m is the mth Bernoulli polynomial, see (Wahba 1990). The square norm is known to be $\int_0^1 (f''(s))^2 ds$. The periodic and integration constraints are removed by adding linear functions to the space and the fitted functions can be shown to be cubic polynomial splines. For more on polynomial splines see (Craven & Wahba 1979) (deBoor 1978) (Wahba 1990). Another popular kernel is the Gaussian kernel, $K(s,t)=\exp\left(-\frac{1}{\sigma^2}\|s-t\|^2\right)$ defined for s,t in Euclidean d space, E^d, where the norm in the exponent is the Euclidean norm. Elements of this space are generated from functions of $s\in E^d$ of the form $K_{t_*}(s)=\exp\left(-\frac{1}{\sigma^2}\|s-t_*\|^2\right)$, for $t_*\in E^d$. Kernels on E^d that depend only on the Euclidean distance between their two arguments are

known as radial basis functions (rbf's). Another popular class of rbf's is the Matern class, see (Stein 1999). Matern kernels have been used to model arterial blood velocity in (Carew, Dalal, Wahba & Fain 2004), after fitting the velocity measurements, estimates of the wall shear stress are obtained by differentiating the fitted velocity model.

We are now ready to write a (special case of) a general theorem about optimization problems in RKHS.

The Representer Theorem (special case)(Kimeldorf & Wahba 1971): Given observations $\{y_i, t(i), i = 1, 2, \cdots, n\}$, where y_i is a real number and $t(i) \in \mathcal{T}$, and given K and (possibly) some particular functions $\{\phi_1, \cdots \phi_M\}$ on $\mathcal{T}$, find f of the form $f(s) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(s) + h(s)$ where $h \in \mathcal{H}_K$ to minimize

$$I_\lambda\{y, f\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t(i))) + \lambda \|h\|^2_{\mathcal{H}_K} \tag{12}$$

where $\mathcal{C}$ is a convex function of f. It is assumed that the minimizer of $\sum_{i=1}^{n} \mathcal{C}(y_i, f(t(i)))$ in the span of the ϕ_ν is unique. Then the minimizer of $I_\lambda\{y, f\}$ has a representation of the form:

$$f(s) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(s) + \sum_{i=1}^{n} c_i K(t(i), s). \tag{13}$$

The coefficient vectors $d = (d_1, \cdots d_M)'$ and $c = (c_1, \cdots, c_n)'$ are found by substituting (13) into the first term in (12), and using the fact that $\| \sum_{i=1}^{n} c_i K_{t(i)}(\cdot)\|^2_{\mathcal{H}_K} = c' K_n c$ where K_n is the $n \times n$ matrix with i, jth entry $K(t(i), t(j))$. The name "reproducing kernel" comes from the fact that $< K_{t_*}, K_{s_*} >= K(t_*, s_*)$.

The minimization of (12) generally has to be done numerically by an iterative descent method, except in the case that $\mathcal{C}$ is quadratic in f, in which case a linear system has to be solved. When $K(\cdot, \cdot)$ is a smooth function of its arguments and n is large, it has been found that excellent approximations to the minimizer of (12) for various $\mathcal{C}$ can be found with functions of the form:

$$f(s) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(s) + \sum_{j=1}^{L} c_{i_j} K(t((i_j), s), \tag{14}$$

where the $t(i_1), \cdots t(i_L)$ are a relatively small subset of $t(1), \cdots, t(n)$, thus reducing the computational load. The $t(i_1), \cdots, t(i_L)$ may be chosen in various ways, as a random subset, by clustering the $\{t(i)\}$ and selecting from each cluster (Xiang & Wahba 1997), or by a greedy algorithm, as for example in (Luo & Wahba 1997), depending on the problem.

2.3 Smoothing spline ANOVA models

Thin plate spline estimates and fits based on the Gaussian kernel and other radial basis functions are (in their standard form) rotation invariant in the sense that rotating the coordinate system, fitting the model, and rotating back do not change anything. Thus they are not appropriate for additive models or for modeling interactions of different orders.

Smoothing spline ANOVA (SS-ANOVA) models provide fits to data of the form $f(t) = C + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$. Here f_α is in some RKHS $\mathcal{H}^\alpha$, $f_{\alpha\beta} \in \mathcal{H}^\alpha \otimes \mathcal{H}^\beta$ and so forth. The components of the decomposition satisfy side conditions which generalize the usual side conditions for parametric ANOVA which make the solutions unique. The f_α integrate to zero, the $f_{\alpha\beta}$ integrate to zero over both arguments, and so forth. f is obtained as the minimizer, in an appropriate function space, of

$$I_\lambda\{y, f\} = \mathcal{L}(y, f) + \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots, \tag{15}$$

where $\mathcal{L}(y, f)$ is the negative log likelihood of $y = (y_1, \cdots, y_n)$ given f, the $J_\alpha, J_{\alpha\beta}, \cdots$ are quadratic penalty functionals in RKHS, the ANOVA decomposition is terminated in some manner, and the λ's are to be chosen. The "spline" in SS-ANOVA models is somewhat of a misnomer, since SS-ANOVA models do not have to consist of splines. The attribute vector $t = (t_1, \cdots, t_d)$, where $t_\alpha \in \mathcal{T}^{(\alpha)}$, is in $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)} \otimes \cdots \otimes \mathcal{T}^{(d)}$ where the $\mathcal{T}^{(\alpha)}$ may be quite general. The ingredients of the model are: For each α, there exist a probability measure $\mu_{(\alpha)}$ on $\mathcal{T}^{(\alpha)}$, and an RKHS of functions $\mathcal{H}^\alpha$ defined on $\mathcal{T}^{(\alpha)}$ such that the constant function is in $\mathcal{H}^\alpha$ and the averaging operator $\mathcal{E}_\alpha f = \int f_\alpha(t_\alpha) d\mu_\alpha$ is well defined for any $f_\alpha \in \mathcal{H}^\alpha$. Then f is in (a subspace of) $\mathcal{H} = \mathcal{H}^1 \otimes \mathcal{H}^2 \cdots \mathcal{H}^d$. The ANOVA decomposition generalizes the usual ANOVA taught in elementary statistics courses via the expansion

$$I = \prod_\alpha (\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) = \prod_\alpha \mathcal{E}_\alpha + \sum_\alpha (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta + \sum_{\alpha<\beta} (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma + \cdots + \prod_\alpha (I - \mathcal{E}_\alpha). \tag{16}$$

The components of this decomposition generate the ANOVA decomposition of f by

$$C = (\prod_\alpha \mathcal{E}_\alpha) f, \quad f_\alpha = ((I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta) f, \quad f_{\alpha\beta} = ((I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma) f, \cdots \tag{17}$$

and so forth. The spaces $\mathcal{H}^\alpha$ are decomposed into the one dimensional spaces of constant functions, and $\mathcal{H}^{(\alpha)}$, whose elements satisfy $\mathcal{E}_\alpha f = 0$. The $\mathcal{H}^{(\alpha)}$ may be further decomposed into low dimensional unpenalized subspaces plus smooth subspaces that will be penalized.

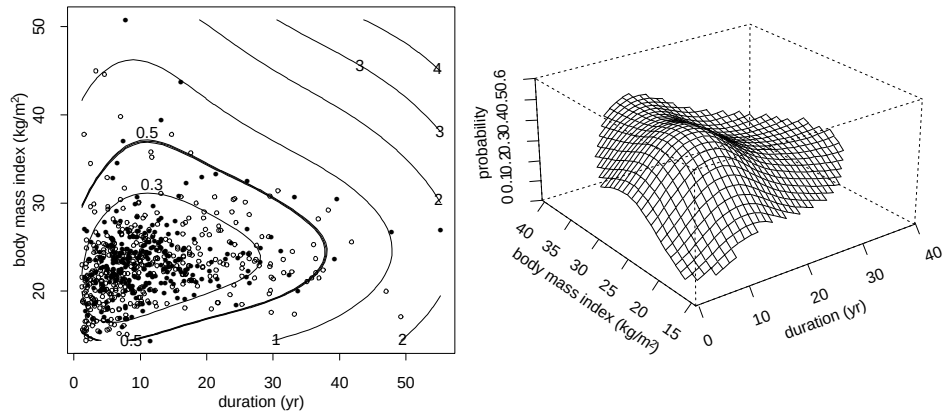

Figure 2: Four year probability of progression of diabetic retinopathy as a function of duration of diabetes at baseline and body mass index, with glycosylated hemoglobin set at its median. ©Ann. Statist.

All this allows the exploitation of the geometry of RKHS to obtain the minimizer of $I_\lambda\{y, f\}$ of (15) in a convenient manner. Reproducing kernels (RKs) for the various subspaces are constructed from Kronecker products of the RKs for functions of one variable. SS-ANOVA models are studied in detain in (Gu 2002). Other references include (Davidson 2006) (Gao, Wahba, Klein & Klein 2001) (Gu & Wahba 1993*b*) (Lin 2000) (Wahba 1990) (Wahba, Wang, Gu, Klein & Klein 1995) (Wang 1998) (Wang, Ke & Brown 2003).

Figure 2 from (Wahba et al. 1995) plots the four year probability of progression of diabetic retinopathy based on three predictor variables, `dur` = duration of diabetes, `gly` = glycosylated hemoglobin, and `bmi` = body mass index. An SS-ANOVA model based on cubic splines was fitted with the result

$$f(t) = C + f_1(\texttt{dur}) + a\texttt{gly} + f_3(\texttt{bmi}) + f_{13}(\texttt{dur}, \texttt{bmi}). \tag{18}$$

In the cubic spline fit, there is no penalty on linear functions. For the `gly` term, the estimated smoothing parameter was sufficiently large so that the fit in `gly` was indistinguishable from linear so that $f_2(\texttt{gly})$ became $a\texttt{gly}$. For the plot, `gly` has been set equal to its median. Software for SS-ANOVA models can be found in the R codes `gss`, which is keyed to (Gu 2002), and `assist`. Software for main effects models is found in the R code `gam`, based on (Hastie & Tibshirani 1986).

2.4 Multicategory penalized likelihood estimates

Multicategory penalized likelihood methods were first proposed in (Lin 1998), see also (Wahba 2002). In this setup, the endpoint is one of several categories; in the works cited, the categories were "alive" or "deceased" by cause of death. Considering $K+1$ possible outcomes, with $K > 1$, let $p_j(x), j = 0, 1, \cdots, K$ be the probability that a subject with attribute vector x is in category j, $\sum_{j=0}^{K} p_j(x) = 1$. The following approach was proposed in (Lin 1998): Let $f_j(x) = \log[p_j(x)/p_0(x)], j = 1, \cdots, K$, where p_0 is assigned to a base class. Then

$$\begin{aligned} p_j(x) &= \frac{e^{f_j(x)}}{1 + \sum_{j=1}^{K} e^{f_j(x)}}, \quad j = 1, \cdots, K \\ p_0(x) &= \frac{1}{1 + \sum_{j=1}^{K} e^{f_j(x)}}. \end{aligned} \tag{19}$$

The class label for the ith subject is coded as $y_i = (y_{i1}, \cdots, y_{iK})$ where $y_{ij} = 1$ if the ith subject is in class j and 0 otherwise. Letting $f = (f_1, \cdots, f_K)$, the negative log likelihood can be written as

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} \{ \sum_{j=1}^{K} -y_{ij} f_j(x(i)) + log(1 + \sum_{j=1}^{K} e^{f_j(x(i))}) \} \tag{20}$$

and an SS-ANOVA model was fitted as a special (main effects) case of (15) with cubic spline kernels.

Figure 3 from (Lin 1998) gives ten year risk of mortality by cause as a function of age. The model included two other risk factors, glycosylated hemoglobin and systolic blood pressure at baseline, and they have been set equal at their medians for the plot. The differences between adjacent curves (from bottom to top) are probabilities for alive, diabetes, heart attack, and other causes. The data are plotted as triangles (alive, on the bottom), crosses (diabetes) diamonds (heart attack) and circles (other).

See also (Zhu & Hastie 2003), who proposed a version of the multicategory penalized likelihood estimate for Bernoulli data that did not have a special base class. The model is

$$p_j(x) = \frac{e^{f_j(x)}}{\sum_{j=1}^{K} e^{f_j(x)}}, \quad j = 1, \cdots, K. \tag{21}$$

This model is overparametrized, but that can be handled by adding a sum-to-zero constraint $\sum_{j=1}^{K} f_j(x) = 0$, as was done in multicategory support vector machine (Lee et al. 2004) discussed later. The authors show that this constraint is automatically satisfied in the optimization problem they propose.

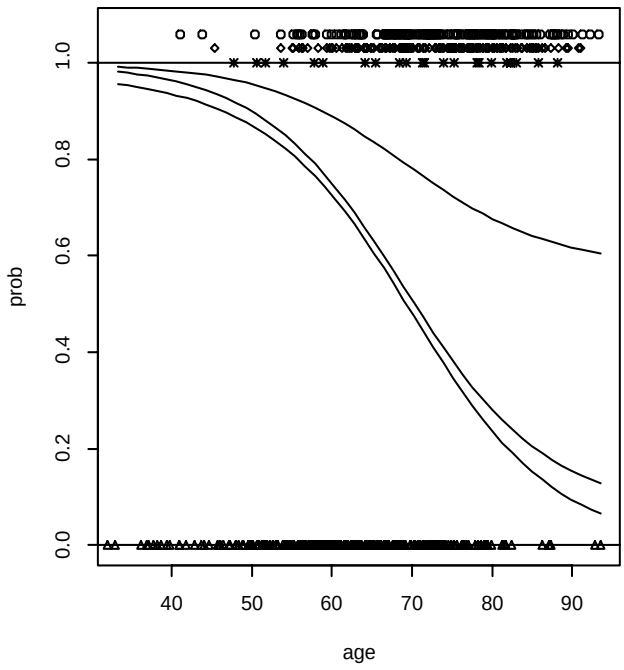

Figure 3: Ten year risk of mortality, by cause. See text for explanation.

2.5 Correlated Bernoulli data: The two eye problem, the multiple sibs problem

In (Gao et al. 2001), a general model including the following is considered: There are n units, each unit has K members, and there is a Bernoulli outcome that is 0 or 1, for each member. There may be member-specific risk factors and unit-specific risk factors. Thus, the responses are vectors $y_i = (y_{i1}, \cdots, y_{iK})$ where $y_{ij} \in \{0, 1\}$ is the response of the jth member of the ith unit. Allowing only first order correlations, a general form of the negative log likelihood is

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} \{ \sum_{j=1}^{K} -y_{ij} f_j(x(i)) - \sum_{j \neq k} \alpha_{jk} y_{ij} y_{ik} + b(f, \alpha) \} \tag{22}$$

where (suppressing the dependence of the f_j on $x(i)$), we have

$$b(f, \alpha) = log \left(1 + \sum_{j=1}^{K} e^{f_j} + \sum_{j \neq k} e^{f_j + f_k + \alpha_{jk}} + \sum_{j \neq k \neq l} e^{f_j + f_k + f_l + \alpha_{jk} + \alpha_{jl} + \alpha_{kl}} + \cdots + e^{(\sum_{j=1}^{K} f_j + \sum_{j \neq k} \alpha_{jk})} \right) \tag{23}$$

The α_{jk} are the log odds ratios ($logOR$) and are a measure of the correlation of the jth and kth outcome when the other outcomes are 0:

$$\alpha_{jk} = log\mathrm{OR}(j, k) = \frac{\Pr(y_j = 1, y_k = 1) \Pr(y_j = 0, y_k = 0)}{\Pr(y_j = 1, y_k = 0) \Pr(y_j = 0, y_k = 1)} | y_r = 0, r \neq j, k. \tag{24}$$

The two eye problem was considered in detail in (Gao et al. 2001) where the unit is a person and the members are the right eye and the left eye. The outcomes are pigmentary abnormality in each eye. There only person-specific predictor variables were considered, so that K is 2, $f_1(x) = f_2(x) = f(x)$ where $x(i)$ is the ith vector of person-specific risk factors, and there is a single $\alpha_{12} = \alpha$. In that work α was assumed to be a constant, f is an SS-ANOVA model, and $I_\lambda(y, f)$ of the form (15) is minimized with $\mathcal{L}(y, f)$ of the form (22). The cross product ratio $a_{12} = logOR(1, 2)$ is a measure of the correlation between the two eyes, taking into account the person-specific risk factors. It may be used to estimate whether, e. g. the second eye is likely to have a bad outcome, given that the first eye already has. The case where the unit is a family and the members are a sibling pair within the family with person-specific attributes, is considered in (Chun 2006), where the dependence on person-specific attributes has the same functional form for each sib. Then $K = 2$ and $f_j(x(i))$ becomes $f(x_j(i))$, where $x_j((i)$ is the attribute vector of the jth sibling, $j = 1, 2$ in the ith family. Again, an optimization problem of the form (15) is solved. If α is large, this indicates correlation within the family, taking account of person-specific risk factors, and may suggest looking for genetic components.

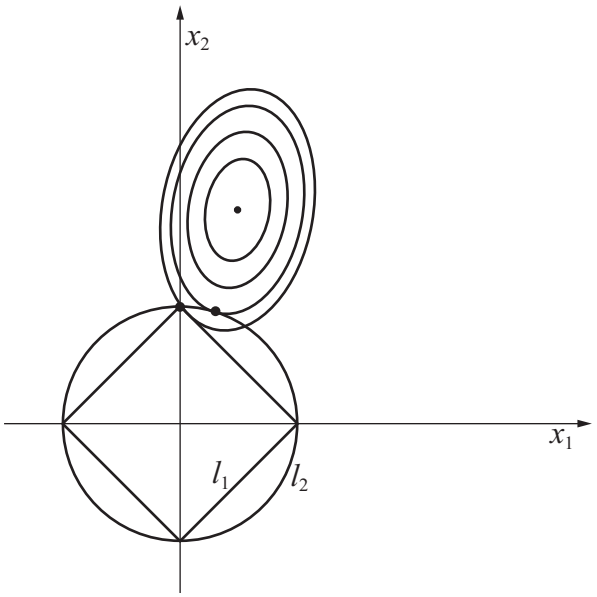

Figure 4: Absolute value penalties lead to solutions at the extreme points of the diamond, which means sparsity in the solution vector.

3 Risk factor estimation: likelihood basis pursuit and the LASSO

3.1 The l_1 penalty

In Section 2 the penalty functionals were all quadratic, being square norms or seminorms[1] in an RKHS. Generally if there are n observations there will be n representers in the solution, for very large n this is not desirable. This may be mitigated as in (14), but it is well known that imposing an absolute value penalty (l_1 penalty) on coefficients of the form $\sum_{i=1}^{n} |c_i|$ (as opposed to a quadratic form in the c's) will tend to provide a sparse solution, that is, many of the c_i will be zero. Figure 4 suggests why. The concentric ellipses are meant to represent the level curves of a quadratic function $Q(x_1, x_2)$ in x_1 and x_2 (with the minimum in the middle) and the circle and inscribed diamond are level curves of $|x|_{l_2} = x_1^2 + x_2^2$ and $|x|_{l_1} = |x_1| + |x_2|$ respectively. If the problem is to minimize $Q(x) + |x|_{l_p}$ for $p = 1$ or 2, it can be seen that with the l_1 norm, the minimum is more likely to be at one of the corners of the diamond. The desirability of sparsity comes up in different contexts; to select a sparser number of basis functions given an overcomplete set of basis functions, or to

[1] A seminorm here is the norm of the projection of f onto a subspace with orthocomplement of low dimension. The orthocomplement is the null space of J. The thin plate penalty functionals are seminorms.

select a smaller number of variables or clusters of variables out of a much larger set to be used for regression or classification. Likelihood basis pursuit (Chen, Donoho & Saunders 1998) and the LASSO (Tibshirani 1996) are two basic papers in the basis function context and variable selection context respectively. There is a large literature in the context of variable selection in linear models, based on the LASSO, which in its simplest form imposes an l_1 penalty on the coefficients in a linear model, see (Efron, Hastie, Johnstone & Tibshirani 2004) (Fan & Li 2001) (Knight & Fu 2000) and others. An overcomplete set of basis functions in a wavelet context was generated in (Chen et al. 1998), who then reduced the number of basis functions in their model via an l_1 penalty on the coefficients. In the spirit of (Chen et al. 1998), (Zhang, Wahba, Lin, Voelker, Ferris, Klein & Klein 2004) generated an overcomplete set of basis functions by the use of of representers in an SS-ANOVA model to do model fitting and variable selection in a flexible way, similarly reducing the number of main effects or interactions by an l_1 penalty on basis function coefficients, The method was used to obtain flexible main effects models for risk factors for eye diseases based on data collected in the Beaver Dam Eye Study (Klein, Klein, Linton & DeMets 1991). Beginning with (Gunn & Kandola 2002) various authors have simultaneously imposed l_1 and quadratic penalties in the context of flexible nonparametric regression/kernel methods, see (Zhang & Lin 2006*b*) (Zhang 2006) and (Zhang & Lin 2006*a*) (who called it "COSSO"). Software for the COSSO may be found at `http://www4.stat.ncsu.edu/~hzhang/software.html`. Later (Zou & Hastie 2005) (calling it "Elastic Net") in the context of (linear) parametric regression, used the same idea of a two term penalty functional, one quadratic the other l_1.

3.2 LASSO-Patternsearch

The LASSO-Patternsearch method of (Shi, Wahba, Lee, Klein & Klein 2006) was designed with specially selected basis functions and tuning procedures to take advantage of the sparsity inducing properties of l_1 penalties to enable the detection of potentially important higher order variable interactions. Large and possibly very large attribute vectors $x = (x_1, \cdots, x_p)$ with entries 0 or 1 are considered, with Bernoulli outcomes. The log odds ratio $f(x) = log[p(x)/(1-p(x))]$ is modeled there as

$$f(x) = \mu + \sum_{\alpha=1}^{p} c_\alpha B_\alpha(x) + \sum_{\alpha<\beta} c_{\alpha\beta} B_{\alpha\beta}(x) + \sum_{\alpha<\beta<\gamma} c_{\alpha\beta\gamma} B_{\alpha\beta\gamma}(x) + \cdots + c_{123\cdots p} B_{123\cdots p}(x) \qquad (25)$$

where $B_\alpha(x) = x_\alpha, B_{\alpha\beta}(x) = x_\alpha x_\beta$ and so forth, and the optimization problem to be solved is: Find f of the form (25) to minimize

$$I_\lambda\{y, f\} = \sum_{i=1}^{n} -y_i f(x(i)) + log(1 + e^{f(x(i))}) + \lambda \sum_{all\ c} |c|, \qquad (26)$$

where the sum taken over all c means the sum of the absolute values of the coefficients (the l_1 penalty). For p small (say, $p = 8$), the series in (25) may be continued to the end, but for large p the series will be truncated. A special purpose numerical numerical algorithm was proposed that can handle a very large number (at least 4000) of unknown coefficients, many of which will turn out to be 0. The "patterns", or basis functions in (25) follow naturally from the log linear representation of the multivariate Bernoulli distribution, see (Shi et al. 2006) (Whittaker 1990). This approach is designed for the case when the direction of all or almost all of the "risky" variables are known and are coded as 1, since then the representation of (25) is most compact then, although this is by no means necessary.. When this and similar problems are tuned for predictive loss, there is a bias towards overestimating the number of basis functions and including some noise patterns. However, at the same time it insures a high probability of including including all the important basis functions, see (Leng, Lin & Wahba 2006) (Zou 2006). The LASSO-Patternsearch is a two-step approach, with the first step global, as opposed to a greedy approach. In the first first step the model is fitted globally and tuned by a predictive loss criteria. Then a second step takes those patterns surviving the first step and enters them a parametric generalized linear model. Finally, all basis functions whose coefficients fail a significance test in this model at level q are deleted, where the value of q is treated as another tuning parameter. This method uncovered an interesting relation between smoking vitamins and cataracts as risk factors in myopia data collected as part of the Beaver Dam Eye study (Klein et al. 1991). The method has also been successfully used to select patterns of SNP's (single nucleotide polymorphisms in DNA data) that can separate cases from controls with a high degree of accuracy. Pseudocode is found in (Shi et al. 2006). Other approaches for finding clusters of important variables include (Breiman 2001) (Ruczinski, Kooperberg & LeBlanc 2002) (Yuan & Lin 2006) (Park & Hastie 2007). These methods rely on sequential, stepwise or greedy algorithms, and tend to work well in a broad range of scenarios, although stepwise algorithms are not guaranteed to always find the best subset. Some preliminary results suggest that under certain kinds of correlated scenarios the global aspects of the LASSO Patternsearch may prove advantageous over stepwise approaches.

4 Classification: Support Vector Machines and related estimates

Support Vector Machines were proposed by Vapnik and colleagues as a nonparametric classification method in the early 90's, see (Vapnik 1995) and references cited there, where it was obtained in an argument quite different than the description we give here. However in the late 90's (Evgeniou,

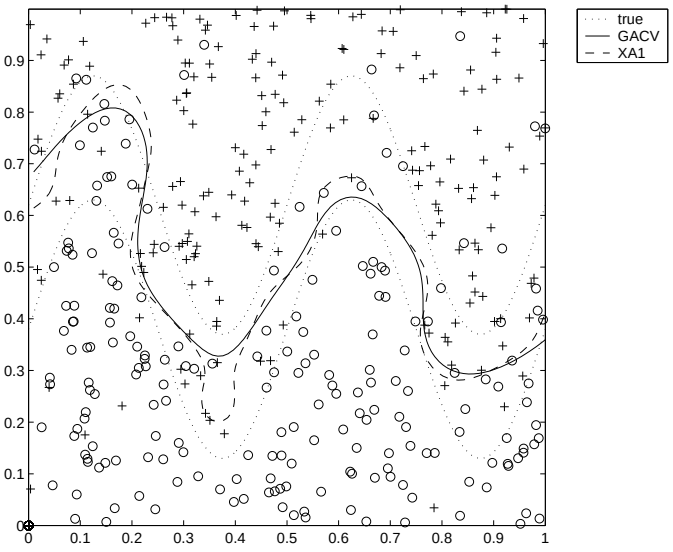

Figure 5: SVM: Toy problem, tuned by GACV and the XiAlpha method.

Pontil & Poggio 2000) (Wahba 1999) it was observed that SVMs could be obtained as the solution to an optimization problem in an RKHS. This made it easy to compare and contrast SVMs with other nonparametric methods involving optimization problems in an RKHS, to develop generalizations, and to examine its theoretical properties. In any case the efficiency of the SVM was quickly recognized in practice, and theory soon followed to explain just why SVMs worked so well. Before giving details, we note the following books: (Cristianini & Shawe-Taylor 2000) (Scholkopf, Burges & Smola 1999) (Scholkopf & Smola 2002) (Scholkopf, Tsuda & J-P.Vert 2004) (Shawe-Taylor & Cristianini 2004) (Smola, Bartlett, Scholkopf & Schuurmans 2000).

4.1 Two category Support Vector Machines

Figure 5 illustrates the flexibility of a (two category) SVM. The locations of the + and o "attribute vectors" were chosen according to a uniform distribution on the unit rectangle. Attribute vectors falling between the two dotted lines were assigned to be + or o with equal probability of .5. Points above the upper dotted (true) line were assigned + with .95 and o with probability .05, and below the lower dotted line the reverse: o with probability .95 and + with probability .05. Thus, any classifier whose boundary lies within the two dotted lines is satisfying the Bayes rule - that is, it will minimize the expected classification error from new observations drawn from the same distribution. In the two category SVM the training data is coded $y_i \pm 1$ according as the ith object is in the + class or the o class. The classifier f is assumed to be of the form $f(s) = d + h(s)$ where the

constant d and $h \in \mathcal{H}_K$ are chosen to minimize

$$I_\lambda\{y, f\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t(i))) + \lambda \|h\|^2_{\mathcal{H}_K} \tag{27}$$

where $\mathcal{C}$ is the so-called hinge function: $\mathcal{C}(y, f) = (1 - yf)_+$ where $(\tau)_+ = 1$ if $\tau > 0$ and 0 otherwise. A new object with $f(x) > 0$ will be classified as in the + class and $f(x) < 0$ in the o class. From the representer theorem, the minimizer of $I_\lambda\{y, f\}$ again has a representation of the form:

$$f(s) = d + \sum_{i=1}^{n} c_i K(t(i), s). \tag{28}$$

$\| \sum_{i=1}^{n} c_i K_{t(i)}(\cdot)\|^2_{\mathcal{H}_K} = c' K_n c$ where K_n is the $n \times n$ matrix with i, jth entry $K(t(i), t(j))$ is substituted into (27). The problem of finding d and $c_1, \cdots, c_n$ is solved numerically by transforming the problem to its dual problem, which results in the problem of minimizing a convex functional subject to a family of linear inequality constraints. Details of this transformation may be found in any of the books cited, in (Evgeniou et al. 2000) (Wahba, Lin & Zhang 2000) and elsewhere.

For the toy problem in Figure 5, the reproducing kernel $K(s, t)$ was taken as the Gaussian kernel $K(s, t) = e^{-\frac{1}{2\sigma^2}\|s-t\|^2}$, so that the two tuning parameters λ and σ^2 have to be chosen. The solid line in Figure 5 is the 0 level curve of f obtained by choosing λ and σ^2 by the GACV method, and the dashed line by choosing λ and σ^2 by Joachim's XiAlpha method, see Section 6.4. The SVM^{light} software is popular code for computing the two class SVM, and the XiAlpha method is implemented in it. See (Joachims 1999), `http://svmlight.joachims.org`. Other codes and references can be found at `http://www.kernel-machines.org`.

Figure 6 is a toy example which demonstrates the difference between SVM and penalized likelihood estimates. The penalized likelihood method provides an estimate of the probability p that an object is in the "1" class. p is above or below .5 according as f is positive or negative. Therefore a classification problem with a representative number of objects in each class in the training set, and equal costs of misclassification can be solved by implementing the Bayes rule, which is equivalent to determining whether the log odds ratio f is positive or negative. The fundamental reason why the SVM works so well is that *it is estimating the sign of the log odds ratio.* See (Lin 2001)(Lin 2002)(Lin, Wahba, Zhang & Lee 2002) for proofs. This is demonstrated in Figure 6. The vertical scale in Figure 6 is $2p - 1$. 300 equally spaced samples in x were selected and assigned the + class with probability p, given by the solid ("truth") line in Figure 5 . The dotted line (labeled "logistic regression") is the penalized likelihood estimate of $2p - 1$ and is very close to the true $2p - 1$. The dashed line is the SVM. The SVM is very close to -1 if $2p - 1 < 0$, and close to $+1$ for $2p - 1 > 0$. Note however that they result in almost exactly the same classifier. The SVM is just

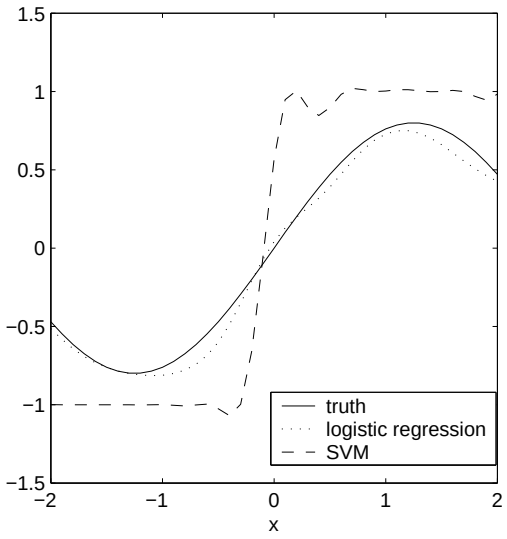

Figure 6: Penalized likelihood and the support vector machine compared.

one member of the class of large margin classifiers. A large margin classifier is one where $\mathcal{C}(y, f)$ depends only on the product yf. When the data are coded as ± 1, then the negative log likelihood becomes $log(1 + e^{yf})$ and so it is also a large margin classifier. From (Lin 2001) (Lin 2002) it can be seen that under very weak conditions on $C(y, f) = C(yf)$, large margin classifiers implement the Bayes rule, that is, the sign of the estimate of f is an estimate of the sign of the log odds ratio. Among the special properties of the SVM, however, is that the hinge function is, in some sense, the closest convex upper bound to the misclassification counter $[-yf]*$, where $[\tau]* = 1$ if $\tau > 0$ and 0 otherwise. Furthermore, due to the nature of the dual optimization problem, the SVM estimate of f tends to have a sparse representation, that is, many of the coefficients c_i are 0, a property not shared by the penalized likelihood estimate.

Regarding the form $f(s) = d + h(s)$ with $h \in \mathcal{H}_K$ of (27), frequently the kernel K is taken as a radial basis function. In some applications, particularly in variable selection problems as we shall see later, it is convenient to choose K as tensor sums and products of univariate rbf's, as in SS-ANOVA models, with one important difference: The null space of the penalty functional should only contain at most the constant function. For technical reasons, the SVM may fail to have a unique solution for larger null spaces.

4.2 Nonstandard Support Vector Machines

The previous (standard) SVM, when appropriately tuned, asymptotically implements the Bayes rule, that is, it minimizes the expected cost, when the training set is representative of the population to be classified in the future, and the costs of each kind of misclassification are the same. The nonstandard SVM of (Lin, Lee & Wahba 2002) is a modification of the standard SVM which implements the Bayes rule when neither of these conditions hold. Let π^+ and $\pi^- = 1 - \pi^+$ be prior probabilities of + and − classes, and let π_s^+ and π_s^- be proportions of + and − classes in the training set, and c^+ and c^- be the costs for false + and false − classifications. Let $g^+(x)$ and $g^-(x)$ be the densities for x in the + class and the 1 class respectively. Let $p(x)$ be $Pr[y = 1|x]$ in the population to be classified. Then

$$p(x) = \frac{\pi^+ g^+(x)}{\pi^+ g^+(x) + \pi^- g^-(x)}. \tag{29}$$

Let $p_s(x)$ be $Pr[y = 1|x]$ in a population distributed as the training sample. Then

$$p_s(x) = \frac{\pi_s^+ g^+(x)}{\pi_s^+ g^+(x) + \pi_s^- g^-(x)}. \tag{30}$$

Then the Bayes rule classifies as + when $\frac{p(x)}{1-p(x)} > \frac{c^+}{c^-}$ and − otherwise, equivalently when $\frac{p_s(x)}{1-p_s(x)} > \frac{c^+}{c^-}\frac{\pi_s^+}{\pi_s^-}\frac{\pi^-}{\pi^+}$. Letting $L(-1) = c^+\pi_s^+\pi^-$ and $L(1) = c^-\pi_s^-\pi^+$, the Bayes rule is then equivalent to classifying as + when $sign(p_s - \frac{L(-1)}{L(-1)+L(1)}) > 0$ and − otherwise. The nonstandard SVM finds f of the form

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i)[(1 - y_i f(x(i))]_+ + \lambda\|h\|^2_{H_K} \tag{31}$$

over functions of the form $f(x) = h(x) + b$. It is shown in (Lin, Lee & Wahba 2002) that the nonstandard SVM of (31) is estimating $sign(p_s - \frac{L(-1)}{L(-1)+L(1)})$, again just what you need to implement the Bayes rule.

4.3 Multicategory Support Vector Machines

Many approaches have been proposed to classify into one of k possible classes by using SVMs. A google search as of 2006 for "multiclass support vector machine" or "multicategory support vector machine" gives over 500 hits. For the moment, letting $y_j \in \{1, \cdots, k\}$ and considering the standard situation of equal misclassification costs and representative training sample, if $P(y = j|x) = p_j(x)$ then the Bayes rule assigns a new x to the class with the largest $p_j(x)$. Two kinds of strategies appear in the literature. The first solves the problem via solving several binary problems, one-vs-rest, one-vs-one, and various designs of several-vs-several. See for example (Allwein, Schapire &

Singer 2000)(Dietterich & Bakiri 1995). The second considers all classes at once. Two examples of this are (Crammer & Singer 2000) (Weston & Watkins 1999) with many variants in the recent literature. Many of these methods are highly successful in general practice, but, in general, situations can be found where they do not implement the Bayes rule, see (Lee et al. 2004).

The multicategory SVM (MSVM) of (Lee & Lee 2003) (Lee et al. 2004) goes as follows: First y_i is coded as a k-dimensional vector $(y_{i1}, \cdots y_{ik})$ with 1 in the jth position if y_j is in class j, and $-\frac{1}{k-1}$ in the other positions, thus $\sum_{r=1}^{k} y_{ir} = 0, i = 1, \cdots n$. Let $L_{jr} = 1$ for $j \neq r$ and 0 otherwise. The MSVM solves for a vector of functions $f_\lambda = (f_\lambda^1, \cdots, f_\lambda^k)$, with $f^r(x) = d^r + h^r(x)$, each h^k in $\mathcal{H}_K$ satisfying the *sum-to-zero constraint* $\sum_{r=1}^{k} f^r(x) = 0$ all x, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k} L_{cat(i)r}(f^r(x(i)) - y_{ir})_+ + \lambda\sum_{j=1}^{k}\|h^j\|^2_{\mathcal{H}_K} \tag{32}$$

equivalently

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)}(f^r(x(i)) + \frac{1}{k-1})_+ + \lambda\sum_{j=1}^{k}\|h^j\|^2_{\mathcal{H}_K} \tag{33}$$

where $cat(i)$ is the category of y_i.

It can be shown that $k = 2$ case reduces to the usual 2-category SVM.

The target for the MSVM is shown in (Lee et al. 2004) to be $f(t) = (f^1(t), \cdots, f^k(t))$ with $f^j(t) = 1$ if $p_j(t)$ is bigger than the other $p_l(t)$ and $f^j(t) = -\frac{1}{k-1}$ otherwise, thus implementing an estimate of the Bayes rule. Similar to the two-class case, there is a nonstandard version of the MSVM. Suppose the sample is not representative, and misclassification costs are not equal. Let

$$L_{jr} = (\pi^j/\pi_s^j)c_{jr}, \quad j \neq r \tag{34}$$

where c_{jr} is the cost of misclassifying a j as an r and $c_{rr} = 0 = L_{rr}$. π^j is the prior probability of category j, and π_s^j is the fraction of samples from category j in the training set. Substituting (34) into (32) gives the nonstandard MSVM, and it is shown in (Lee et al. 2004) that the nonstandard MSVM has as its target the Bayes rule. That is, the target is $f_j(x) = 1$ if j minimizes

$$\sum_{\ell=1}^{k} c_{\ell j} p_\ell(x)$$

equivalently

$$\sum_{l=1}^{k} L_{\ell j} p_\ell^s(x),$$

and $f_j(x) = -\frac{1}{k-1}$ otherwise.

To illustrate the use of the MSVM (Lee et al. 2004) revisited the small round blue cell tumors (SRBCTs) of childhood data set in (Khan, Wei, Ringner, Saal, Ladanyi, Westermann, Berthold,

Schwab, Atonescu, Peterson & Meltzer 2001). There are 4 classes: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), and the data were cDNA gene expression profiles. There was a training set of 63 samples, (NB: 12, RMS: 20, BL: 8, EWS: 23), and a test set of 20 SRBCT cases (NB: 6, RMS: 5, BL: 3, EWS: 6) and five non SRBCTs. The gene expression profiles contained observations on 2308 genes, after several preprocessing steps the observations were reduced to those on 100 genes, and the final data set for classification consisted of a vector of three principal components based on the 100 gene observations for each profile. The principal components turned out to contain enough information for nearly perfect classification.

The four class labels are coded according as EWS: $(1, -1/3, -1/3, -1/3)$,
BL: $(-1/3, 1, -1/3, -1/3)$, NB: $(-1/3, -1/3, 1, -1/3)$ and RMS: $(-1/3, -1/3, -1/3, 1)$.

The top four panels in Figure 7 show the predicted decision vectors (f_1, f_2, f_3, f_4) at the test

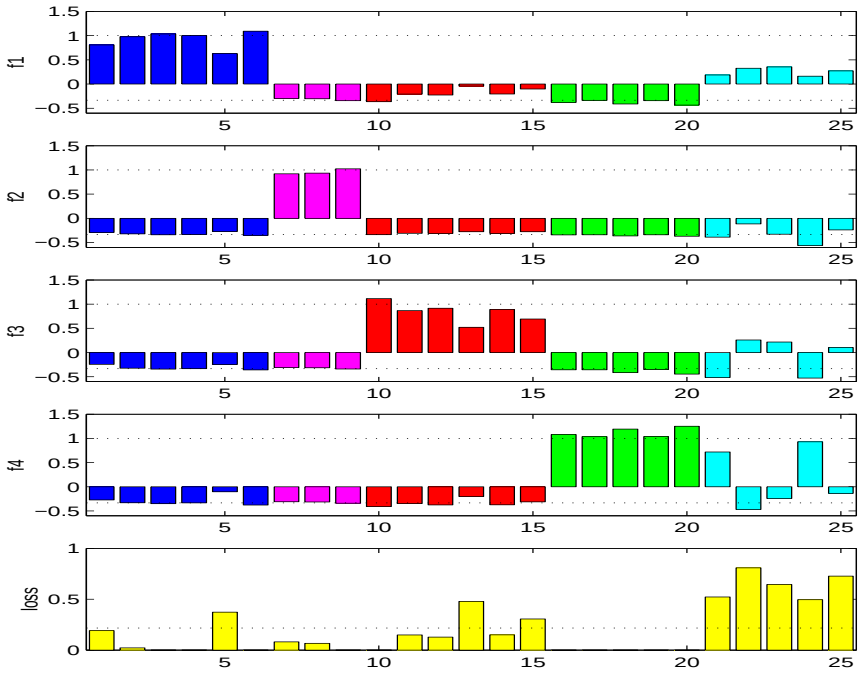

Figure 7: Predicted four dimensional decision vectors for 20 test samples in four classes and 5 test samples in "none of the above". ©Oxford University Press

examples. The first 6 columns are the estimated class labels for the 6 EWS cases in the test set: ideally they will be $(1, -1/3, -1/3, -1/3)$. As can be seen, of these six cases the f_1 bars (top panel) are all close to 1, and in the three next lower panels, the f_2, f_3, f_4 bars are all negative, so that these six members of the test set are all identified correctly. The next three columns are the three

BL cases in the test set, ideally their estimates are $(-1/3, 1, -1/3, -1/3)$ —in the second panel they are all about 1, and in the first, third and fourth panel they are all negative, so that these BL cases are all classified correctly. In the next 6 columns, the 6 members of the NB class are classified correctly, that is, f_3 is close to 1 and the other components are negative, and the next 5 RMS cases are all classified correctly. The last five columns are the 5 nonSRBT cases, and with one exception none of the bars are close to one, with the exceptional case having both f_1 and f_4 positive, leading to a dubious classification ("none of the above"). The bottom panel gives a measure of the weakness of the classification, obtained from a bootstrap argument, and it is suggesting that the classification of all of the "none of the above" cases is weak. Software for the MSVM can be found at `http://www.stat.ohio-state.edu/~yklee/software.html`.

4.4 Support Vector Machines with variable selection

In dealing with classification problems with very large observation vectors such as occur, for example in microarray (gene chip) or SNP data, classification is only part of the problem. It is typically of scientific interest to know which genes out of the thousands obtained from the gene chip data are important for the classification, or, which SNPs from the thousands that are observed, are important. Google provides thousands of hits for "Variable Selection" and SVM. Here we briefly provide the flavor of three recent papers appropriate to these situations. We describe only two-category SVM's, but most of the results generalized to the MSVM.

In (Zhang 2006), f is modeled as a (low order) SS-ANOVA model which can be written:

$$f(x_1, \cdots, x_d) = d + \sum_{\alpha=1}^{d} h_\alpha(x_\alpha) + \sum_{\alpha<\beta} h_{\alpha\beta}(x_\alpha, x_\beta) + \cdots \tag{35}$$

with $h_\alpha \in \mathcal{H}^\alpha$, $h_{\alpha\beta} \in \mathcal{H}^\alpha \otimes \mathcal{H}^\beta \cdots$ and so forth. The proposed SVM optimization problem becomes

$$min_{f\in\mathcal{F}} \sum_{i=1}^{n} [1 - y_i f(x(i))]_+ + \tau[\sum_{\alpha=1}^{d} \|h_\alpha\|_{\mathcal{H}^\alpha} + \sum_{\alpha<\beta} \|h_{\alpha\beta}\|_{\mathcal{H}^\alpha\otimes\mathcal{H}^\beta} + \cdots] \tag{36}$$

where $x = (x_1, \ldots, x_d)$. Note that (36) uses norms rather than squared norms in the penalty functional. This formulation is shown to be equivalent to

$$min_{f\in\mathcal{F}} \sum_{i=1}^{n} [1 - y_i f(x(i))]_+ + [\sum_{\alpha=1}^{d} \theta_\alpha^{-1} \|h_\alpha\|^2_{\mathcal{H}^\alpha} + \sum_{\alpha<\beta} \theta_{\alpha\beta}^{-1} \|h_{\alpha\beta}\|^2_{\mathcal{H}^\alpha\otimes\mathcal{H}^\beta} + \cdots] + \lambda[\sum \theta_\alpha + \sum \theta_{\alpha\beta} + \cdots] \tag{37}$$

where the θs are constrained to be non-negative. Lee et al ((Lee, Kim, Lee & Koo 2006) also considered the approach of (37), in the context of the MSVM of (Lee et al. 2004) and applied the

method to the data of (Khan et al. 2001) that was used there, to select influential genes. The home pages of both these first authors cited contain related software relevant to this problem.

Mukherjee and Wu (Mukherjee & Wu 2006) perform variable selection via an algorithm which learns the gradient of the response with respect to each variable - if the gradient is small enough then the variable is deemed not important. They applied their method to the same two-class leukemia data of (Golub, D.Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, C.Bloomfield & Lander 1999) that was analyzed in (Lee & Lee 2003).

5 Dissimilarity data and kernel estimates

In many problems direct attribute vectors are not known, or are not convenient to deal with, while some sort of pairwise dissimilarity score between pairs of objects in a training set is known. Examples could be subjective pairwise differences between images as provided by human observers, pairwise differences between graphs, strings, sentences, microarray observations, protein sequences, etc. Given pairwise dissimilarity scores we describe two approaches to obtaining a kernel, which can then be used in an SVM for classifying protein sequence data.

5.1 Regularized Kernel Estimation

The Regularized Kernel Estimation (RKE) method (Lu, Keles, Wright & Wahba 2005) goes as follows: Given K, a non-negative definite $n \times n$ matrix, the squared distance $\hat{d}_{ij}$ between the ith and jth object in a set of n objects can be defined by $\hat{d}_{ij}(K) = K(i,i) + K(j,j) - 2K(i,j)$, where $K(i,j)$ is the (i,j) entry of K. Given a set of noisy, possibly incomplete, set of pairwise distances $\{d_{ij}\}$ between n objects, the regularized kernel estimation problem is to find an $n \times n$ non-negative definite matrix which minimizes

$$\min_{K \succeq 0} \sum_{(i,j) \in \Omega} |d_{ij} - \hat{d}_{ij}(K)| + \lambda \, \mathrm{trace}(K). \tag{38}$$

Here Ω is a set of pairwise distances which forms a connected set, that is, a graph connecting the included pairs is connected. This problem can be solved numerically for K by a convex code algorithm, see (Benson & Ye 2004) (Lu et al. 2005) (Tütüncü, Toh & Todd 2003).

Letting $K = K_\lambda$ be the minimizer of (38), the eigenvalues of K_λ are set to zero after the pth largest, resulting in $K_{\lambda,p}$, say. Pseudo data $z(i), i = 1, \cdots, n$ for the n objects can be found by letting $z(i) = (z_1(i), \cdots, z_p(i))$ where $z_\nu(i) = \sqrt{\lambda_\nu}\phi_\nu(i), \nu = 1, \cdots p.$ with the λ_ν and ϕ_ν being the eigenvalues and eigenvectors of $K_{\lambda,p}$. Given labels on (a subset of) the n objects, a support vector machine can be built on the pseudodata. To classify a new object, a "newbie" algorithm is

used to obtain the pseudodata $z(n+1)$ for the $n+1$st object. The newbie algorithm obtains an $(n+1)\times(n+1)$ kernel K_{n+1} of the form

$$\tilde{K}_{n+1} = \begin{bmatrix} K_n & b^T \\ b & c \end{bmatrix} \succeq 0, \tag{39}$$

(where $b \in \mathcal{R}^n$ and c is a scalar) that solves the following optimization problem:

$$\begin{aligned} &\min_{c\geq 0, b} \textstyle\sum_{i\in\Psi} \left| d_{i,n+1} - \hat{d}_{i,n+1}(K_{n+1}) \right| \\ &\text{such that } \; b \in \text{Range}(K_n), \;\; c - b^T K_n^+ b \geq 0, \end{aligned} \tag{40}$$

where K_n^+ is the pseudo-inverse of $K_n = K_{\lambda,p}$ and Ψ is a suitably rich subset of $\{1,2,\ldots,n\}$. Pseudodata $z(n+1)$ is found upon observing that $z(i)^T z(n+1) = K(i,n+1) = b_i$. Figure 8 from (Lu et al. 2005) gives the 280 eigenvalues for K based on dissimilarity scores from protein sequence alignment scores from 280 protein sequences. The eigenvalues of K_λ were truncated after $p = 3$, and a three dimensional black and white plot of the pseudo-data is given in Figure 9. The four classes can be seen, although the original color plot in (Lu et al. 2005) is clearer.

This approach can easily tolerate missing data, in fact only about 36 % of the pairs were used, and it is robust to very noisy or binned dissimilarity data, for example dissimilarity information given on a scale of 1,2,3,4 or 5.

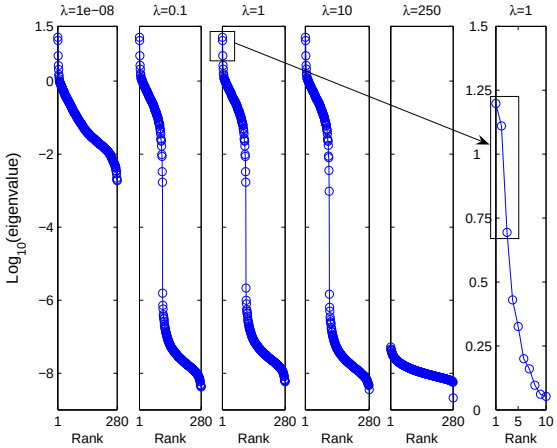

Figure 8: Left five panels: log scale eigensequence plots for five values of λ. As λ increases, smaller eigenvalues begin to shrink. Right panel: first ten eigenvalues of the $\lambda = 1$ case displayed on a larger scale. ©PNAS

The RKE can be used in the semisupervised situation, where the kernel is built on both labeled and unlabeled data, and then used to classify both the unlabeled data used to build it as well as new observations that were not.

Data from multiple sources, some of which may involve dissimilarity data and some direct attribute data, can be combined in an SVM once kernels are given for each source. Let z be a pseudo-attribute vector of length p, obtained from the $n \times n$ kernel K_Z which was derived from dissimilarity data and then had its eigenvalues truncated after the pth, and let x be an attribute vector, with an associated kernel $K_X(x, x')$ to be chosen, (for example a Gaussian kernel). We can define a composite attribute vector as $t^T = (z^T : x^T)$ and build a support vector machine on the domain of the composite attribute vectors based on the kernel $K_\mu(t, t') = \mu_Z K_Z(z, z') + \mu_X K_X(x, x')$, where μ_Z and μ_X are non-negative tuning parameters. $K_Z(z, z') = (z, z')$, the Euclidean inner product, from the way that z was obtained, but some other kernel, for example, a Gaussian or SS-ANOVA kernel could be built on top of the psudodata. Then the (two category) SVM finds d and $c = (c_1, \cdots, c_n)$ to minimize

$$\sum_{i=1}^{n} [1 - y_i f(x(i)]_+ + \lambda c' K_\mu c. \tag{41}$$

as before where

$$f(t) = d + \sum_{i=1}^{n} c_i K_\mu(t(i), t) \tag{42}$$

and $\mu = (\mu_Z, \mu_X)$ are to be chosen. Generalizations to the MSVM can also be defined.

5.2 Kernels from constructed attribute vectors

In (Lanckriet, Cristianini, Bartlett, ElGhoui & Jordan 2004) a detailed study was carried out using data from several sources, including both direct data and dissimilarity data. For dissimilarity data they used a kernel constructed from n-dimensional attribute vectors whose components are themselves dissimilarity measures. The method is described in (Liao & Noble 2003) and elsewhere. It goes as follows: The training set consists of n objects, with $\binom{n}{2}$ dissimilarity scores d_{ij} available between all pairs. The ith object is assigned an n dimensional vector $x(i)$ whose rth component is d_{ir}. Then $K(i, j)$ is defined as $(x(i), x(j))$, where the inner product is the Euclidean inner product.

6 Tuning methods

6.1 Generalized Cross Validation

This article has concentrated on Bernoulli and categorical data, since this kind of data is typically assumed when "Statistical Learning" is the topic. However, to best explain several of the tuning methods used in conjunction with Bernoulli and categorical data, it is easiest to begin by describing tuning for nonparametric function estimation with Gaussian data. The model is

$$y_i = f(x(i)) + \epsilon_i, \quad i = 1, \cdots, n \tag{43}$$

where $x \in \mathcal{T}$ (some domain), $f \in \mathcal{H}_K$ and the ϵ_i are i.i.d Gaussian random variables with common unknown variance σ^2. The estimate f_λ is obtained as the solution to the problem find $f \in \mathcal{H}_K$ to minimize

$$I_\lambda\{y, f\} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x(i)))^2 + \lambda J(f) \tag{44}$$

where $J(f) = \|f\|^2_{\mathcal{H}_K}$ or a seminorm in $\mathcal{H}_K$. The target for choosing λ is to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(f(x(i) - f_\lambda(x(i)))^2 \tag{45}$$

where f is the "true" f in the model. The GCV (Generalized Cross Validation) to be described (Craven & Wahba 1979)(Golub et al. 1979) is derived from a leaving-out- one estimate for λ which goes as follows: Let $f_\lambda^{[-k]}(x(k))$ be the estimate of f based on the data omitting the k th data point. The leaving-out-one function $V_0(\lambda)$ is defined as

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda^{[-k]}(x(k)))^2 \tag{46}$$

and the minimizer of V_0 is the leaving-out-one estimate. Let $A(\lambda)$ be the $n \times n$ influence matrix, which satisfies

$$(f_\lambda(x(1)), \cdots, f_\lambda(x(n))^T = A(\lambda)(y_1, \cdots, y_n)^T, \tag{47}$$

which exists since the estimate is linear in the data. It is known from the leaving-out-one lemma (Craven & Wahba 1979) that

$$V_0(\lambda) \equiv \frac{1}{n}\sum_{k=1}^{n}\frac{(y_k - f_\lambda(x(k))^2}{(1 - a_{kk}(\lambda))^2} \tag{48}$$

where the $a_{kk} \in (0, 1)$ are the diagonal elements of $A(\lambda)$. The GCV function $V(\lambda)$ is obtained by replacing each a_{kk} in (48) by their average, namely $\frac{1}{n}traceA(\lambda)$, to get

$$V(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - f_\lambda(x(i))^2}{(1 - \frac{1}{n}trA(\lambda))^2} \tag{49}$$

and the estimate of λ is the minimizer of $V(\lambda)$. Theoretical properties are discussed in (Li 1986), and the important randomized trace technique for calculating $trA(\lambda)$ can be found in (Girard 1989) (Girard 1995) (Hutchinson 1989). A different calculation method is found in (Golub & vonMatt 1997). For comparison to the methods described below, we note that when $I_\lambda\{y, f\}$ is as in (44), that is, J is a quadratic form in $(f_\lambda(x(i), \cdots f_\lambda(x(n))$ then $A(\lambda)$ is the inverse Hessian of I_λ of (44) with respect to $f_i \equiv f_\lambda(x(i), i = 1, \cdots, n$.

6.2 Generalized Approximate Cross Validation, Bernoulli data, RKHS penalties

The GACV (Generalized Approximate Cross Validation) for Bernoulli data and reproducing kernel squared norms or seminorms as penalties was provided in (Xiang & Wahba 1996). As in Section 2 I_λ is of the form

$$I_\lambda\{y, f\} = \frac{1}{n}\sum_{i=1}^{n} -y_i f(x(i)) + log(1 + e^{f(x(i))}) + \lambda J(f), \tag{50}$$

where $J(f)$ is a squared norm or seminorm in an RKHS. The target for the GACV is the expected value of the so-called Comparative Kullback Liebler distance (CKL) between the true and estimated probability distribution, and is

$$CKL\lambda) = \frac{1}{n}\sum_{i=1}^{n} -p(x(i))f_\lambda(x(i)) + log(1 + e^{f_\lambda(x(i))}) \tag{51}$$

where $p(x)$ is the true but unknown probability that $y = 1|x$. The leaving-out-one estimate of the CKL is

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n} -y_k f_\lambda^{[-k]}(x(k)) + log(1 + e^{f_\lambda(x(k)})). \tag{52}$$

The GACV is obtained from $V_0(\lambda)$ by a series of approximations followed by averaging over the diagonal elements of a matrix which plays the role of the influence matrix, and the result is

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} -y_i f)_\lambda(x(i)) + log(1 + e^{f_\lambda(x(i))}) + \frac{1}{n} trA(\lambda)\frac{\sum_{i=1}^{n} y_i(y_i - p_\lambda(x(i)))}{(n - trW^{1/2}A(\lambda)W^{1/2})}. \tag{53}$$

Here $A(\lambda) = A(\lambda, f_\lambda)$ is the inverse Hessian of $I_\lambda\{y, f\}$ with respect to $f_i \equiv f_\lambda(x(i)), i = 1, \cdots, n$, and $W = W(\lambda, f_\lambda)$ is the diagonal matrix with iith entry $p_\lambda(x(i))(1-p_\lambda(x(i))$, which is the variance of the estimated Bernoulli distribution as well as the second derivative of $log(1 + e^{f_\lambda(x(i))})$. Figure 10 from (Xiang & Wahba 1996)

gives two plots comparing the true $CKL(\lambda)$ with $GACV(\lambda)$ in a simulation experiment where $p(x)$ is known. Numerous experimental works show that the minimizer of the $GACV$ provides a

good estimate of the minimizer of the CKL, but theoretical results analogous to those in (Li 1986) for GCV remain to be found. A generalization of the GACV to the two-eye problem of Section 2.5 based on leaving-out-one-unit is found in (Gao et al. 2001).

6.3 Generalized Approximate Cross Validation, Bernoulli data, l_1 penalties

A general version of GACV targeted at the CKL adapted for LASSO-type optimization problems appears in (Zhang et al. 2004). A special case, for optimization problems like that of the LASSO-Patternsearch (Shi et al. 2006) goes as follows. For each trial value of λ, there will be, say, $N = N(\lambda)$ basis functions in the model with non-zero coefficients. Let B be the $n \times N$ design matrix for the N basis functions and W be as before. Let $A(\lambda, f_\lambda) = B(B^T W B)^{-1} B^T$ and observe that $trW^{1/2}A(\lambda)W^{1/2} = N$. The GACV becomes

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} -y_i f(x(i)) + log(1 + e^{f(x(i))} + \frac{1}{n} trA(\lambda)\frac{\sum_{i=1}^{n} y_i(y_i - p_\lambda(x(i)))}{(n-N)}. \qquad (54)$$

6.4 Support Vector Machines

A large number of competing methods have been proposed for tuning SVMs. When sufficiently large data sets are available, a common practice is to divide the data into three parts: a training set, a tuning set for choosing λ and any other tuning parameters, and a test set for evaluating the results. Five-fold and ten-fold cross validation are both popular. Several tuning methods related in some way to cross validation ideas are described in (Chapelle, Vapnik, Bousquet & Mukherjee 2002) (Gold & Sollich 2003). Tuning methods based on structural risk minimization appear in (Lanckriet et al. 2004). A perturbation method which perturbs both inputs and outputs is proposed in (Wang & Shen 2006). A popular method is Joachims' XiAlpha method (Joachims 2000), which is part of the SVM^{light} package at `http://svmlight.joachims.org/`. A GACV method was derived in (Wahba 1999) by methods analogous to those in Section 6.2. The XiAlpha and GACV methods are seen to be related (Wahba, Lin, Lee & Zhang 2001), where a generalization of both methods to the nonstandard case is proposed. A GACV for the multicategory support vector machine of Lee, Lin and Wahba is in (Lee et al. 2004).

6.5 Regularized Kernel Estimates

A leaving out pairs algorithm can be obtained to choose λ in the RKE estimate, although K_λ appears to be insensitive to λ over a fairly broad range. To date the choice of p has been made

visually by plotting eigenvalues, but when the pseudo-data is used for classification one possibility is to chose it simultaneously with the SVM parameters. A definitive automatic procedure is yet to be obtained.

7 Regularization, Empirical Bayes, Gaussian Processes Priors and Reproducing Kernels

It is well known that there is a duality between zero mean Gaussian processes and RKHS: For every positivie definite function K there is a unique RKHS with K as its reproducing kernel, and for every positive definite function K there is an associated zero mean Gaussian process prior with K as its covariance, see (Aronszajn 1950) (Kimeldorf & Wahba 1971) (Parzen 1970) (Wahba 1990). When the first term in the optimization problem is a negative log likelihood $\mathcal{L}\{y, f\}$ and the penalty term involves RKHS squared norms then for fixed tuning parameters the estimate is a Bayes estimate with a Gaussian Process prior. These remarks extend to the case when the penalty term involves squared seminorms, which correspond to an improper prior, see (Kimeldorf & Wahba 1971) (Wahba 1990). Similarly, in the LASSO class of estimates, the l_1 penalty corresponds to negative exponential priors on the coefficients. In typical regularization methods like those described here the tuning parameters are chosen by generalization and model selection arguments, in "frequentist" style. There is a large literature labeled Empirical Bayes methods, as well as Gaussian Process Priors methods, and the discerning reader may consider the relationships between those and regularization methods.

Acknowledgments

Grace Wahba's research is at the time of this writing supported by NSF Grants 0505636 and 0604572, NIH Grant EY09946, and ONR Grant N00014-06-1-0095.

The author wishes to acknowledge her gratitude to her collaborators, former students, and present students. Former and present students are listed on her home page. Special help of David Callan is acknowledged.

References

Allwein, E. L., Schapire, R. E. & Singer, Y. (2000), Reducing multiclass to binary: A unifying approach for margin classifiers, *in* 'Proc. 17th International Conf. on Machine Learning', Morgan Kaufmann, San Francisco, CA, pp. 9–16.

Aronszajn, N. (1950), 'Theory of reproducing kernels', *Trans. Am. Math. Soc.* **68**, 337–404.

Benson, S. & Ye, Y. (2004), DSDP5: A software package implementing the dual-scaling algorithm for semidefinite programming, Technical Report ANL/MCS-TM-255, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL.
*http://www-unix.mcs.anl.gov/~benson/dsdp/dsdp5userguide.pdf

Bookstein, F. (1997), *Morphometric Tools for Landmark Data: Geometry and Biology*, Cambridge University Press.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. & Haussler, D. (2000), 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proceedings of the National Academy of Science* **97**, 262–267.

Carew, J., Dalal, R., Wahba, G. & Fain, S. (2004), A nonparametric method for estimation of arterial wall shear stress, *in* 'Proceedings Intl Soc Mag Reson Med 12', International Society for Magnetic Resonance in Medicine, Berkeley CA, p. 1924.

Chapelle, O., Vapnik, V., Bousquet, O. & Mukherjee, S. (2002), 'Choosing multiple parameters for support vector machines', *Machine Learning* **46**, 131–159.

Chen, S., Donoho, D. & Saunders, M. (1998), 'Atomic decomposition by basis pursuit', *SIAM J. Sci. Comput.* **20**, 33–61.

Chun, H. (2006), 'Smoothing spline ANOVA model for bivariate Bernoulli observations', Abstract, JSM 2006 program.

Crammer, K. & Singer, Y. (2000), On the learnability and design of output codes for multiclass problems, *in* 'Computational Learing Theory', pp. 35–46.
*citeseer.nj.nec.com/article/crammer00learnability.html

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation', *Numer. Math.* **31**, 377–403.

Cristianini, N. & Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge University Press.

Davidson, L. (2006), 'Compring tongue shapes from ultrasound imaging using smoothing spline analysis of variance', *J. Acoust. Soc. Am.* **120**, 407–415.

deBoor, C. (1978), *A Practical Guide to Splines*, Springer-Verlag, New York.

Dietterich, T. G. & Bakiri, G. (1995), 'Solving multiclass learning problems via error-correcting output codes', *Journal of Artificial Intelligence Research* **2**, 263–286.

Duchon, J. (1977), Splines minimizing rotation-invariant semi-norms in Sobolev spaces, *in* 'Constructive Theory of Functions of Several Variables', Springer-Verlag, Berlin, pp. 85–100.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Ann. Statist.* **32**, 407–499.

Evgeniou, T., Pontil, M. & Poggio, T. (2000), 'Regularization networks and support vector machines', *Advances in Computational Mathematics* **13**, 1–50.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Gao, F., Wahba, G., Klein, R. & Klein, B. (2001), 'Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion', *J. Amer. Statist. Assoc.* **96**, 127–160.

Girard, D. (1989), 'A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data', *Numer. Math.* **56**, 1–23.

Girard, D. (1995), 'The fast Monte-carlo cross-validation and C_L procedures: Comments, new results and application to image recovery problems', *Computational Statistics* **10**, 205–231.

Gold, C. & Sollich, P. (2003), 'Model selection for support vector machine classification', *Neurocomputing* **55**, 221–249.

Golub, G., Heath, M. & Wahba, G. (1979), 'Generalized cross validation as a method for choosing a good ridge parameter', *Technometrics* **21**, 215–224.

Golub, G. & vonMatt, U. (1997), 'Generalized cross-validation for large-scale problems', *J. Comput. Graph. Statist.* **6**, 1–34.

Golub, T., D.Slonim, Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., C.Bloomfield & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer.

Gu, C. & Wahba, G. (1991), 'Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method', *SIAM J. Sci. Statist. Comput.* **12**, 383–398.

Gu, C. & Wahba, G. (1993*a*), 'Semiparametric analysis of variance with tensor product thin plate splines', *J. Royal Statistical Soc. Ser. B* **55**, 353–368.

Gu, C. & Wahba, G. (1993*b*), 'Smoothing spline ANOVA with component-wise Bayesian "confidence intervals"', *J. Computational and Graphical Statistics* **2**, 97–117.

Gunn, S. & Kandola, J. (2002), 'Structural modelling with sparse kernels', *Machine Learning* **48**, 137–163.

Hastie, T. & Tibshirani, R. (1986), 'Generalized additive models', *Statistical Science* **1**, 297–318.

Hutchinson, M. (1989), 'A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines', *Commun. Statist.-Simula.* **18**, 1059–1076.

Joachims, T. (1999), Making large-scale svm learning practical, *in* B. Scholkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods-Support Vector Learning', MIT Press, pp. 69–88.

Joachims, T. (2000), Estimating the generalization performance of an SVM efficiently, *in* 'Proceedings of the International Conference on Machine Learning', Morgan Kaufman, San Francisco.

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Atonescu, C., Peterson, C. & Meltzer, P. (2001), 'Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks', *Nature Medicine* **7**, 673–679.

Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *J. Math. Anal. Applic.* **33**, 82–95.

Klein, R., Klein, B. E. K., Linton, K. & DeMets, D. (1991), 'The Beaver Dam eye study: Visual acuity', *Ophthalmology* **98**, 1310–1315.

Knight, K. & Fu, W. (2000), 'Asymptotics for LASSO-type estimators', *Ann. Statist* **28**, 1356–1378.

Lanckriet, G., Cristianini, N., Bartlett, P., ElGhoui, L. & Jordan, M. (2004), 'Learning the kernel matrix with semidefinite programming', *J. Mach. Learn. Res.* **5**, 27–72.

Lee, Y., Kim, Y., Lee, S. & Koo, J. (2006), 'Structured multicategory support vector machines with analysis of variance decomposition', *Biometrika* **93**, 555–571.

Lee, Y. & Lee, C.-K. (2003), 'Classification of multiple cancer types by multicategory support vector machines using gene expression data', *Bioinformatics* **19**, 1132–1139.

Lee, Y., Lin, Y. & Wahba, G. (2004), 'Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data', *J. Amer. Statist. Assoc.* **99**, 67–81.

Leng, C., Lin, Y. & Wahba, G. (2006), 'A note on the LASSO and related procedures in model selection', *Statistica Sinica* **16**, 1273–1284.

Li, K. C. (1986), 'Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing', *Ann. Statist.* **14**, 1101–1112.

Liao, L. & Noble, W. (2003), 'Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships', *J. Computational Biology* **10**, 857–868.

Lin, X. (1998), Smoothing spline analysis of variance for polychotomous response data, Technical Report 1003, PhD thesis, Department of Statistics, University of Wisconsin, Madison WI. Available via G. Wahba's website.

Lin, Y. (2000), 'Tensor product space ANOVA models', *Ann. Statist* **28**, 734–755.

Lin, Y. (2001), 'A note on margin-based loss functions in classification', *Statistics and Probability Letters* **68**, 73–82.

Lin, Y. (2002), 'Support vector machines and the Bayes rule in classification', *Data Mining and Knowledge Discovery* **6**, 259–275.

Lin, Y., Lee, Y. & Wahba, G. (2002), 'Support vector machines for classification in nonstandard situations', *Machine Learning* **46**, 191–202.

Lin, Y., Wahba, G., Zhang, H. & Lee, Y. (2002), 'Statistical properties and adaptive tuning of support vector machines', *Machine Learning* **48**, 115–136.

Lu, F., Keles, S., Wright, S. & Wahba, G. (2005), 'A framework for kernel regularization with application to protein clustering', *Proceedings of the National Academy of Sciences* **102**, 12332–12337. Open Source at www.pnas.org/cgi/content/full/102/35/12332.

Luo, Z. & Wahba, G. (1997), 'Hybrid adaptive splines', *J. Amer. Statist. Assoc.* **92**, 107–114.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models, Second Edition*, Chapman and Hall.

Meinguet, J. (1979), 'Multivariate interpolation at arbitrary points made simple', *J. Appl. Math. Phys. (ZAMP)* **30**, 292–304.

Mukherjee, S. & Wu, Q. (2006), 'Estimation of gradients and coordinate covariation in classification', *J. Machine Learning Research* **7**, 2481–2514.

O'Sullivan, F., Yandell, B. & Raynor, W. (1986), 'Automatic smoothing of regression functions in generalized linear models', *J. Amer. Statist. Assoc.* **81**, 96–103.

Park, M. & Hastie, T. (2007), 'Penalized logistic regression for detecting gene interactions', manuscript.

Parzen, E. (1970), Statistical inference on time series by rkhs methods, *in* R. Pyke, ed., 'Proceedings 12th Biennial Seminar', Canadian Mathematical Congress, Montreal. 1-37.

Ruczinski, I., Kooperberg, C. & LeBlanc, M. (2002), Logic regression — methods and software, *in* D. Denison, M. Hansen, C. Holmes, B. Mallick & B. Yu, eds, 'Nonlinear Estimation and Classification', Springer, pp. 333–344.

Scholkopf, B., Burges, C. & Smola, A. (1999), *Advances in Kernel Methods-Support Vector Learning*, MIT Press.

Scholkopf, B. & Smola, A. (2002), *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press.

Scholkopf, B., Tsuda, K. & J-P.Vert (2004), *Kernel Methods in Computational Biology*, MIT Press.

Shawe-Taylor, J. & Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press.

Shi, W., Wahba, G., Lee, K., Klein, R. & Klein, B. (2006), LASSO-Patternsearch algorithm with application to ophthalmalogy data, Technical Report 1131, Department of Statistics, University of Wisconsin, Madison WI.

Smola, A., Bartlett, P., Scholkopf, B. & Schuurmans, D. (2000), *Advances in Large Marin Classifiers*, MIT Press.

Stein, M. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer-Verlag.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. Roy. Stat. Soc, B* **58**, 267–288.

Tütüncü, R. H., Toh, K. C. & Todd, M. J. (2003), 'Solving semidefinite-quadratic-linear programs using SDPT3', *Mathematical Programming* **95**(2), 189–217.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer.

Wahba, G. (1981), 'Numerical experiments with the thin plate histospline', *Commun. Statist.-Theor. Meth.* **A10**, 2475–2514.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G. (1999), Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, *in* B. Scholkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods-Support Vector Learning', MIT Press, pp. 69–88.

Wahba, G. (2002), 'Soft and hard classification by reproducing kernel Hilbert space methods', *Proceedings of the National Academy of Sciences* **99**, 16524–16530.

Wahba, G., Lin, Y., Lee, Y. & Zhang, H. (2001), On the relation between the GACV and Joachims' $\xi\alpha$ method for tuning support vector machines, with extensions to the non-standard case, Technical Report 1039, Statistics Department University of Wisconsin, Madison WI.

Wahba, G., Lin, Y. & Zhang, H. (2000), Generalized approximate cross validation for support vector machines, *in* A. Smola, P. Bartlett, B. Scholkopf & D. Schuurmans, eds, 'Advances in Large Margin Classifiers', MIT Press, pp. 297–311.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *Ann. Statist.* **23**, 1865–1895. Neyman Lecture.

Wahba, G. & Wendelberger, J. (1980), 'Some new mathematical methods for variational objective analysis using splines and cross-validation', *Monthly Weather Review* **108**, 1122–1145.

Wang, J. & Shen, X. (2006), 'Estimation of generalization error: random and fixed inputs', *Statistica Sinica* **16**, 569–588.

Wang, Y. (1998), 'Mixed-effects smoothing spline ANOVA', *J. Roy. Statist. Soc. B* **60**, 159–174.

Wang, Y., Ke, C. & Brown, M. (2003), 'Shape-invriant modeling of circadian rhythms with random effects and smoothing spline anova decompositions', *Biometrics* **59**, 241–262.

Weston, J. & Watkins, C. (1999), Support vector machines for multiclass pattern recognition, *in* 'Proceedings of the Seventh European Symposium On Artificial Neural Networks'. *citeseer.nj.nec.com/article/weston99support.html

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics*, Wiley.

Xiang, D. & Wahba, G. (1996), 'A generalized approximate cross validation for smoothing splines with non-Gaussian data', *Statistica Sinica* **6**, 675–692.

Xiang, D. & Wahba, G. (1997), Approximate smoothing spline methods for large data sets in the binary case, Technical Report 982, Department of Statistics, University of Wisconsin, Madison WI. Proceedings of the 1997 ASA Joint Statistical Meetings, Biometrics Section, pp 94-98 (1998).

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *J. Roy. Statist. Soc. B* **68**, 49–67.

Zhang, H. (2006), 'Variable selection for SVM via smoothing spline ANOVA', *Statistica Sinica* **16**, 659–674.

Zhang, H. & Lin, Y. (2006*a*), 'Component selection and smoothing for nonparametric regression in exponential families', *Statistica Sinica* **16**, 1021–1042.

Zhang, H. & Lin, Y. (2006*b*), 'Component selection and smoothing in multivariate nonparametric regression', *Ann. Statist.* **34**, 2272–2297.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. & Klein, B. (2004), 'Variable selection and model building via likelihood basis pursuit', *J. Amer. Statist. Assoc.* **99**, 659–672.

Zhu, J. & Hastie, T. (2003), 'Classification of gene microarrays by penalized logistic regression', *Biostatistics* **5**, 427–443.

Zou, H. (2006), 'The adative LASSO and its oracle properties', *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. Roy. Statist. Soc. B* **67, Part 2**, 301–320.

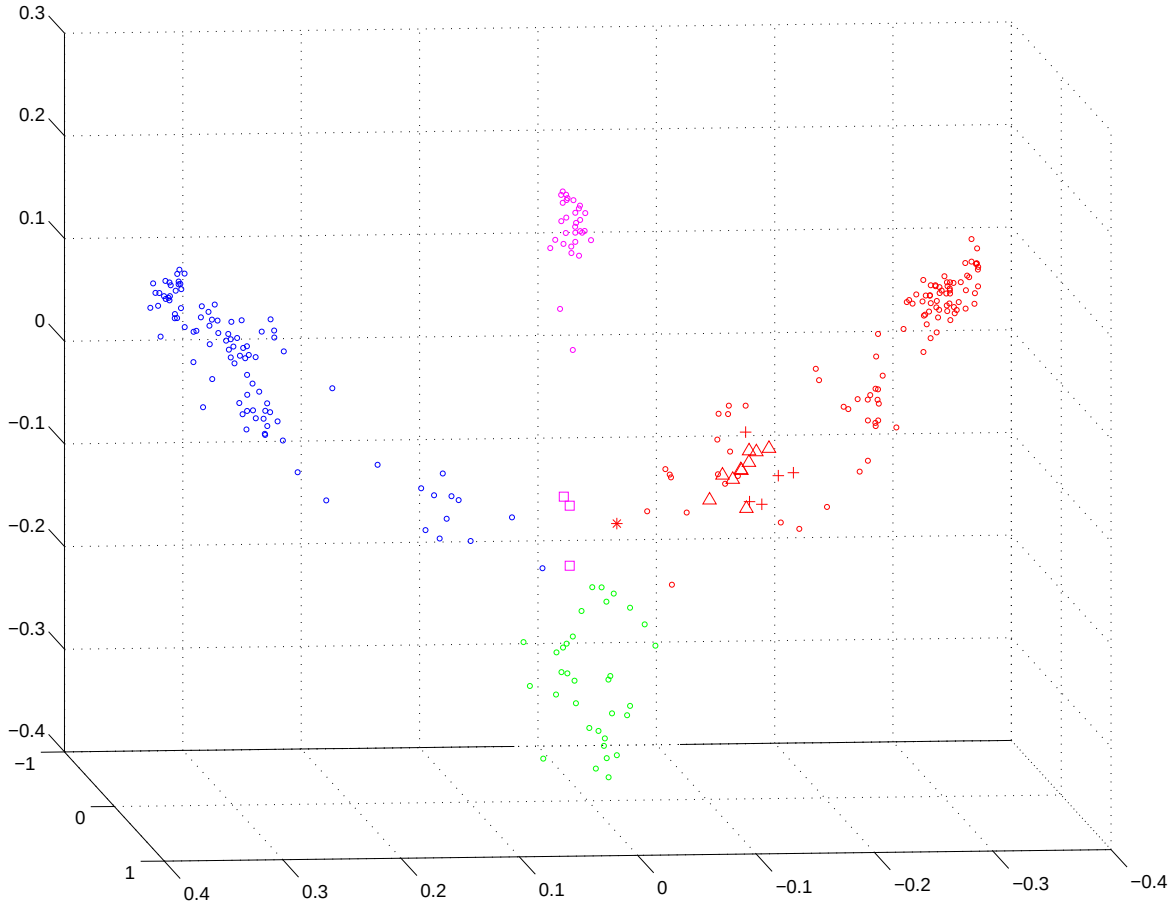

Figure 9: 3D representation of the sequence space for 280 proteins from the globin family.©PNAS

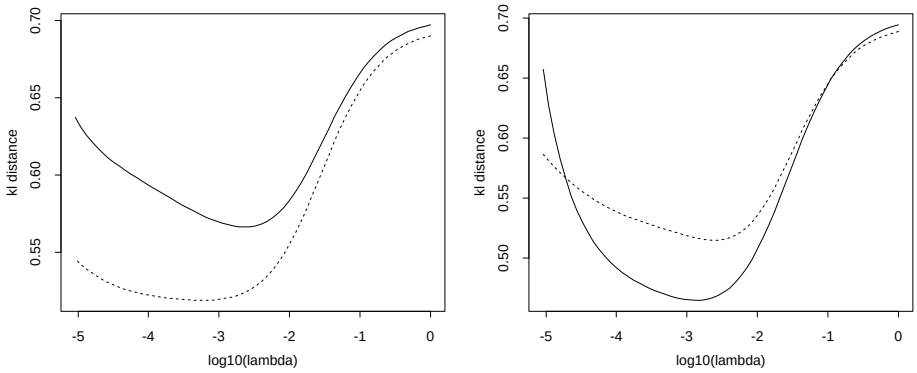

Figure 10: Two $GACV(\lambda)$ (solid lines) and $CKL(\lambda)$ (dotted lines) curves. ©Statistica Sinica