Stanford 50: State of the Art and Future Directions of Computational Mathematics and Numerical Computing

March 29-31, 2007

Stanford University

Grace Wahba

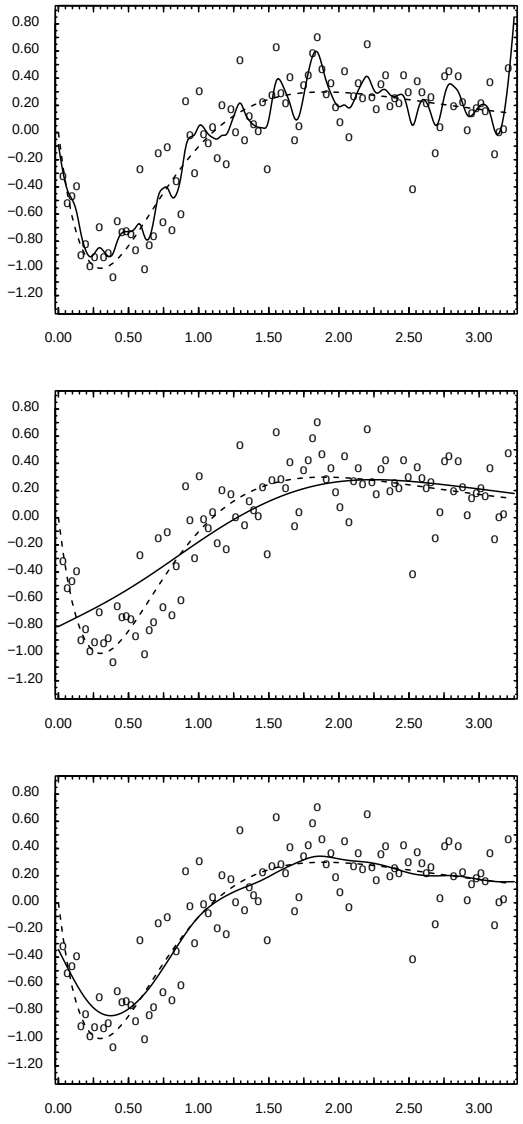
A Statistician's Debt to Numerical Analysts

These slides at
`http://www.stat.wisc.edu/~wahba/` $\rightarrow$ TALKS

Papers/preprints at
`http://www.stat.wisc.edu/~wahba/` $->$ TRLIST

3.18

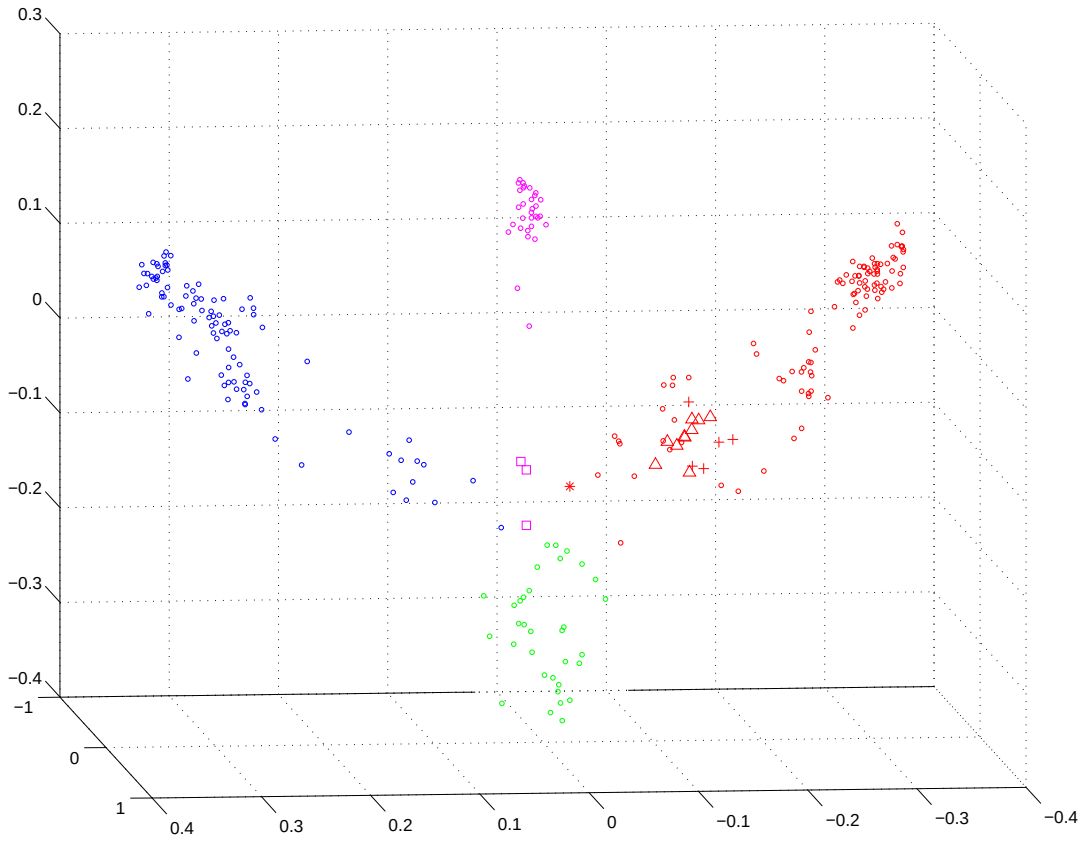
Statisticians, including this one owe a huge debt to numerical analysts. Where would we be without the Singular Value Decomposition, Spline Algorithms, Matrix Computations (Golub and Van Loan)? After briefly noting my collaboration with Gene and Michael Heath on Generalized Cross Validation (1979) which laid the foundation for much later work, I will describe some more recent work of my own and collaborators which relies on mathematical programming and convex cone algorithms for the numerical solution of large optimization problems. These include ♣ Regularized Kernel Estimation for data sets with dissimilarity data rather than attribute data, and ♠ the LASSO-Patternsearch algorithm for finding patterns of high order interactions in risk factor models with large and extremely large attribute vectors. These are special cases of ♣♠ regularization algorithms.



The Cross Validated Smoothing Spline, λ too big, λ too small, and λ just right.

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

♣ Regularized Kernel Estimation.



3D representation of the sequence space for 280 proteins from the globin family. Red:α-globins, blue: β-globins, purple: myoglobins, green: a heterogeneous group of proteins from other small subfamilies within the globin family. Note that there are no units on the axes. The coordinate system has been derived from (noisy, scattered, incomplete) dissimilarity data (BLAST scores) via Regularized Kernel Estimation (RKE).

♣ Dissimilarity Information and RKE

Given a set of N objects, suppose we have obtained a measure of dissimilarity, d_{ij}, for certain object pairs (i, j). Regularized Kernel Estimation (RKE): Finds K:

$$\min_{K \in S_N} \sum_{(i,j) \in \Omega} L\Big(d_{ij}, \widehat{d}_{ij}(K)\Big) + \lambda J(K), \qquad (1)$$

S_N is the convex cone of all real nonnegative definite matrices of dimension N, Ω is the set of pairs with dissimilarity information d_{ij}. L measures the discrepancy between the observed and induced dissimilarity, where the induced dissimilarity $\widehat{d}_{ij}$ is

$$\widehat{d}_{ij} = K(i, i) + K(j, j) - 2K(i, j),$$

$K(i, j)$ being the (i, j) entry of K. L and J are convex in K and λ is a tuning parameter balancing fit to the data and the penalty or complexity on K.

No restrictions on the set of pairs other than requiring that the graph of the objects with pairs connected by edges be connected.

Observed dissimilarity information may be incomplete, may not satisfy the triangle inequality, may be noisy. It also may be crude, as for example when it encodes a small number of coded levels such as “very close”, “close”, “distant”, and “very distant”. A dimension is not specified in advance.

F. Lu, S. Keles, S. Wright, and G. Wahba. A framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102:12332–12337, 2005.

♣ Special Case: l_1 loss and trace penalty

The special case used here was the l_1 loss function and trace penalty:

$$\min_{K \succeq 0} \sum_{(i,j) \in \Omega} |d_{ij} - \widehat{d}_{ij}(K)| + \lambda \, \text{trace}(K). \qquad (2)$$

This formulation can be posed as a special case of a general convex cone optimization problem for which efficient software is available. The sum of squares loss function with trace penalty can also be solved with convex cone software. The problem needs to be solved for a range of λ. Objective methods for choosing λ are under study.

The solution of a general convex cone problem can be obtained numerically using publicly available software such as SDPT3 or DSDP5.

SDPT3: Tütüncü, R. H., Toh, K. C. & Todd, M. J. (2003) *Mathematical Programming* **95**, 189–217.

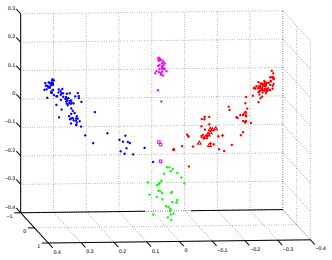
DSDP5: Benson, S. J. & Ye, Y. (2004) DSDP5: A software package implementing the dual-scaling algorithm for semidefinite programming, (Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL), Technical Report ANL/MCS-TM-255.

Setting all the eigenvalues of K after the largest p to 0 results in the p dimensional coordinates of the ith object as

$$x(i) = (\sqrt{\lambda_1}\phi_{i1}, \cdots, \sqrt{\lambda_p}\phi_{ip})$$

where the λ_ν are the first p eigenvalues, and the $\phi_{i\nu}$ are the ith components of the νth eigenvector.

The figure plotted each of the 280 proteins in $p = 3$ dimensional space.



The four differenty classes of proteins here are easily separated to a high degree of accuracy based on their first three coordinates. We can thus build a multicategory support vector machine (another optimization problem!) on this training data to classify future objects. To do this we need to be able to enter a new, unlabeled protein into this coordinate system.

♣ The ”Newbie” Algorithm

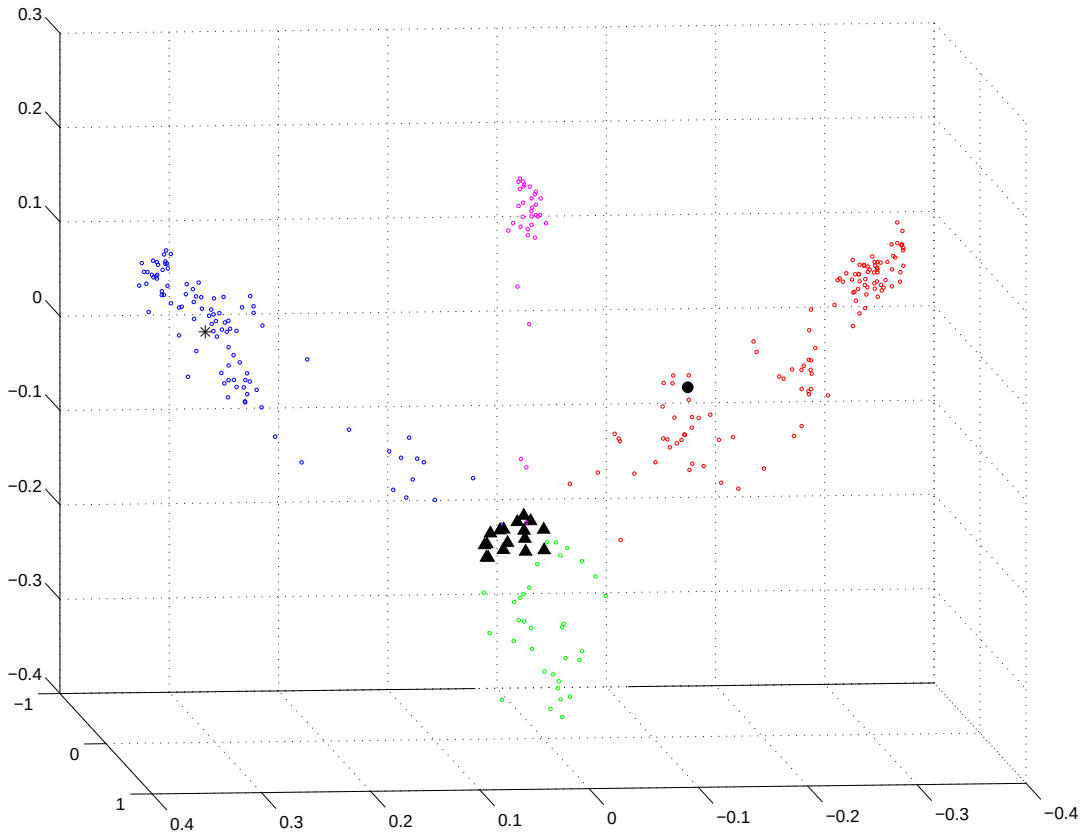
Suppose a solution K_N has been found for a ”training” set of N objects. To account for a new object, we wish to augment the optimal kernel (by one row and column), without changing any of its existing elements. That is, find a new “pseudo-optimal” kernel $\tilde{K}_{N+1}$ of the form

$$\tilde{K}_{N+1} = \begin{bmatrix} K_N & b^T \\ b & c \end{bmatrix} \succeq 0,$$

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

$$\min_{b,c \geq 0} \sum_{i \in \Omega_{newbie}} \left| d_{i,N+1} - \hat{d}_{i,N+1}(K_{N+1}) \right|$$
$$\text{s.t. } b \in \text{Range}(K_N), \quad c - b^T K_N^{\dagger} b \geq 0.$$

$\widehat{d}_{i,N+1}$ is the distance of the newbie to the ith object in the training set, thus, once this is given, the coordinates of the newbie are found easily.



Positioning test globin sequences in the coordinate system of 280 training sequences from the globin family. The newbie algorithm is used to locate one Hemoglobin zeta chain (black circle), one Hemoglobin theta chain (black star), and seventeen Leghemoglobins (black triangles) into the coordinate system of the training globin sequence data.

♠ The LASSO-Patternsearch Algorithm.

Applied to to "progression of myopia" from the Beaver Dam Eye Study, BDES 1 to BDES2, $n = 876$ records of persons aged 60-69 at BDES1. A person whose 'worse eye' scored at a decrease of .75 Diopters or more is labeled $y = 1$, and 0 otherwise. Which variables or clusters of variables are predictive of this outcome?

Table 1: Trial Variables and Cutpoints

	variable	description	binary cut point (higher risk) $X = 1$)
X_1	sex	sex	Male
X_2	inc	income	< 30
X_3	jomyop	juvenile myopia	< 21
X_4	catct	cataract	4-5
X_5	pky	packyear	>30
X_6	asa	aspirin	not taking
X_7	vtm	vitamin	not taking

There are 2^7 possible subsets (clusters) of variables that could be important.

Given $\{y, x = (x_1, \cdots, x_p)\}, y \in \{0, 1\}, x_j \in \{0, 1\}$ for n people:

$$p(x) = Prob(y = 1|x) = e^{f(x)}/(1 + e^{f(x)})$$

$$f(x) = \sum_{\ell=1}^{N} c_\ell B_\ell(x)$$

$$minC(y, f) + \lambda \sum_{\ell=1}^{N} |c_\ell|,$$

where $C(y, f)$ is the negative log likelihood:

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

For the LASSO-Patternsearch the basis functions will be all products of the x_r up to order q:

$$B_{j_1,j_2,..,j_r}(x) = \prod x_{j_1} x_{j_2} ... x_{j_r}, r = 1, \cdots, q.$$

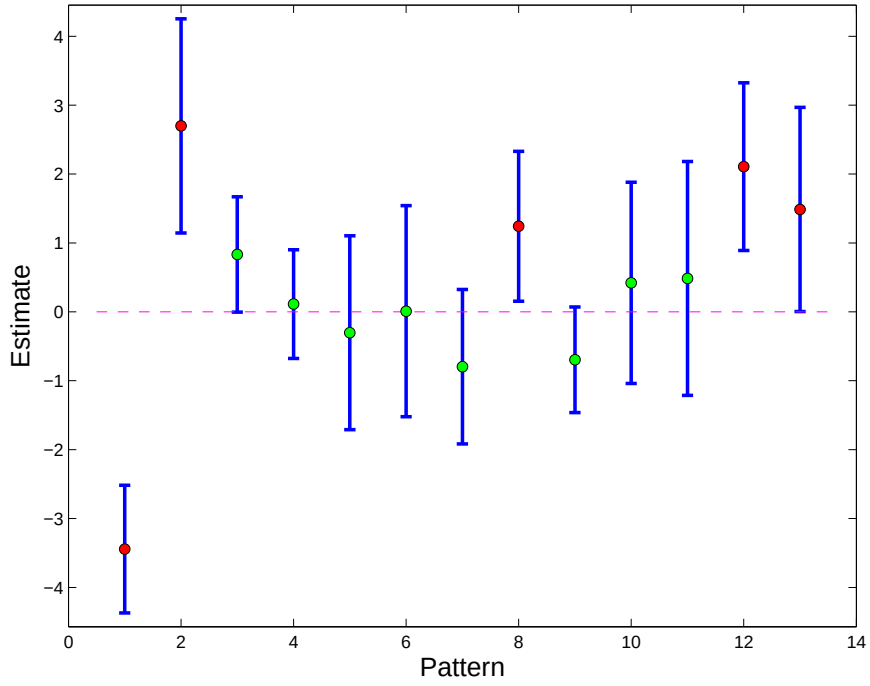
Thus, $B_{j_1,j_2,...,j_r}(x) = 1$ if x is a p-vector which has ones in each of the $j_1, j_2, \cdots, j_r$ positions, and $B_{j_1,...,j_r}(x) = 0$ otherwise. The number N of basis functions is then

$$N = \binom{p}{0} + \binom{p}{1} + \binom{p}{2} + ... + \binom{p}{q}.$$

For $q = p$, (all possible patterns), $N = 2^p$.

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

Patterns and their confidence intervals:



Pattern Number

Significant patterns:

1. Constant
2. catct (Cataract)
8. pky vtm (Packyear $>$ 30 and not taking vitamins)
12. sex inc jomyop asa (Male, low income, juvenile myopia, not taking aspirin)
13. sex inc catct asa (Male, low income, cataract, not taking aspirin)

Having done some ”data mining”, the investigators can go back and look at classes of people who may not have been examined separately before.

catct	pky	not take vitamins	risk of progression
1	1	1	17/23 = 0.7391
1	1	0	7/14 = 0.5000
0	1	1	22/137 = 0.1606
0	1	0	2/49 = 0.0408
1	0	1	18/51 = 0.3529
1	0	0	19/36 = 0.5278
0	0	1	22/363 = 0.0606
0	0	0	13/203 = 0.0640

Looking at the smokers: smokers with cataract are relatively protected by taking vitamins, and smokers without cataract are also relatively protected by taking vitamins. For non smokers taking or not taking vitamins makes no (significant) difference. The fun of discovery.

Physiologically meaningful - recent literature suggests:

a) Certain vitamins are good for eye health.
b) Smoking depletes the serum and tissue vitamin level, especially Vitamin C and Vitamin E.

(Although as usual, a ”randomized controlled clinical trial would provide the best evidence of any effect of vitamins on progression of myopia in smokers”)

Want to be able to solve this problem for $N = 128$ as here, but for N many thousands, as in genetic data, e. g. single nucleotide polymorphisms (SNPs), where $p \sim 9000$, $\binom{9000}{2}$ very large.. Minimize a convex functional subject to a humongous number of linear inequality constraints.

♣♠ The "Regularization Class" of Statistical Models.

$$y \in \mathcal{Y},\ x \in \mathcal{X},\ f \in \mathcal{F}$$

Observe $\{y_i, x(i), i = 1, \cdots, n\}$ where $\mathcal{X}$ can be quite general: Small or large vectors whose components are categories, are in E^d, are on manifolds, are trees, graphs, $\cdots$ etc. $\mathcal{F}$ is a specified class of functions whose domain is $\mathcal{X}$, and y is a random vector whose distribution depends on f, $y_i \sim F_{f(x(i))}$, where $f \in \mathcal{F}$.

The goal is to find $f \in \mathcal{F}$ to mininmize

$$\sum_{i=1}^{n} C(y_i, f(x(i))) + J_\lambda(f)$$

C measures some goodness of fit of f to the observations, while J constrains the complexity of the solution, and λ controls this tradeoff.

Various forms of C include

- Negative log likelihood (for y a member of of an exponential family-Gaussian, Bernoulli, Poisson, Gamma, etc)
- Quantile estimation (estimates quantiles of $\mathcal{F}_{f(x)}$)
- Robust estimation (insensitive to outliers)
- ”Hinge functions” (Support Vector Machines, for classification)
- Multivariate generalizations of all of the above.

A large variety of J_λ are appearing in the literature:

$$J_\lambda(f) = \sum_{\ell=1}^{q} \lambda_\ell J_{\ell,\theta}(f)$$

where the $J_{\ell,\theta}$ are convex functionals, norms, semi-norms in various spaces, depending possibly on nuisance parameters θ. The optimization problem generally has to be solved for many values of $\lambda = (\lambda_1, \cdots, \lambda_q)$ in order to estimate an optimum λ.

With $\mathcal{Y}, \mathcal{X}, \mathcal{F}$ and J_λ becoming more complex, and n very large, there is real need for efficient large scale optimization code.