President's Invited Address

Meeting theme: "Statistics, harnessing information"

Regularization Methods in Statistical Model Building: Statisticians, Computer Scientists, Classification and Machine Learning

Grace Wahba

Joint Statistical Meetings

Salt Lake City, July 30 2007

These slides at

`http://www.stat.wisc.edu/~wahba/` → TALKS

Papers/preprints at

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

Abstract

- We give an overview of a broad class of statistical model building tools-regularization methods-that have come from, and populate both Statistics and Computer Sciences.
- Relations between Bayes estimates and this class is noted.
- Tuning of these models for prediction and for model selection will be noted.
- Interplay between Statisticians and Computer Scientists in extending this rich class of methods is noted.
- We proceed by example.

1. The Regularization Class of Statistical Models.
2. Example: The cubic smoothing spline.
3. Cost functions.
4. Penalty functionals.
5. Relation to Bayes estimates.
6. More on quadratic penalty functionals.
7. Example: Spline ANOVA models:local global warming trends.
8. Classification: The support vector machine (SVM).
9. Example: Classification of satellite radiance profiles.
10. l_1 penalties, the LASSO, Basis Pursuit, LASSO-Patternsearch.
11. Examples: Risk of progression of myopia, classification of rheumatoid arthritis SNP data.
12. Comments and conclusions.

Regularization Class of Statistical Models

- $y \in \mathcal{Y}$: The observations, $y_1, \cdots, y_n$.
- $x \in \mathcal{X}$: The attribute vectors, $x(1), \cdots, x(n)$.
- $f \in \mathcal{H}$: The model, to be found, relates $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. $\mathcal{H}$ is the class of functions in which f is to be found.
- $\mathcal{C}(y, f)$: The cost-measures goodness of fit of the model to the data.
- $J_\lambda(f)$: Penalty functional on f, constrains complexity/degrees of freedom of the model.

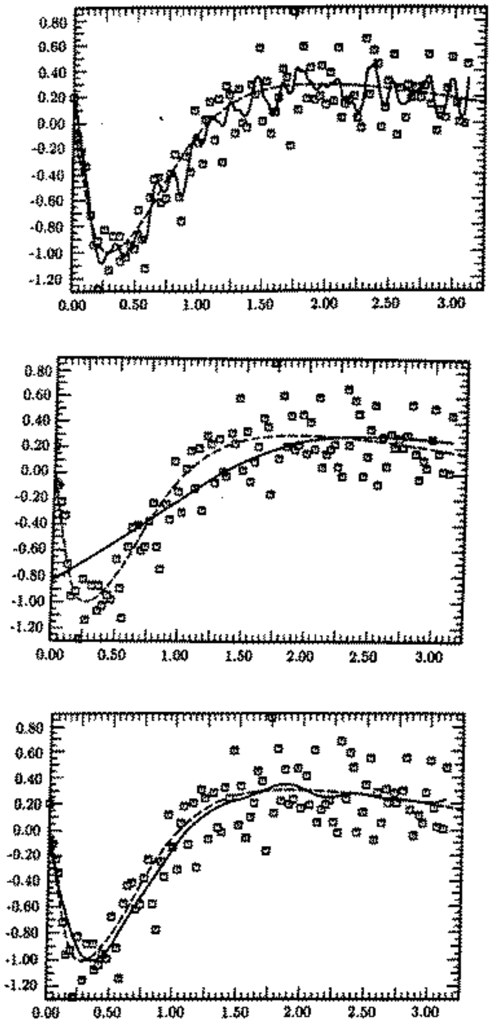
The model f is found as the solution to: min $f \in \mathcal{H}$:

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

The (set of) parameter(s) λ controls the tradeoff between fit and complexity, a. k. a bias-variance in some contexts.

One simple example leads to the cubic smoothing spline.

- y a real number.
- $x \in [0, 1]$.
- $f \in W_2^2$ (Sobolev space of functions with square integrable second derivative).
- $\mathcal{C}(y, f) = (y - f(x))^2$.
- $J_\lambda(f) = \int_0^1 (f''(x))^2 dx$.



Top: λ too small; Middle λ too big; Bottom λ just right, chosen by Generalized Cross Validation *GCV*. Dotted line = "truth". Golub, Heath and Wahba, 1979 SVD, Craven and Wahba, 1979).

Cost Functions (Univariate)	$\mathcal{C}(y, f)$
Regression:	
Gaussian data	$(y - f)^2$
Bernoulli, $f = log[p/(1-p)]$	$-yf + log(1 + e^f)$
Other exponential families	other log likelihoods
Data with outliers	robust functionals
Quantile functionals	$\rho_q(y - f), \rho_q(\tau) = \tau(q - I(\tau \leq 0))$
Classification: $y \in \{-1, 1\}$	
Support vector machines	$(1 - yf)_+$, $(\tau)_+ = \tau, \tau \geq 0, 0$ otherwise
Other "large margin classifiers"	e^{-yf} and other functions of yf

Multivariate (vector-valued y) versions of the above.

Penalty Functionals	$J_\lambda(f)$
Quadratic (RKHS) Penalties: $x \in \mathcal{T}$, some domain, can be very general. $f \in \mathcal{H}_K$, a Reproducing kernel Hilbert Space (RKHS) of functions, characterized by some positive definite function $K(s,t)$, $s,t \in \mathcal{T}$.	$\lambda\|f\|^2_{\mathcal{H}_K}$, etc.
l_p **Penalties:** $x \in \mathcal{T}$, some domain, can be very general. $f \in$ span $\{B_r(x), r = 1, \cdots, N\}$, a specified set of basis functions on $\mathcal{T}$. $f(x) = \sum_{r=1}^N c_r B_r(x)$	$\lambda \sum_{r=1}^N \|c_r\|^p$

$\lambda \rightarrow (\lambda_1, \cdots, \lambda_q)$ Combinations of RKHS and l_p penalties.

Bayes Estimates: Let $\mathcal{C}$ be a log likelihood.

- RKHS penalties: $\lambda\|f\|^2_{\mathcal{H}_K}$. Let f_λ be the minimizer of

  $$\sum_{i=1}^{n} \mathcal{C}(y_i, f(x(i))) + \lambda\|f\|^2_{\mathcal{H}_K}.$$

  f_λ is a Bayes estimate for the zero-mean Gaussian prior with covariance some multiple of $K(s,t), s,t \in \mathcal{T}$, controlled by λ.

- l_1 penalties: $\lambda\sum_{r=1}^N |c_r|$. f_λ is a Bayes estimate for some multiple of independent prior double negative exponential distributions on the c_r, controlled by λ.

- Remark: (Low rank) improper priors are allowed in RKHS penalties and there are then no penalties on the components - in the cubic smoothing spline example the estimate shrinks to linear in the large λ case. (Similar for double neg. exponential.)

Some Tuning References

Not complete. May not be the earliest reference. Not guaranteed.

- Unbiased Risk C. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.
- AIC H. Akaike. A new look at the statistical identification model. *IEEE Trans. Auto. Control*, 19:716–723, 1974.
- Leaving-out-one G. Wahba and S. Wold. A completely automatic French curve. *Commun. Stat.*, 4:1–17, 1975.
- GCV-illposed G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14:651–667, 1977.
- Unbiased Risk M. Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.*, 6:473–484, 1978.
- BIC G. Schwartz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.
- GCV G. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–224, 1979.
- GCV P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.
- GML G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, 13:1378–1402, 1985.
- Randomized Trace D. Girard. A fast ‘Monte-Carlo cross-validation’ procedure for large least squares problems with noisy data. *Numer. Math.*, 56:1–23, 1989.
- Randomized Trace M. Hutchinson. A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun. Statist.-Simula.*, 18:1059–1076, 1989.
- GACV D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6:675–692, 1996.

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

- SVM T. Joachims. Estimating the generalization performance of an SVM efficiently. In *Proceedings of the International Conference on Machine Learning*, San Francisco, 2000. Morgan Kaufman.

- GACV-SVM G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–311. MIT Press, 2000.

- GACV-clustered outcomes F. Gao, G. Wahba, R. Klein, and B. Klein. Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion. *J. Amer. Statist. Assoc.*, 96:127–160, 2001.

- SVM O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002.

- GACV-l_1 H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein. Variable selection and model building via likelihood basis pursuit. *J. Amer. Statist. Assoc.*, 99:659–672, 2004.

- GACV-multicat-SVM Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *J. Amer. Statist. Assoc.*, 99:67–81, 2004.

- B.Efron. The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 81:619–642. (with discussion), 2005.

- GACV-l_1,BGACV 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.

- M. Yuan. GACV for quantile smoothing splines. *Comp. Stat. Data Anal.*, 50:813–829, 2006.

More on Quadratic (RKHS) Penalties

Are defined by a positive definite function $K(s,t), s,t \in \mathcal{T}$. The domain $\mathcal{T}$ can be any domain on which you can define a positive definite function:

$$\begin{aligned}
\mathcal{T} &= (\ldots, -1, 0, 1, \ldots) \\
\mathcal{T} &= [0,1] \\
\mathcal{T} &= E^d \qquad\qquad \text{(Euclidean } d\text{-space)} \\
\mathcal{T} &= \mathcal{S} \qquad\qquad \text{(the unit sphere)} \\
\mathcal{T} &= \text{ the atmosphere} \\
\mathcal{T} &= \{\Diamond, \triangle, \heartsuit\} \quad \text{(unordered set)}
\end{aligned}$$

More...

More domains:

$$\begin{aligned}
\mathcal{T} &= \text{A Riemannian manifold} \\
\mathcal{T} &= \text{A collection of trees} \\
\mathcal{T} &= \text{A collection of graphs} \\
\mathcal{T} &= \text{A collection of proteins} \\
\mathcal{T} &= \text{A collection of gene microarray chips}
\end{aligned}$$

Vector sums and products of domains are allowed.

Quadratic (RKHS) penalties:

- To every positive definite function $K(s,t), s,t \in \mathcal{T}$ there corresponds a unique RKHS and vice versa.
- Consider the optimization problem: find $f \in \mathcal{H}_K$ to minimize

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|f\|^2_{\mathcal{H}_K}.$$

The representer theorem says, under mild conditions on a convex $\mathcal{C}$, the minimizer has the form

$$f_\lambda(t) = \sum_{i=1}^{n} c_i K(t_i, t)$$

for some $c = (c_1, \cdots, c_n)$.

- Furthermore $\|f\|^2_{\mathcal{H}_K}$ is then given by $\sum_{i,j} c_i c_j K(t_i, t_j)$ so that f may be computed by solving a convex optimization problem with a finite number of unknowns.

Tensor sums and products of positive definite functions are positive definite. Spline ANOVA models:

A Time and Space Model for Global Warming.

$t = (x, P)$, $x = 1, \cdots 30$, (year) $P =$ space =(latitude, longitude).

$$\mathcal{H} = \underbrace{\left[[1^{(1)}] \oplus [\phi] \oplus \mathcal{H}_s^{(1)} \right]}_{time} \quad \otimes \quad \underbrace{\left[[1^{(2)}] \oplus \mathcal{H}_s^{(2)} \right]}_{space}$$

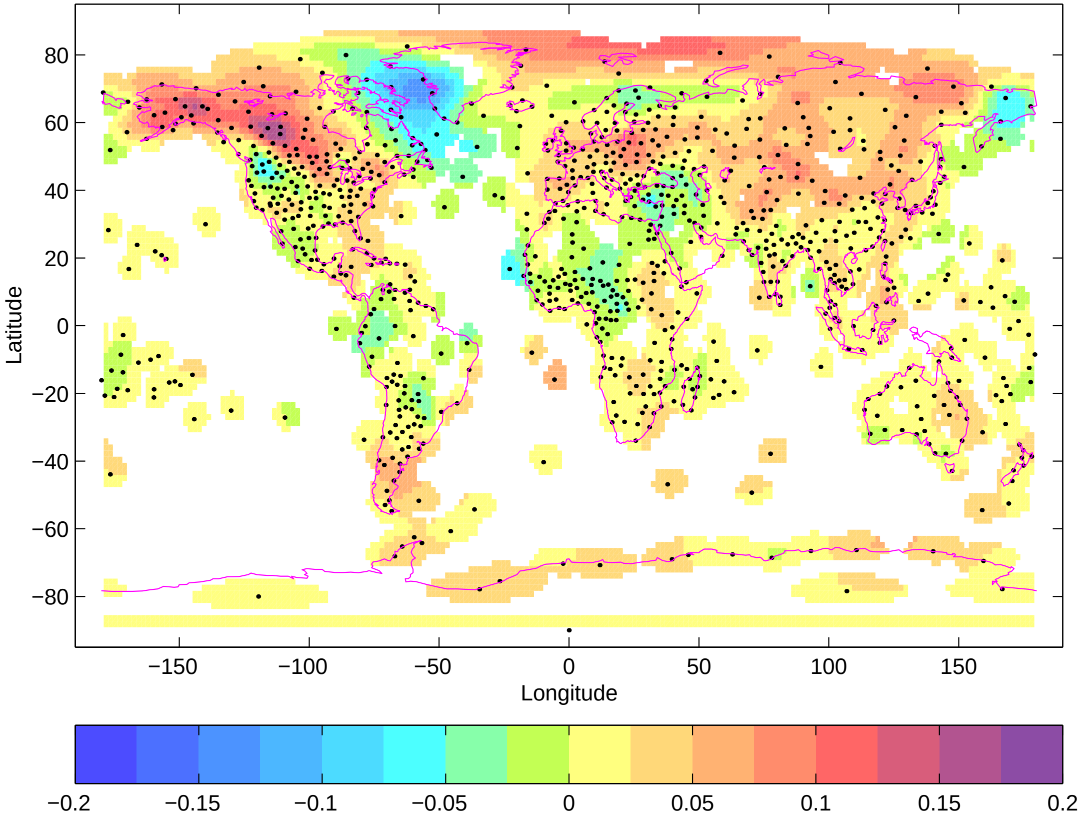
ϕ is linear in time orthogonal to the one dimensional constant function $[1^{(1)}]$.

Expands out to 6 terms - next slide.

The trend by space term below will show how the warming trend varies with latitude and longitude

$$\begin{array}{ccccccccc}
\mathcal{H} & = & [1] & \oplus & [\phi] & \oplus & [\mathcal{H}_s^{(1)}] & \oplus & [\mathcal{H}_s^{(2)}] \\
f(x,P) & = & \mu & + & d\phi(x) & + & f_1(x) & + & f_2(P) \\
 & = & \mathit{mean} & + & \mathit{global} & + & \mathit{time} & + & \mathit{space} \\
 & & & & \mathit{time} & & \mathit{main} & & \mathit{main} \\
 & & & & \mathit{trend} & & \mathit{effect} & & \mathit{effect}
\end{array}$$

$$\begin{array}{ccccc}
\oplus & [[\phi] \otimes \mathcal{H}_s^{(2)}] & \oplus & [\mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)}] \\
+ & \phi(x) f_{\phi,2}(P) & + & f_{12}(x,P) \\
+ & \mathit{trend} & + & \mathit{space}- \\
 & \mathit{by\ space} & & \mathit{time} \\
 & \mathit{effect} & & \mathit{interaction}
\end{array}$$



Trend by Space Effect-Global Warming

Average Nov. Dec. Jan. surface temperature, 1961-1990. Local trend as a function of latitude and longitude. Note warmer swath from the Upper Midwest to Alaska, noticeably affecting the X-C ski season in the Upper Midwest. 1000 stations (dots) for 30 years with missing data. Solve a large linear system.

Luo, Wahba and Johnson, Spatial-temporal analysis of temperature using smoothing spline ANOVA, *J. Climate* **11**, 1998, Chiang et. al. 1999

Classification: The Support Vector Machine (SVM)

The Support Vector Machine for classification is very popular among Computer Scientists, for good reason, to be explained.

The simplest case is the "Standard" two class situation $y = \pm 1$, a "representative" training set is available, and costs for the two types of misclassifications are the same. Given $\{y_i, t_i\}, i = 1, \cdots, n$, find $f \in \mathcal{H}_K$ to min

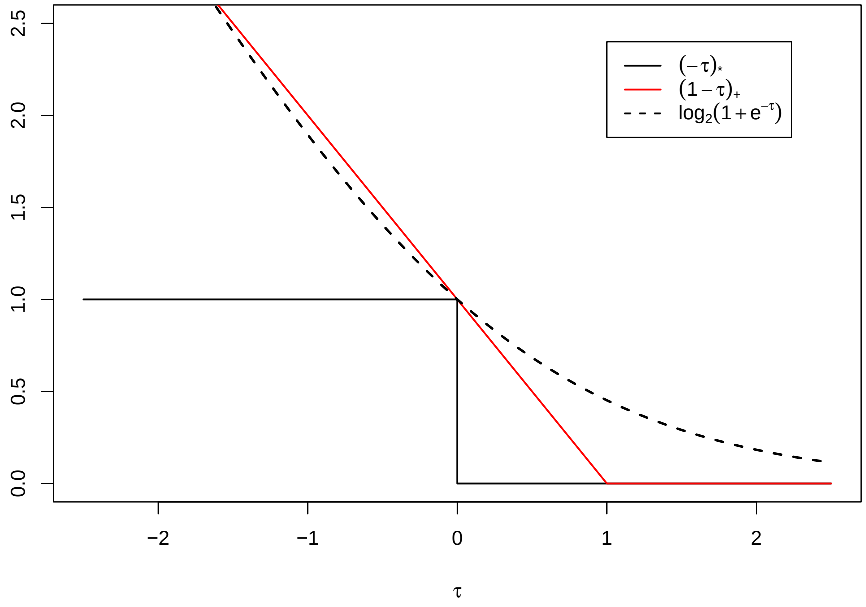
$$\sum_{i=1}^{n} (1 - y_i f(t_i))_+ + \lambda \|f\|^2_{\mathcal{H}_K},$$

where $(\tau)_+ = \tau$ if $\tau > 0$ and 0 otherwise.

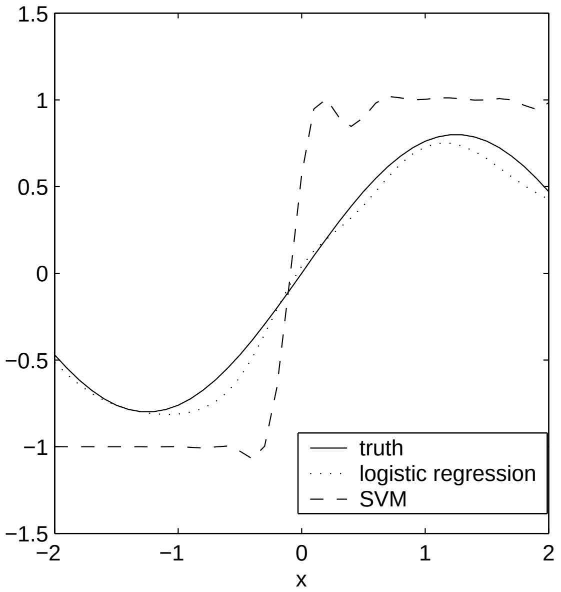
To classifiy a future object with attribute t_*: the $+$ class if $f_\lambda(t_*) > 0$ and the $-$class if $f_\lambda(t_*) < 0$. Minimize a quadratic functional subject to a family of linear inequality constraints. Solutions tend to be sparse.

Classification: The SVM (continued).

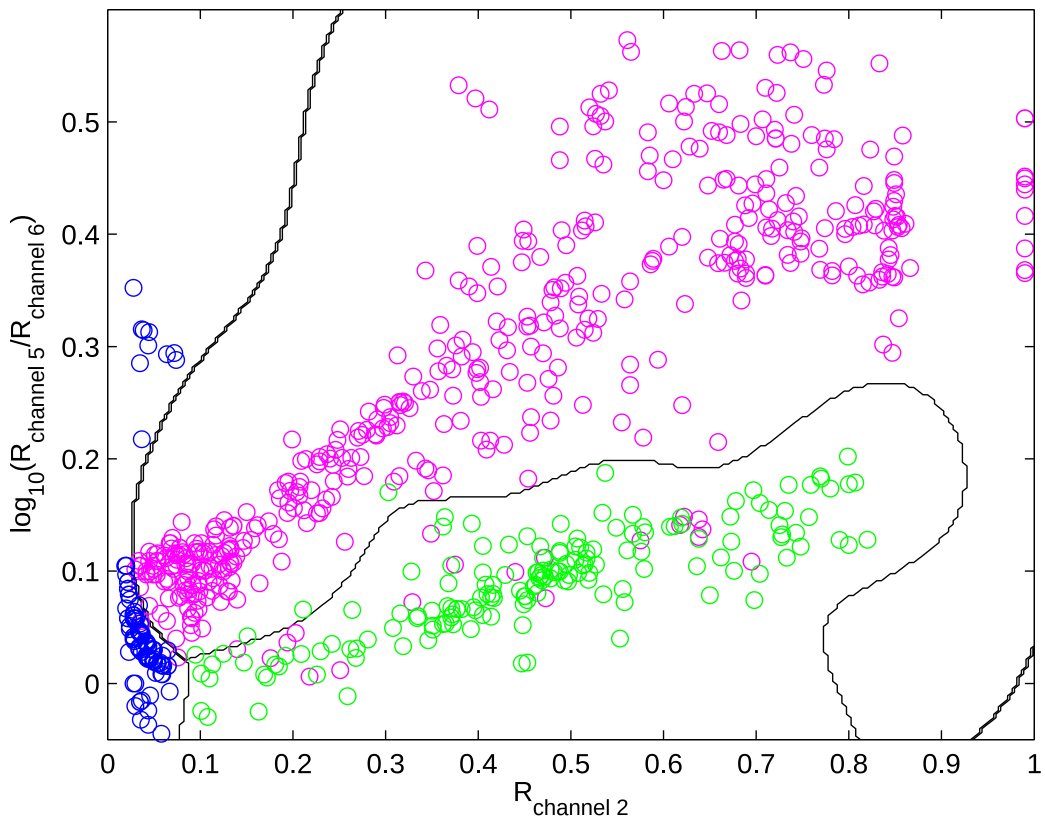
The SVM may be compared with the penalized likelihood estimate, which estimates the probability that an object is in the $+1$ class, by recoding Bernoulli data ($y \in \{0,1\}$) as $y \in \{-1,+1\}$. With this recoded data the Bernoulli log likelihood $-yf + log(1+e^f)$ becomes $log(1+e^{-yf})$. For classification the log likelihood is replaced by the so-called hinge function $(1-yf)_+$. Observe that now both the log likelihood and the hinge function depend on yf, the so-called "margin".



Let $\mathcal{C}(y, f) = c(yf)) = c(\tau), \tau = yf$. Comparison of the misclassification function $c(\tau) = (-\tau)_* = 1$ if $\tau \leq 0$, and 0 otherwise, the hinge function $(1 - \tau)_+$ and the log likelihood function $log_2(1 + e^{-\tau})$. Any strictly convex function that goes through 1 at $\tau = 0$ will be an upper bound on the misclassification function $(-\tau)_*$ and will be a looser bound than some SVM (hinge) function $(1 - \theta\tau)_+$.



Penalized log likelihood estimates the log odds ratio $f(x) = log[p(x)/(1 - p(x)]$, while the SVM is estimating the sign of the log odds ratio, just what you need for classification. (Yi Lin 2002). Vertical scale is $2p(x) - 1$ for "truth" and logistic regression and $f_\lambda(x)$ for the (tuned) SVM. Data were 300 equally spaced Bernoulli observations from "truth".



The (tuned) multicategory SVM of Lee, Lin and Wahba (2004). Data are functions of radiance observation vectors from the MODIS satellite and the categories are clear sky, water clouds, and ice clouds. Note a few outliers-probably human error.

Some comments on the SVM and penalized likelihood:

- If you want the *probability* that an object is in class 1, then you want to use the penalized log likelihood. The SVM does not estimate a probability, despite many erroneous remarks to this effect in the literature.

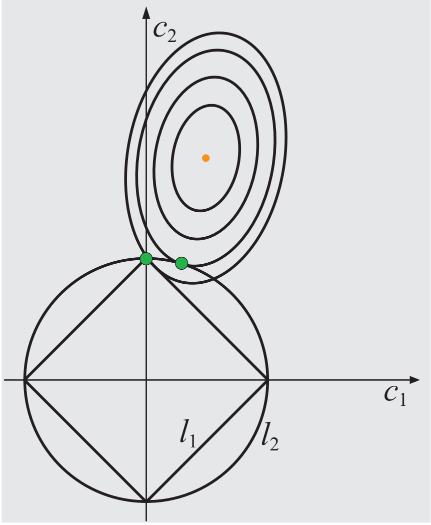
- However, when the classes are (nearly) separable, or you are in high dimensions, then the penalized likelihood estimate will cause computational trouble, since, as $p(x)$ tends to 0 or 1, $f(x)$ tends to $\pm\infty$.

- Various flavors of the SVM have become highly popular in the applied machine learning literature they work very well in high dimensions and/or, when classes are (nearly) separable.

- Many in-sample competing tuning methods can be found in the literature. When n large, a withheld tuning set is popular. Much activity-graphs, trees, heterogenous data sources ...

l_1 penalties, the LASSO, Basis Pursuit (1994).

l_1 penalties give sparse solutions. Note the ellipse meets the diamond at a vertex- where $c_1 = 0$.



$f(x) = \sum_{r=1}^{N} c_r B_r(x)$, where the basis functions, $\{B_r\}$ are chosen with respect to the practical problem at hand.

Given $\{y_i, x(i), i = 1, \cdots n\}$, find $c = (c_1, \cdots c_N)$ to min

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(x(i))) + \lambda \sum_{r=1}^{N} |c_r|$$

It is well know that the l_1 penalty tends to give sparse solutions, that is, many of the $c_r = 0$, unlike l_p penalties for $p > 1$.

Thus the LASSO is popular when it is appropriate to start with a large number of basis functions (when $N >> n$, known as basis pursuit) in the expectation that only relatively few of the c_r are non-zero. The number of non-zero coefficients is controlled by the choice of λ. Many variations have been proposed including multiple tuning parameters, weighted and grouped coefficients. A growing literature exists on the theoretical properties of this class when $\mathcal{C}$ is squared error, but no doubt extends to the general case.

There are a number of interesting practical issues, not completely solved. They include:

- Choice of λ. Before thinking about this it is good to consider what the objective of the analysis is. Assuming that the “true” model (or the nearest model of the assumed form to some “true” model) is sparse, it is good to recognize that choosing λ for best prediction is not necessarily the same as choosing λ to extract just the right non-zero coefficients. (One way to think about this is to contemplate the difference between AIC (prediction) and BIC (variable selection)).

- Numerical methods for minimizing a convex functional with possibly a very large number of linear inequality constraints, or other formulation of this optimization problem. Global vs. sequential methods for choosing the non-zero coefficients in the LASSO very large N context.

The LASSO-Patternsearch Algorithm

We will describe how the two issues:

- Prediction vs. Model Selection, and
- Large scale computation

were handled via two recent examples with Bernoulli response data, $\mathcal{C}(y, f) = -yf + log(1 + e^f)$,

Example 1. Risk of progression of myopia in a demographic study.
Example 2. Classification of rheumatoid arthritis SNP data in a case-control study.

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.

There are several assumptions behind the algorithm we will call "LASSO-Patternsearch".

- $y \in \{0,1\}, x$ a p-vector of zeroes and ones, possibly long.
- For all or nearly all of the components of x the possibly "risky" direction is known *a priori*, and is coded as 1.
- It is desired to see if high order interactions/synergy between the components of x are present.

Example 1: Risk of progression of myopia in a demographic study.

Applied to to "progression of myopia" from the Beaver Dam Eye Study, BDES 1 to BDES 2, five years apart. $n = 876$ records of persons aged 60-69 at BDES 1. 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? Consider seven variables of possible interest and want to see if there are high order interactions among the variables. The continuous variables are dichotomized so as to be able to do this.

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.

For the LASSO-Patternsearch the basis functions will be all products of the components of $x = (x_1, x_2, \cdots, x_p)$ 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), $N = 2^p$. For the myopia data there are $2^7 = 128$ coefficients, or, not counting the constant, 127 possible “patterns”.

Note that the conditional distribution of one Bernoulli random variable y given p other Bernoulli random variables $x_1, \cdots, x_p$ has 2^p paramteters and can be expanded in complete generality in these basis functions. The representation will be most compact, however, if all the risky variables are coded with the risky direction as 1.

A special purpose algorithm which can handle N up to 4000 on our 3.4 GHz cpu and 4Gb memory workstation has been designed by our CS collaborator Stephen Wright working with Weiliang Shi is in Shi *et. al.*, see also Steve Wright's June 07 talk "Solving l_1 Regularized Regression Problems" in his talks directory at `http://pages.cs.wisc.edu/~swright/`

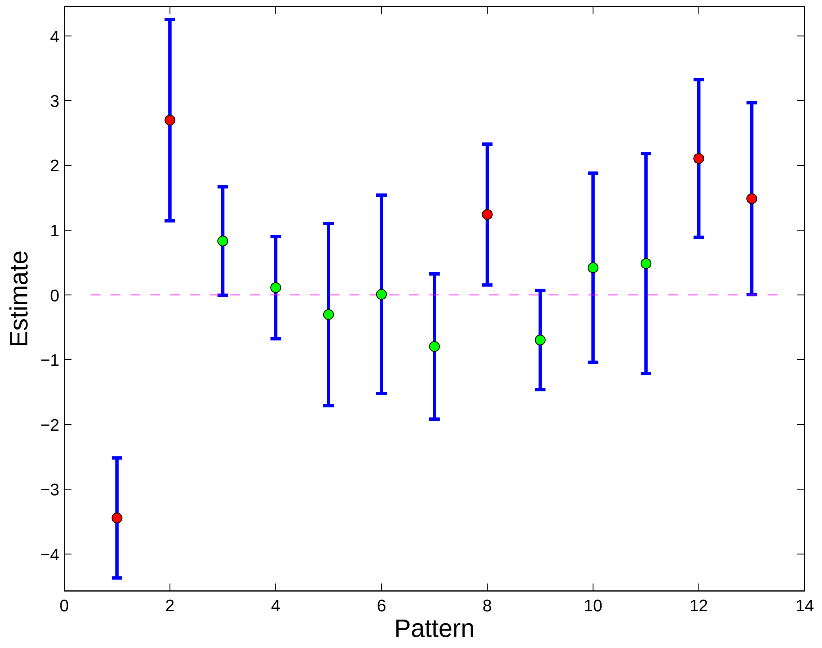
The LASSO-Patternsearch has the following steps:

Step 1. Minimize $\sum_{i=1}^{n} \mathcal{C}(y_i, f(x(i))) + \lambda \sum_{\ell=1}^{N} |c_\ell|$, choose λ by $GACV$-what is GACV?

Step 2. Enter all basis functions with $\ell : |c_\ell| > 0$ into a parametric logistic regression model:

$$f(x) = \sum_{\ell : c_\ell > 0} a_\ell B_\ell(x)$$

and fit. Twelve patterns passed Step 1. Four of them are significant at significance level q =96.92%. Choose q by $BGACV$-what is BGACV?



Step 2. continued. Estimates and $q = 96.92\%$ confidence intervals in parametric logistic regression for the twelve patterns (plus constant) that passed Step 1, $|c_r| > 0$. Red dots at the estimates indicate four significant coeficients.

- What is GACV? What is BGACV? Begin with ordinary leaving-out-one cross validation $CV(\lambda)$:

$$CV(\lambda) = \sum_{k=1}^{n} \mathcal{C}(y_k, f_\lambda^{[k]}(x(k)))$$

where $f^{[k]}$ is the estimate with the kth data point left out.

- Make several approximations. Perform some averaging.
- End result is: GACV: find λ to min

$$GACV(\lambda) = \sum_{i=1}^{n} \mathcal{C}(y_i, f_\lambda(x(i))) + trH(\lambda) \sum_{i=1}^{n} y_i(y_i - p_\lambda(x(i)))/(n - N_B),$$

where $H(\lambda)$ is the inverse Hessian of the variational problem, p_λ is the estimated probability of a 1, and N_B is the number of basis functions.

- BGACV: find λ to min

$$BGACV(\lambda) = \sum_{i=1}^{n} \mathcal{C}(y_i, f_\lambda(x(i)) + log\ n/2 tr H(\lambda) \sum_{i=1}^{n} y_i(y_i - p_\lambda(x(i))/(n - N_B).$$

Minimizer of $GACV$ is a good estimator of the minimum of $E_{f_{true}} \sum_{i=1}^{n} \mathcal{C}(y_{new}, f_{true}(x(i))$,(1995). Many simulation results, could use some theory. $BGACV$ - adhoc insertion of $log\ n/2$ for model selection, motivated by the difference between AIC and BIC. Needs theory. Many works on model selection. Here the 12 patterns passing Step 1. will be lined up in significance order and $BGACV$ minimized to get q=96.92%.

Step 3. Select all ℓ for which a_ℓ are significant at the $q\% = 96.92\%(BGACV)$ level, to fit the final model. The patterns passing this test are:

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)

Step 3.(continued) Fit the final model with the five patterns significant at the 96.92% (BGACV) level.

$$f(x) = \sum_{\ell : a_\ell \quad significant} b_\ell B_\ell(x).$$

The (refitted) model is

$$f(catct, pky, vtm, sex, inc, jomyop, asa)$$

$$-3.29 + 2.42 * cact + 1.18 * pky * vtm$$

$$+ 1.84 * sex * inc * jomyop * asa + 1.08 * sex * inc * cat * asa.$$

Step 4. Having done some “data mining”, the investigators can go back and look at classes of people who may not have been examined separately before. For example:

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: $(1, 1, 1, 1)$:

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.

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")

Message: Aside from theoretical issues, there are practical questions of scientific import in dealing with real problems.

To check on the "significance" of the patterns, randomly scramble the ys while keeping the x's fixed, and apply the entire LASSO-Patternsearch algorithm to see how often false patterns are generated. Repeat 600 times. (Statistical theory is not clear on properties of multistep procedures)

Detection of noise patterns found in scrambled data compared to observed p values:

Log p values of the patterns found (out of 600) are plotted (l. to r. top to bottom) for observed patterns of size 1,2,3,4. Red lines are for the observed p-values for *catct*, *pky vtm*, none, and *sex inc jomyop asa* (lower) and *sex inc catct asa* (upper).

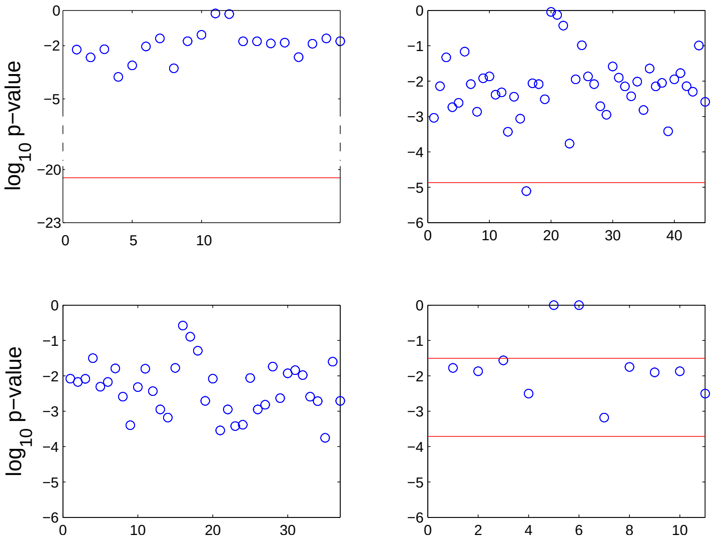


Figure 1: Upper red line suggests that *sex inc catct asa* is borderline significant.

Genetic Data (realistic simulation from the Genetic Analysis Workshop 15, 2006).

y = phenotype Rheumatoid Arthritis or not.
x = SNPs, alleles, covariates, $p = 9192$ components.

Train: 1500 cases, 2000 controls
Tune: 1500 cases, 2000 controls
Test: 1500 cases, 2000 controls.

Pre-screen step: 9192 variables reduced to $N = 2559$ basis functions for the LASSO step. Final model has 8 main effects and 3 interactions. Using $p = .5$ as a classifier, a competitive 12.6% error rate was obtained. Identified a SNP near most of the genes that were used to generate the data.

W. Shi, K. Lee, and G. Wahba. Detecting disease causing genes by LASSO-Pattern search algorithm. TR 1140, Department of Statistics, University of Wisconsin, Madison WI, 2007.

Message:

Regarding the special purpose algorithm we could do the LASSO step with 2559 basis functions from the prescreen simultaneously. There are many algorithms for doing this or similar problems but most or all are greedy/sequential (at least up to 2006). We think a global method here has advantages over a sequential method when searching over a very large number of basis functions, and/or for when high order interactions are present, assuming that the coding is reasonably correct. Some preliminary simulations suggest that the approach is relatively advantageous when extraneous variables are correlated with relevant variables. All this rasises practical and theoretical questions regarding numerical methods for solving very to extremely large optimization problems regarding global vs. sequential methods.

In Conclusion:

- We looked at the cubic smoothing spline, a spline ANOVA model, the support vector machine and the LASSO-Patternsearch algorithm. RKHS and l_p penalties were noted.
- The regularization class of statistical model-building methods can be used in an extremely wide variety of contexts.
- Tuning for prediction and for model selection do not necessarily lead to the same results- choice of optimality criteria can be an important issue
- Real applications require some understanding of the science behind the data. By the nature of this class of methods this understanding helps suggest reasonable penalties, basis functions, optimality criteria.
- Computer scientists have developed and continue to develop

computational tools which are crucially important to statisticians.

- Theoretical and applied results are available for the practical data analyst as well as the computer scientist, but many challenging problems remain, along with opportunities for novel applications, especially with complex data structures and very large attribute vectors.