Statistics 860: Estimation of Functions From Data (a.k.a Statistical Machine Learning)

Grace Wahba wahba@stat.wisc.edu
http://www.stat.wisc.edu/~wahba

Course directory: http://www.stat.wisc.edu/~wahba/stat860

This course is about statistical model building and supervised machine learning, based primarily on the regularization class of statistical models. We mainly consider reproducing kernel Hilbert space methods (a.k.a. "kernel methods"), but also look at the LASSO (l_1) class of methods. We will be concerned with model tuning, primarily based on crossvalidation based methods; variable selection methods, and methods for combining various kinds of information.

Regularization Class of Statistical Models

- $y \in \mathcal{Y}$: The observations, $y_1, \cdots, y_n$.
- $x \in \mathcal{X}$: The attribute vectors, $x(1), \cdots, x(n)$. (sometimes $t \in \mathcal{T}$.
- $f \in \mathcal{H}$: The model, 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 f
- $J_\lambda(f)$: Penalty functional on f, constrains complexity/degrees of freedom of the model.

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

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

λ controls the tradeoff between fit and complexity.

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).
- $\mathcal{C}(y, f) = (y - f(x))^2$.
- $J_\lambda(f) = \int_0^1 (f''(x))^2 dx$.

Find $f \in W_2^m[0,1]$ to minimize

$$\frac{1}{n}\sum \big(y_i - f(t(i))\big)^2 + \lambda \int_0^1 (f^{(m)}(t))^2 dt.$$

λ is known as the smoothing parameter or, alternately in more complex models as a tuning parameter.

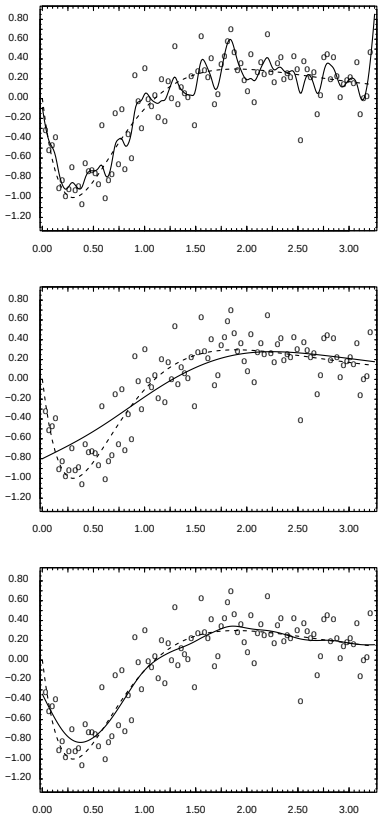


Figure 1: Cubic smoothing spline ($m = 2$) with three different choices of smoothing parameter. λ too small, λ too large, λ just right (chosen by GCV)

.

The domain $\mathcal{X}$ may be almost anything. When we mean a general domain, we may switch from $x \in \mathcal{X}$ to $t \in \mathcal{T}$.

$$
\begin{array}{lrll}
f(t), t \in \mathcal{T} & \mathcal{T} & = & (1, 2, \ldots, N) \\
 & \mathcal{T} & = & (\ldots, -1, 0, 1, \ldots) \\
 & \mathcal{T} & = & [0, 1] \\
 & \mathcal{T} & = & E^d \qquad \text{(Euclidean } d\text{-space)} \\
 & \mathcal{T} & = & \mathcal{S} \qquad \text{(the unit sphere)} \\
 & \mathcal{T} & = & \mathcal{S} \otimes [0, 1] \qquad \text{(the atmosphere)} \\
 & \mathcal{T} & = & \mathcal{S} \otimes [0, 1], \mathcal{S} \otimes [0, 1], \cdots, \mathcal{S} \otimes [0, 1] \\
 & \mathcal{T} & = & \text{vector of SNPs} \\
 & \ldots & &
\end{array}
$$

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	$\rho_q(y - f)$
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	$J_\lambda(y, f)$
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 $\mathcal{T}$.	
$f(x) = \sum_{r=1}^N c_r B_r(x)$	$\lambda \sum_{r=1}^N \|c_r\|^p$
$\lambda \to (\lambda_1, \cdots, \lambda_q)$, Combinations of RKHS and l_p penalties.	

Types of information:

1. Noisy values of point functionals
   $y_i = f(t_i) + \epsilon_i, i = 1, \cdots, n, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

2. Values of derivatives, integrals, and other bounded linear functionals
   $y_i = f^{(j)}(t_i) + \epsilon_i$
   $y_i = \int K(t_i, f(s)ds) + \epsilon_i$
   $y_i = L_i f + \epsilon_i$

3. Penalized GLIM
   $y_i \sim F_{f(t_i)}, \sim F_{L_i f}$
   Ex: $F_{f(t_i)}$ Poisson with mean $\Lambda_i = f(t_i)$
   Ex: $F_{f(t_i)}$ Bernoulli $B(1, p_i)$,
   with $f(t_i) = log[p_i/(1 - p_i)]$ = log odds ratio a.k.a logit

Bernoulli data: The goal is to estimate $f(t)$, and then recover $p(t) =$ probability of outcome 1, given t.

4. Categorical information
   $\{y_i, t_i\}, y_i = 1, 2, ..., k$ classes.
   The goal is to build a classification model.

5. Noisy dissimilarity information
   $d_{ij} = dissim($Object i, Object $j)$
   From this information, embed the (training) objects in a Euclidean space, then build models using the Euclidean coordinates as attributes. Classification of protein sequences, incorporation of pedigree information in models.

6. Examining independence of $x_A \in \mathcal{X}_A = E^p$ and $x_B \in \mathcal{X}_B = E^q$ nonparametrically using only pairwise distances (Distance Correlation)

   Distance Correlation is based only on $|x_A(i) - x_A(j)|$ and $|x_B(i) - |x_B(j)|, i, j = 1 \cdots n$ allowing independence tests between variables for which only pairwise distances are known.

The course has five themes

1. Unified approach to the estimation of functions and the building of statistical/learning models, given various kinds of observations, via regularization methods. Tuning and variable selection included.
2. Theoretical properties.
3. Numerical methods for implementation.
4. Use of software (a) to study properties of the method by Monte Carlo methods and (b) to analyze data.
5. Applications in risk factor estimation, machine learning and classification, ill posed inverse problems, protein sequence analysis, extremely large SNP sequence analysis, meteorological data analysis, others.. .

Examples:

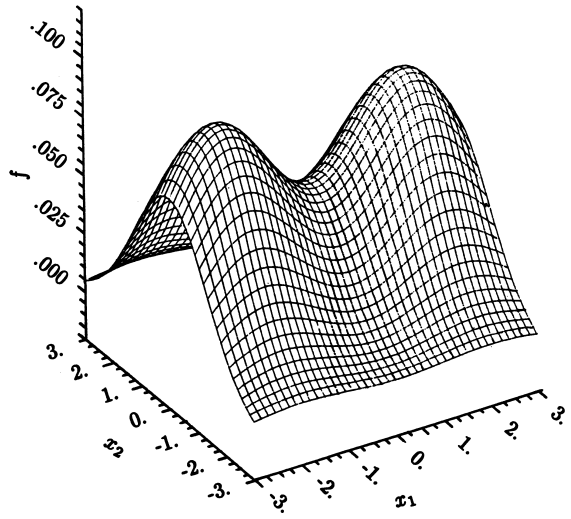
$$\mathcal{T} = E^2 \qquad t = (x_1, x_2)$$

$$y_i = f(x_1(i), x_2(i)) + \epsilon_i$$

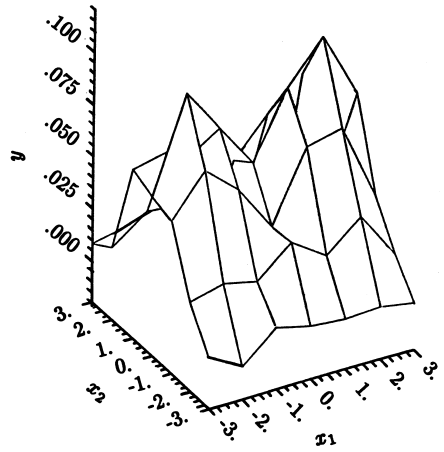
Find $f \in \mathcal{X}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\big(f(x_1(i), x_2(i)) - y_i\big)^2 + \lambda \iint_{-\infty}^{\infty} f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2 dx_1 dx_2$$

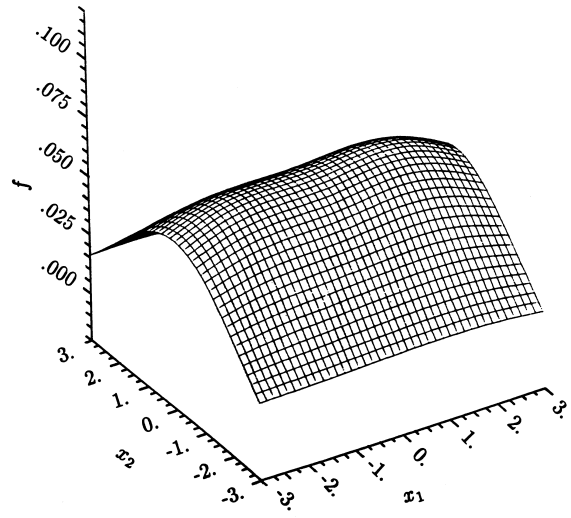
Leads to thin plate splines of dimension 2 and order 2.



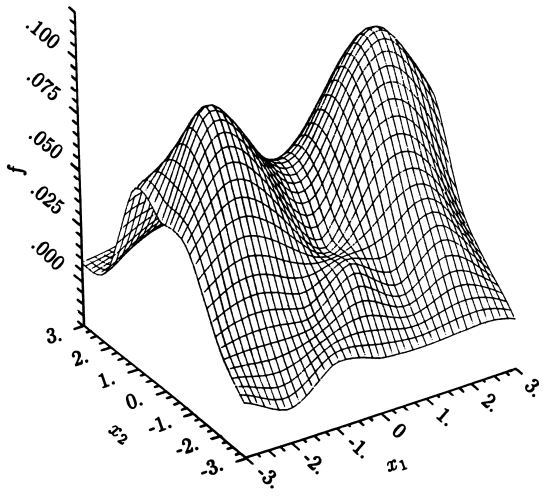
The actual surface.



The data.

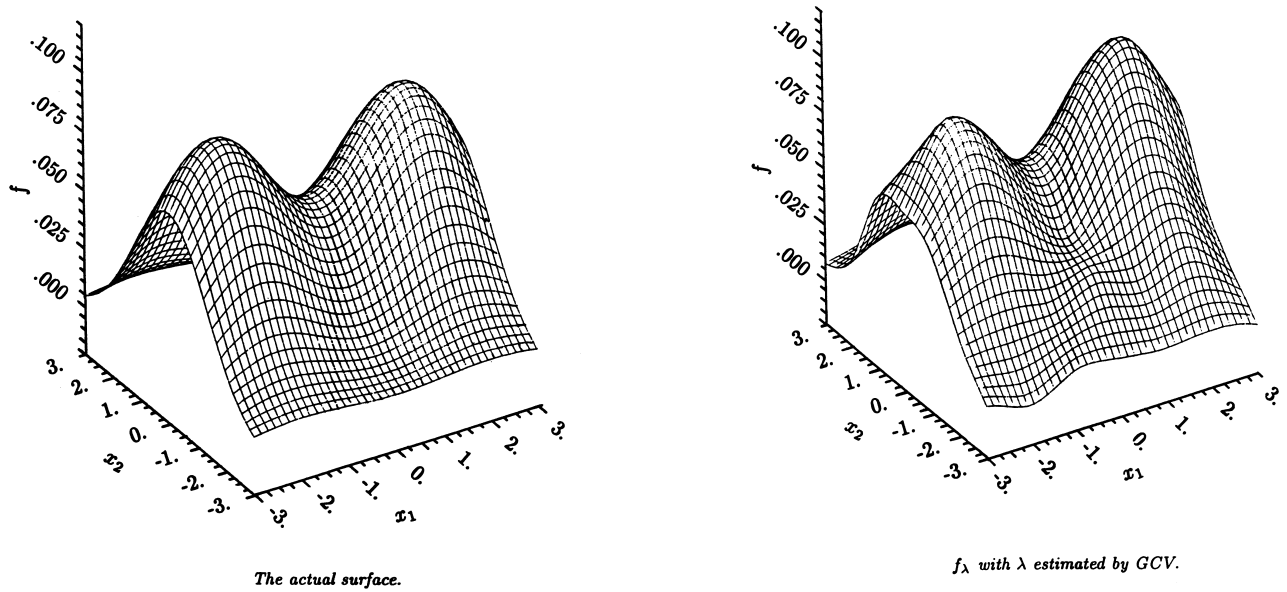


f_λ with λ too large, $\lambda = 100\hat{\lambda}$.



f_λ with λ too small, $\lambda = .01\hat{\lambda}$.

Left to right and top to bottom: The actual surface, the data, λ too big, λ to small.



The actual surface.

f_λ with λ estimated by GCV.

Left to right: The actual surface (again), the estimate with λ chosen by GCV.

$\mathcal{T} = S$ sphere (500 mb height, head)

$t = (\lambda, \phi)$ lat. long.

$$\frac{1}{n}\sum_{i=1}^{n}\big(f(\lambda(i),\phi(i)) - y_i\big)^2 + \lambda\int_S (\Delta^m f)^2 dP$$

$$\Delta f = \frac{1}{\sin^2\lambda} f_{\phi\phi} + \frac{1}{\sin\lambda}(\sin\lambda\ f_\lambda)_\lambda$$

when $\mathcal{T} = E^2,\ \Delta f = f_{x_1x_1} + f_{x_2x_2}$

Integral Equations

$$y_i = \int K(t(i), s, f(s))ds + \epsilon_i$$

(possibly nonlinear)

Satellite Tomography; K involves the radiative transfer equation and the ith filter bandpass

Partial Spline Models (a.k.a Mixed Effect Models)

$$y_i = \underbrace{\sum \theta_j \phi_j(t(i), z_i)}_{\text{parametric}} + \underbrace{f(t(i))}_{\text{smooth}} + \epsilon_i$$

semi-parametric

(i) Example: Electricity demand—smooth in temperature, linear in price, income, etc

(ii) To model jumps and breaks in images (intervention splines)

ANOVA In Function Spaces

model response as a function of several variables:

$$\underbrace{y_i}_{\substack{\text{degree of}\\ \text{retinopathy}}} = f(\underbrace{x_1(i)}_{\text{dur}}, \underbrace{x_2(i)}_{\text{gly}}, \underbrace{x_3(i)}_{\text{bmi}}) + \epsilon_i$$

ANOVA functional decomposition

$$f(t) = \mu + \sum_{\alpha} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \dots$$

$$\underbrace{y_i}_{\substack{\text{lake}\\ \text{acidity}}} = f(\underbrace{x_1(i)}_{\substack{\text{calcium}\\ \text{content}}}, \underbrace{x_2(i)}_{\substack{\text{lat.}\\ \text{long.}}})$$

$$f(calcium, lat.-long.) = f_1(calcium) + f_2(lat.-long.)$$

Global Historical Climate data

Monthly mean temperatures from ~ 1700 stations

winter–Dec Jan Feb 1961–1990

$x =$ year, $P =$ lat.-long.

$$f(x,P) = \mu + d\phi(x) + g_1(x) + g_2(P) + g_{\phi,2}(P)\phi(x) + g_{12}(x,P)$$

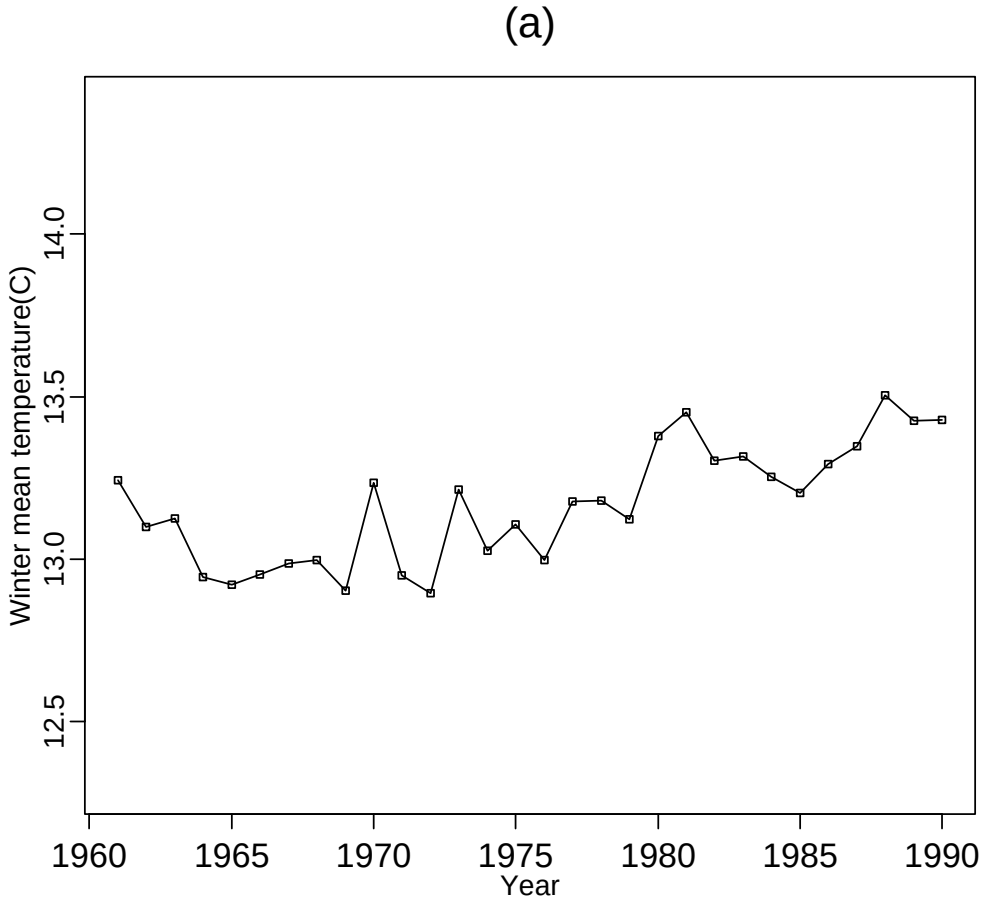
$$y_i = f(x_i, P(i)) + \epsilon_i$$

$$\phi(x) = x - \frac{1}{2} \quad \text{time trend scaled to } [0,1]$$

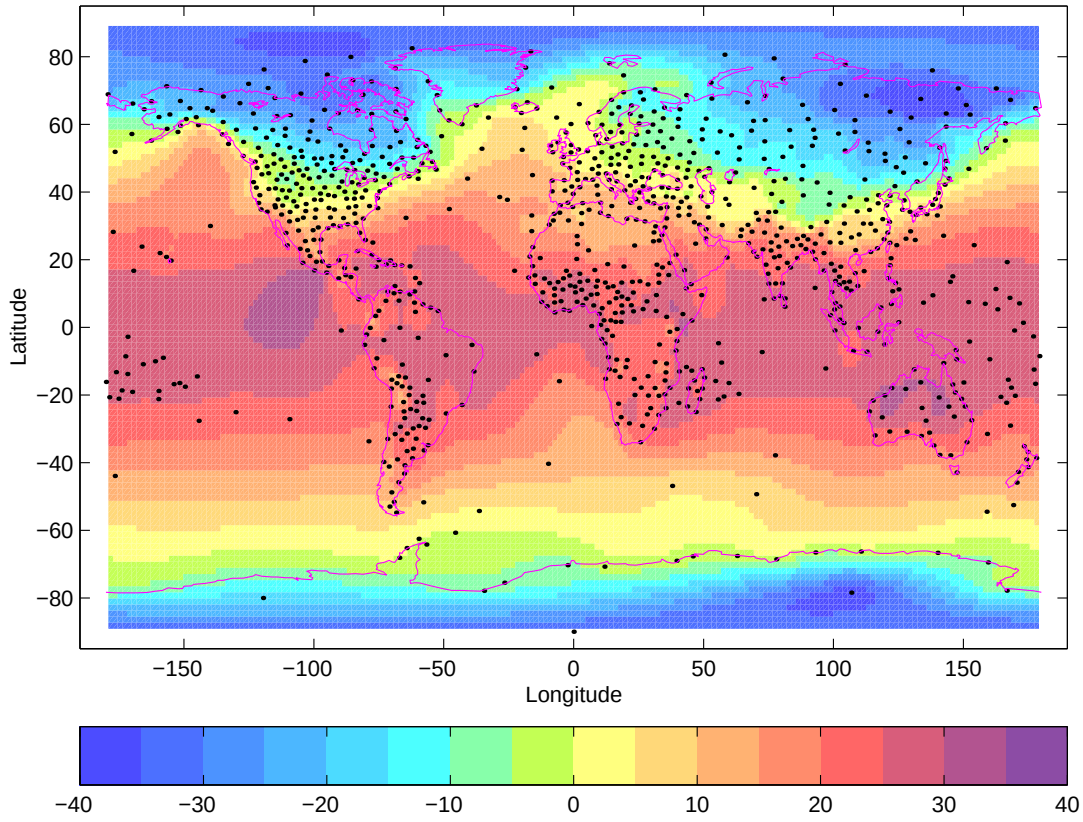
$$\int \phi(x) = \int g_1(x)dx = \int g_2(P)dP = 0$$

$$\int g_{\phi,2}(P)dP = 0$$

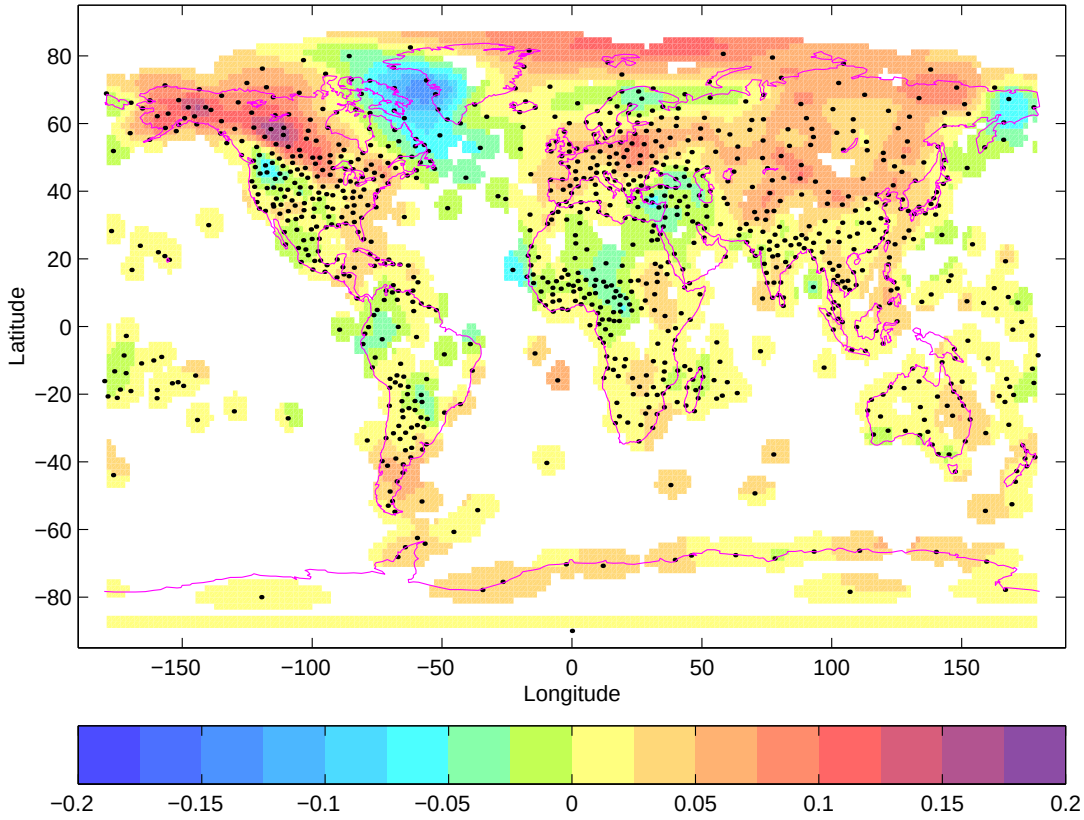
$$\int g_{12}(x,P)dx = 0$$



Yearly average winter temperature (degrees Centigrade). From Chiang, Wahba, Tribbia and Johnson TR1010. (1999)



Mean of the historical average winter temperature (^{o}C), 1961-1990.



Linear trend of the historical average winter temperature (^{O}C/yr), 1961-1990.

Non-Gaussian Data

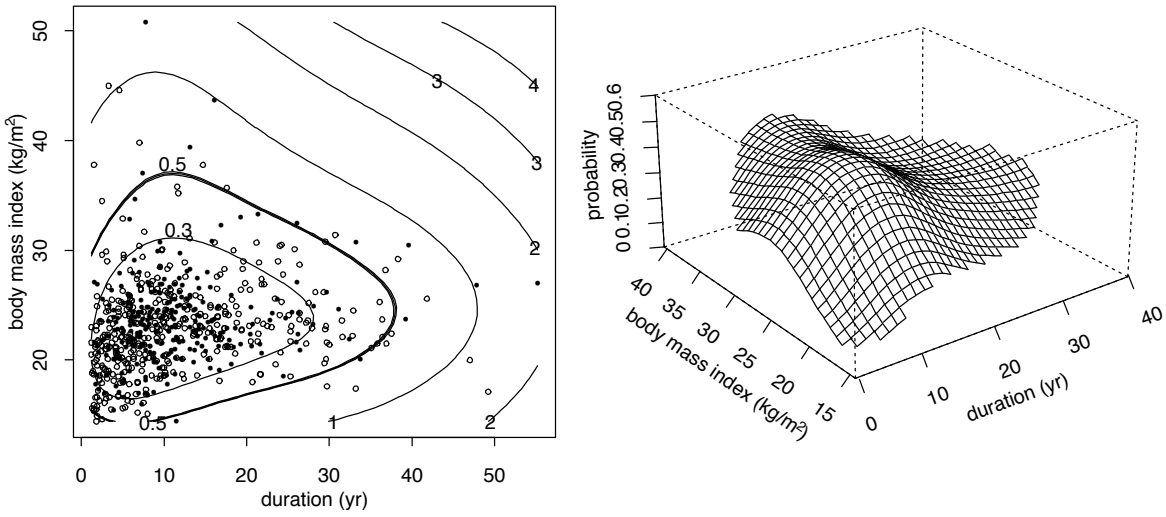
$$y_i \sim \text{Binomial}(1, p(t(i))$$

$$f = \log \frac{p}{1-p}$$

Risk of progression of diabetic retinopathy:

$$f(\text{dur}, \text{gly}, \text{bmi}) =$$

$$\mu + f_1(\text{dur}) + a_2\text{gly} + f_3(\text{bmi}) + f_{13}(\text{dur}, \text{bmi})$$



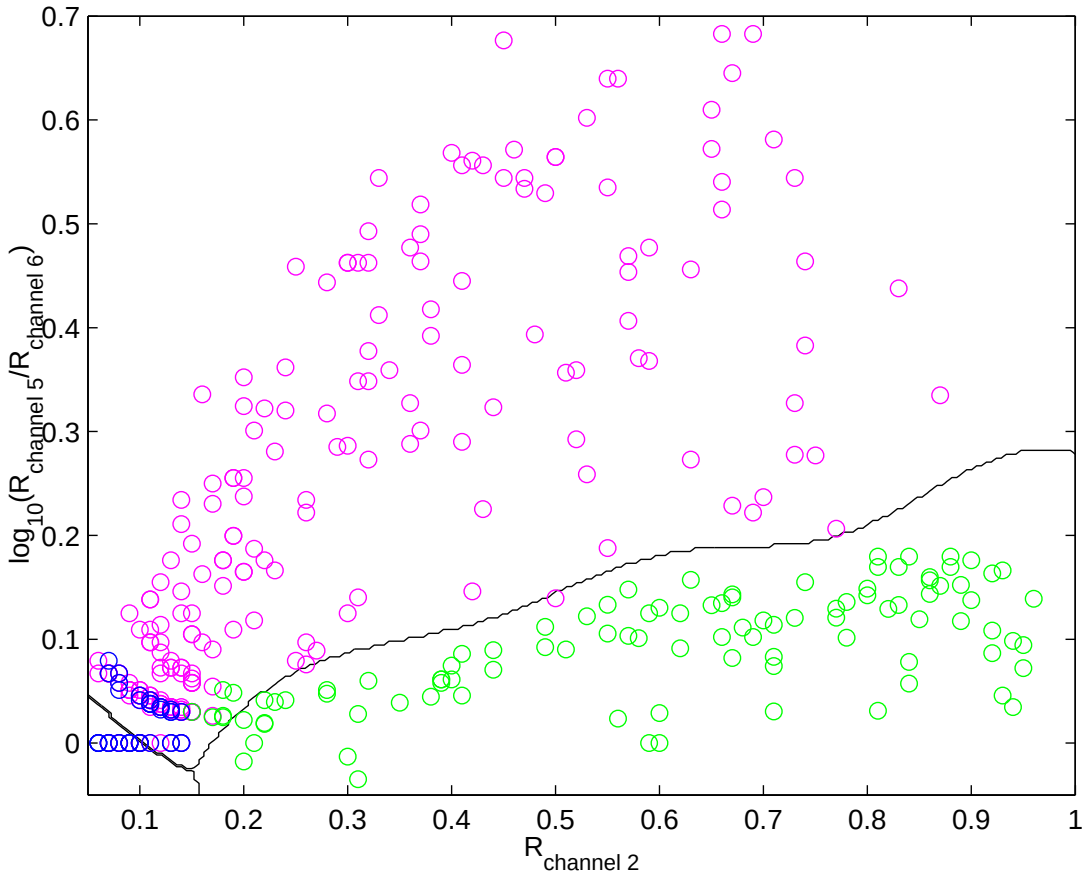
Estimated probability of five-year progression of diabetic retinopathy as a function of duration and body mass index for a measure of blood sugar fixed at its median. From Wahba, Wang, Gu, Klein and Klein Ann. Statist(1995)

Classification: Find $f \in \mathcal{X}$ to minimize

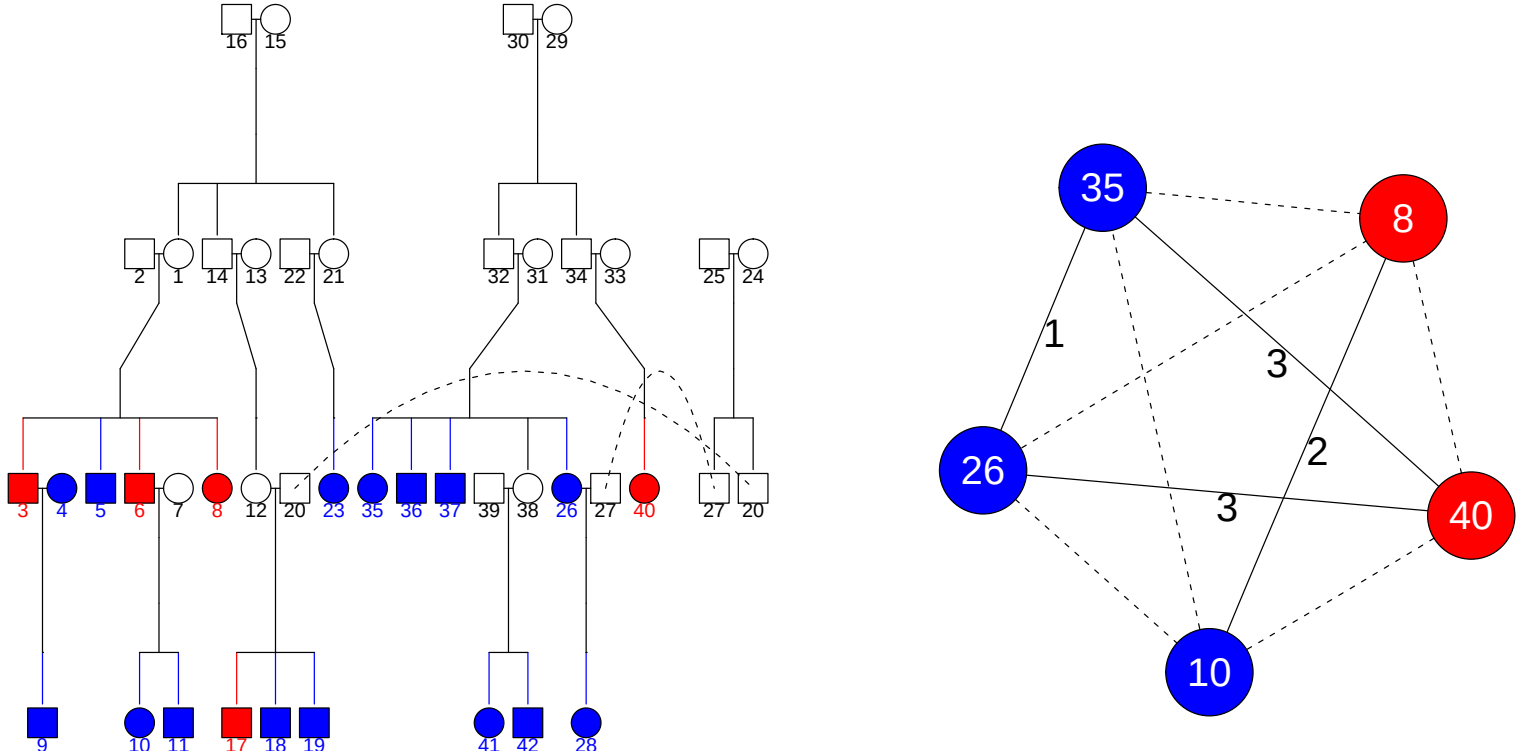
$$\frac{1}{n}\sum_i (1 - y_i f(t(i))_+ + \lambda J(f) \tag{1}$$

where $(\tau)_+ = \tau, \tau > 0, = 0$ otherwise. $J(f)$ a penalty functional, possibly one of the preceeding. The classification algorithm for an item with attribute t is given by the sign of $f(t)$.

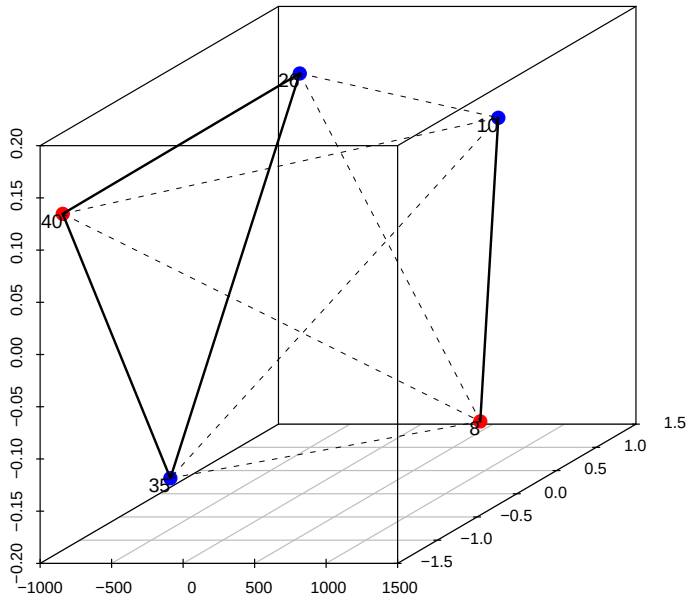
Support Vector Machine(SVM) - Classification tool: $y_i = \pm 1$ according to the $+$ or $-$ class. Figure to come is the Multicategory Support Vector Machine: For k classes, the algorithm returns $f_1(t), \cdots, f_k(t)$ satisfying $\sum_l f_l(t) \equiv 0$. The largest component determines the classification.



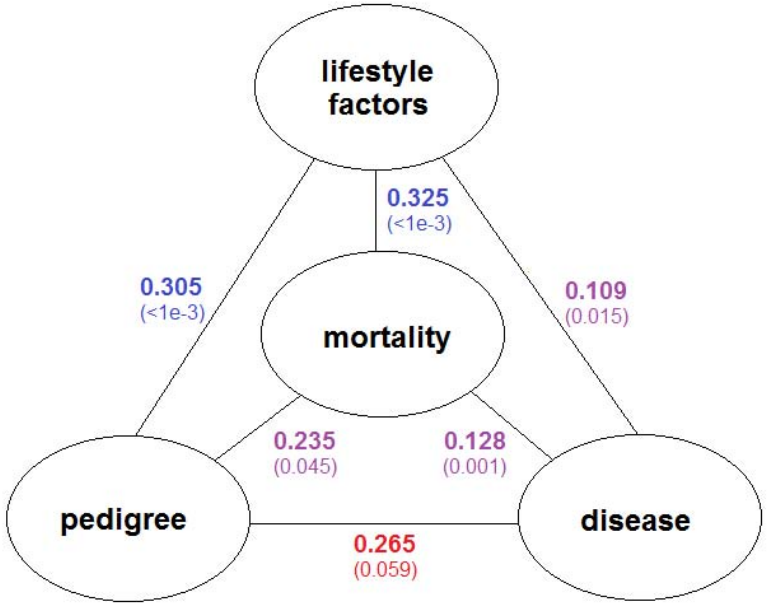
Cloud classification of satellite radiance data determined by the MSVM on 374 test samples, built on 370 labeled training samples. Clear, ice cloud, water cloud. From Lee, Wahba Ackerman, J. TECH (2004).



Dissimilarity information in pedigrees: Left: A pedigree: women circles, men squares, red= 1, blue= 2. Right: Relationship graph: Child/parent = 1, aunt/niece= 2, first coursins = 3. Corrada Bravo *et al*, PNAS (2009).



Embedding of the relationship graph. From Corrada Bravo *al* PNAS (2009).



Distance Correlation Between Lifestyle Factors, Pedigrees, Diseases and Mortality. From Kong, Klein, Klein, Lee, Wahba PNAS (2012)

A Duality properties between optimization problems and Bayes estimates

B Convergence properties

C "Bayesian" confidence intervals

D Use of side info

positivity

monotonicity

Bayesian info e.g. as in atmosphere

Non-parametric but not context-free

Five Themes (again)

1. Unified approach to estimation of functions and the building of statistical/learning models, given various kinds of observations, via regularization methods. Tuning and variable selection included. are an important special case.
2. Theoretical properties
3. Numerical methods for implementation
4. Use of software (a) to study properties of the methods by Monte Carlo methds and (b) to analyze data.
5. Applications in risk factor estimation, machine learning and classification, illposed inverse problems, protein sequence analysis, extremely large SNP sequence analysis, meteorological data analysis, others..