Smoothing Spline ANOVA Models I. Overview of Optimization Problems in Reproducing Kernel Hilbert Spaces: Varieties of Data Types, Models, Tuning Parameters

Grace Wahba
`http://www.stat.wisc.edu/~wahba`
$\rightarrow$ *TRLIST*

Joint Statistical Meetings
San Francisco, CA
August 5, 2003

Former Students

Mark C. K. Yang (1970) Bernard Viort (1972) Robert Kuhn (1974) Guido del Pino (1976) Heather Lucas Gamber (1978) Wing-Hung Wong(1980) James Wendelberger(1982) Douglas Nychka (1983) Finbarr O'Sullivan (1983) Miguel Villalobos (1983) Jyh-Jen Shiau (1985) Zehua Chen (1989) Chong Gu (1989) Feng Gao (1993) Yuedong Wang (1994) Ronaldo Dias (1994) Dong Xiang (1996) Zhen Luo (1996) Jianjian Gong (1996) Xiwu Lin (1998) Alan Y. H. Chiang (1999) Fangyu Gao (1999) Hao Helen Zhang (2002) Yoonkyung Lee (2002)

Present Students

Chenlei Leng, Ming Yuan, Xianhong Xie, John Carew, Fan Lu

Some Early Contributors

Gene Golub, Paul Speckman, Bernard Silverman, Jim Ramsay, John Rice, Mike Hutchinson, Didier Girard, Dennis Cox, Randy Eubank, Robert Kohn, Bob Anderson, Peter Bloomfield, many others.

Many Recent Collaborations

Yi Lin

And last, but not least, Where I learned about RKHS

Manny Parzen

Abstract

Overview of regularization methods in reproducing kernel Hilbert spaces and the representer theorem.Varieties of cost functions, the bias-variance tradeoff, complex penalty functionals and smoothing spline ANOVA models. An application to modeling of climate (global warming) data. Models with complex input and output structure.

OUTLINE

1. Review of positive definite matrices and functions.

2. Reproducing kernel Hilbert spaces (RKHS) and Gaussian processes.

3. Regulariztion problems in RKHS and the representer theorem.

4. Varieties of cost functions. (Univariate case)

5. The bias-variance tradeoff and adaptive tuning.

6. More complex penalty functionals (abstract version).

7. SS-ANOVA, or, ANalysis Of VAriance in RKHS.

8. A time and space model for global warming.

9. Some models with multivariate complex structure.

References

[CWTJ99] Chiang, A., Wahba, G., Tribbia, J., and Johnson, D. R. Quantitative Study of Smoothing Spline-ANOVA Based Fingerprint Methods for Attribution of Global Warming ” TR 1010, July 1999.

[GWKK01] Gao, F., Wahba, G., Klein, R. and Klein, B. Smoothing Spline ANOVA for Multivariate Bernoulli Observations, With Application to Ophthalmology Data. J.Amer.Statist. Assoc. 96 (2001) 127-160, with discussion.

[LLW02] Lee, Y, Lin, Y. and Wahba, G. Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data ” TR 1064, September 2002. TR1064r, May 2003. The revision contains further results concerning the relation of other multicategory methods to the Bayes rule.

[LWJ98] Luo, Z. , Wahba, G, and Johnson, D. R. Spatial-Temporal Analysis of Temperature Using Smoothing Spline ANOVA ” J. Climate 11, 18-28 (1998).

[LWZL02] Lin, Y., Wahba, G., Zhang, H., and Lee, Y. Statistical Properties and Adaptive Tuning of Support Vector Machines. Machine Learning, 48, 115-136, 2002.

[W92] Wahba, G. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels, in “Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity”, Proc. Vol. XII, Eds. M. Casdagli and S. Eubank, Addison-Wesley, 95-112 (1992).

[WLL02] Wahba, G., Lin, Y., and Leng, C. Penalized Log Likelihood Density Estimation via Smoothing-Spline ANOVA and ranGACV - Comments to Hansen and Kooperberg ‘Spline Adaptation in Extended Linear Models’. Statistical Science 17:33-37, 2002.

[WLLZ01] Wahba, G., Lin, Y., Lee, Y. and Zhang, H. On the Relation Between the GACV and Joachims' $\xi\alpha$ Method for Tuning Support Vector Machines, With Extension to the Nonstandard Case " TR 1039, June 2001.

[XL98] Lin, Xiwu. Smoothing Spline Analysis of Variance for Polychotomous Response Data " December 1998, Ph.D. Thesis.

[XW96] Xiang, D. and Wahba, G. A Generalized Approximate Cross Validation for Smoothing Splines with Non-Gaussian Data. Statistica Sinica, 6, 1996, pp.675-692.

[W81]Wahba, G. Spline interpolation and smoothing on the sphere. SIAM J. Sci. Statist. Comp. 2 (1981) 5-16. Erratum 3 (1982), 385-386.

Many other contributors!

♣♣ 1. Positive definite matrices and functions.

Let $\mathcal{T}$ be an index set. A symmetric function of two variables, $K(s,t), s,t \in \mathcal{T}$ is said to be positive definite (pd) if, for every n and $t_1, \cdots, t_n \in \mathcal{T}$, and every $a_1, \cdots, a_n$,

$$\sum_{i,j=1}^{n} a_i a_j K(t_i, t_j) \geq 0.$$

In the case $\mathcal{T} = \{1, 2, \cdots, N\}$ K reduces to an $N \times N$ matrix. But we will be interested in a (limitless) variety of other index sets-anything on which you can construct a positive definite function:

$\mathcal{T} = (\ldots, -1, 0, 1, \ldots)$
$\mathcal{T} = [0, 1]$
$\mathcal{T} = E^d$ (Euclidean *d*-space)
$\mathcal{T} = \mathcal{S}$ (the unit sphere)
$\mathcal{T} =$ the atmosphere
$\mathcal{T} = \{\diamondsuit, \triangle, \heartsuit\}$ (unordered set)
$\mathcal{T} =$ A Riemannian manifold
$\mathcal{T} =$ A collection of trees

etc, etc.

♣♣ 1. Positive definite matrices and functions (cont.).

For matrices A and B of appropriate dimensions, the sum $(A \oplus B)$, and the (Kronecker) product,

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ a_{21}B & \dots & a_{2n}B \\ \vdots & & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{pmatrix}$$

are pd, and this carries over to positive definite functions on arbitrary domains: Let

$$u, u' \in \mathcal{T}^{(1)}, v, v' \in \mathcal{T}^{(2)}$$

$$s = (u, v), t = (u', v') \in \mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)}$$

$$K_1(u, u'),\ K_2(v, v') \text{ be pd.}$$

Then $K \equiv K_1 \otimes K_2$:

$$K(s, t) = K_1(u, u')K_2(v, v')$$

is pd on $\mathcal{T} \otimes \mathcal{T}$. Thus tensor sums and products of pd functions on arbitrary domains provide an inexhaustible source of models.

♣♣ 2. Reproducing kernel Hilbert spaces (RKHS) and Gaussian processes.

Recall: An RKHS (reproducing kernel Hilbert space) is a Hilbert space $\mathcal{H}_K$of functions on a domain $\mathcal{T}$ with all the evaluation functionals $t : f \rightarrow f(t)$ bounded. That is, for each $t \in \mathcal{T}$ there exists a *representer* $\eta_t \in \mathcal{H}_K$ such that $f(t) =< \eta_t, f >_{\mathcal{H}_K}$.

Furthermore, let $K(s,t) =< \eta_s, \eta_t >_{\mathcal{H}_K}$. Thus, K is a uniquely determined pd function, and the famous Moore-Aronszajn theorem says that the converse is true: to each positive definite function on $\mathcal{T} \otimes \mathcal{T}$ there corresponds a unque RKHS $\mathcal{H}_K$, with

$$\eta_t(\cdot) = K(t, \cdot).$$

η_t is the so-called 'representer of evaluation' at t.

Remark: K is also the covariance of a zero mean Gaussian process (GP): (Bayesian interpretations, large GP literature in machine learning.)

♣♣ 3. Regularization Problems in RKHS.

The canonical regularization problem in RKHS: Given

$$\{y_i, t_i\}, y_i \in \mathcal{Y}, t_i \in \mathcal{T},$$

and $\{\phi_1, \cdots, \phi_M\}$, M special functions defined on $\mathcal{T}$. Find f of the form

$$f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$$

with $h \in \mathcal{H}_K$ to minimize

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

$\mathcal{C}$ is a convex function of f for each $y_i \in \mathcal{Y}$ and it is required that the minimizer of $\frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i))$ in the span of the ϕ_ν be unique. $f(t_i)$ may be replaced by $L_i(f)$, where $L_i(f)$ is a bounded linear functional on $\mathcal{H}_K$ and well defined on the ϕ_ν: For example:

$$L_i(f) = \int_{\mathcal{T}} H_i(s) f(s) ds.$$

For some $\mathcal{H}$, observed derivatives can also be used. So a wide variety of observation types can be used.

♣♣ 3. Regularization Problems in RKHS (cont.), the representer theorem.

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

$\mathcal{C}$ measures "fit to data", $\|h\|^2_{\mathcal{H}_K}$ is "complexity" and λ governs their tradeoff. The minimizer of $\mathcal{I}\{f,y\}$ has a representation of the form:

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

$d = (d_1, \cdots d_M)'$ and $c = (c_1, \cdots, c_n)'$ are found using

$$\|\sum_{i=1}^{n} c_i K(t_i, \cdot)\|^2_{\mathcal{H}_K} = c' K_n c$$

where K_n is the $n \times n$ matrix with i,jth entry $K(t_i, t_j)$. If $f(t_i)$ is replaced by $L_i(f)$ then $K(t_i, \cdot)$ is replaced by ξ_i obtained by applying L_i to one of the arguments in K, for example if $L_i(f) = \int_{\mathcal{T}} H_i(s) f(s) ds$ then

$$L_i(K(t,\cdot)) = \int_{\mathcal{T}} H_i(s) K(t,s) ds = \xi_i(t)$$

♣♣ 4. Varieties of Cost Functions (Univariate Case).

	$\mathcal{C}(y, f)$
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	$(1-yf)_+$
Other "large margin classifiers"	e^{-yf} and other functions of (yf)
.....	
(MV) Density estimation: $y \equiv 1$	$-yf + \int e^f$

(Here $(\tau)_+ = \tau, \tau \geq 0, = 0$ otherwise,
$\rho_q(\tau) = \tau(q - I(\tau \leq 0))$.

♣♣ 5. The bias-variance tradeoff and adaptive tuning.

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

Letting f_λ be the minimizer:

$$f_\lambda(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i K(t_i, t).$$

As $\lambda \to \infty$, $f_\lambda \to$ the minimizer of $\frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i))$ in span $\{\phi_\nu\}$, and as $\lambda \to 0$, $f_\lambda \to$ interpolate the data. (if K_n is strictly pd). λ controls the bias-variance tradeoff.

♣♣♣ 5. The bias-variance tradeoff and adaptive tuning (cont.).

Methods for choosing λ from the data:

- Gaussian Data: Generalized Cross Validation (GCV), Generalized Maximum Likelihood (GML)(aka REML), Unbiassed risk (UBR), others (google "methods" "choose" "smoothing parameter" gave 2850 hits)

- Bernoulli Data: Generalized Approximate Cross Validation (GACV) (XW96),other earlier related

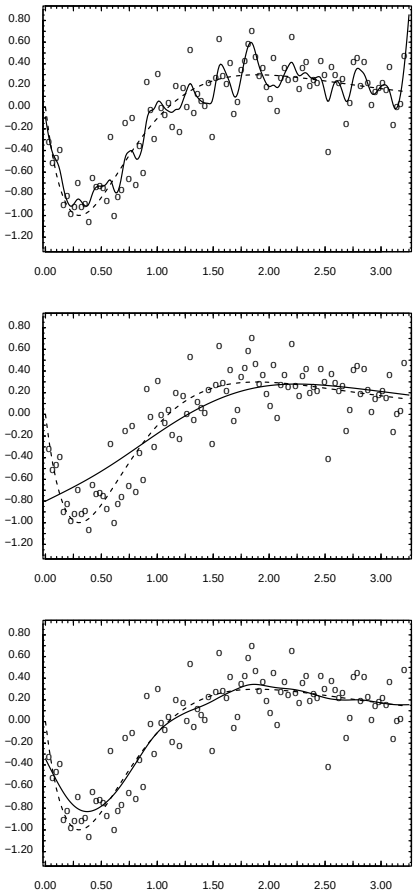
- Support Vector Machines: GACV for SVM's (WLZ00) other related, esp. Joachim's $\xi\alpha$ method.

- Multivariate Density Estimation: GACV for density estimation. (WLL02)

- All problems: Leaving-out-one, k-fold cross validation

♣♣ 5. The bias-variance tradeoff and adaptive tuning (cont.).

1979 Figure: A cubic smoothing spline-minimizer of $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int (f''(s))^2 ds$



Top to bottom: λ too small, λ too large, λ 'just right'. Gaussian white noise, λ estimated by GCV.

♣♣ 6. More complex penalty functionals, multiple smoothing parameters (abstract version).

Let $\mathcal{H}$ be the direct sum of p orthogonal subspaces,

$$\mathcal{H}_K = \sum_{\beta=1}^{p} \oplus \mathcal{H}_\beta$$

In the penalty functional $I_\lambda\{y, f\}$, replace $\lambda\|h\|^2_{\mathcal{H}_K}$ by

$$\sum_{\beta=1}^{p} \lambda_\beta \|P^\beta h\|^2_{\mathcal{H}_K} \equiv \sum_{\beta=1}^{p} \lambda_\beta \|P^\beta h\|^2_{\mathcal{H}_\beta}$$

where P^β is the orthogonal projection of h onto $\mathcal{H}_\beta$. The representer theorem along with some rescaling of components of K can be used to obtain the desired representers with the multiple smoothing parameters $\{\lambda_\beta\}$ explicitly available for tuning.

♣♣ 7. Smoothing spline ANOVA, or, analysis of variance in RKHS (SS-ANOVA).

$$t \equiv (t_1, \cdots, t_d) \in \mathcal{T} \equiv \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$$

$$f(t) = f(t_1, \cdots, t_d)).$$

Let $d\mu_\alpha$ be a probability measure on $\mathcal{T}^{(\alpha)}$ and define the averaging operator $\mathcal{E}_\alpha$ on $\mathcal{T}$ by

$$(\mathcal{E}_\alpha f)(t) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \cdots, t_d) d\mu_\alpha(t_\alpha),$$

giving the SS-ANOVA decomposition of f:

$$f(t_1, \cdots, t_d) = \mu + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$$

$$\begin{aligned}
\mu &= \prod_\alpha \mathcal{E}_\alpha f \\
f_\alpha &= (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta f \\
f_{\alpha\beta} &= (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha,\beta} \mathcal{E}_\gamma f \\
&\vdots \qquad \vdots \qquad \mathcal{E}_\alpha f_\alpha = 0, \;\; \mathcal{E}_\alpha \mathcal{E}_\beta f_{\alpha\beta} = 0, etc.
\end{aligned}$$

♣♣ 7. Smoothing spline ANOVA, or, analysis of variance in RKHS (cont.).

The idea behind SS-ANOVA is to construct an RKHS $\mathcal{H}$ of functions on $\mathcal{T}$ so that the components of the SS-ANOVA decomposition represent an orthogonal decomposition of f in $\mathcal{H}$. Then RKHS methods can be used to explicitly impose smoothness penalties of the form $\sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, (where, however, the series will be truncated at some point.)

♣♣ 7. Smoothing spline ANOVA, or, analysis of variance in RKHS (cont.).

Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha = 0$ for $f_\alpha(t_\alpha) \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$.

Construct $\mathcal{H}$ as

$$\mathcal{H} = \otimes_{\alpha=1}^{d} \left[[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)} \right]$$

$$= \otimes_{\alpha=1}^{d} [1^{(\alpha)}] \oplus \sum_j \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots ,$$

Factors of the form $[1^{(\alpha)}]$ are omitted whenever they multiply a term of a different form. Thus $\mathcal{H}^{(1)}$ is shorthand for $\mathcal{H}^{(1)} \otimes [1^{(2)}] \otimes \cdots \otimes [1^{(d)}]$ (which is a subspace of $\mathcal{H}$).

The components of the ANOVA decomposition will be in mutually orthogonal subspaces of $\mathcal{H}$.

♣♣ 7. Smoothing spline ANOVA, or, analysis of variance in RKHS (cont.).

Consider

$$I = \prod_{\alpha=1}^{d} [\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)] =$$

$$\prod_{\alpha=1}^{d} \mathcal{E}_\alpha + \sum_{\alpha=1}^{d} (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta$$

$$+ \sum_{\alpha<\beta} (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha,\beta} \mathcal{E}_\gamma + \cdots + \prod_{\alpha=1}^{d} (I - \mathcal{E}_\alpha).$$

and note that the the terms match up with the expansion of

$$\mathcal{H} = \otimes_{\alpha=1}^{d} \left[[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)} \right]$$

$J_\alpha(f) = \|P^{\mathcal{H}^{(\alpha)}} f\|^2$. (Details allowing for unpenalized terms omitted here.)

♣♣ 8. A Time and Space Model for Global Warming.

$t = (t_1, t_2) = (x, P)$, $x = 1, \cdots 30$, (year) $P = \mathcal{S}$ (latitude, longitude).

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

ϕ is linear in time orthogonal to $[1^{(1)}]$. $\mathcal{H}$ and f have the (six term) decompositions given below:

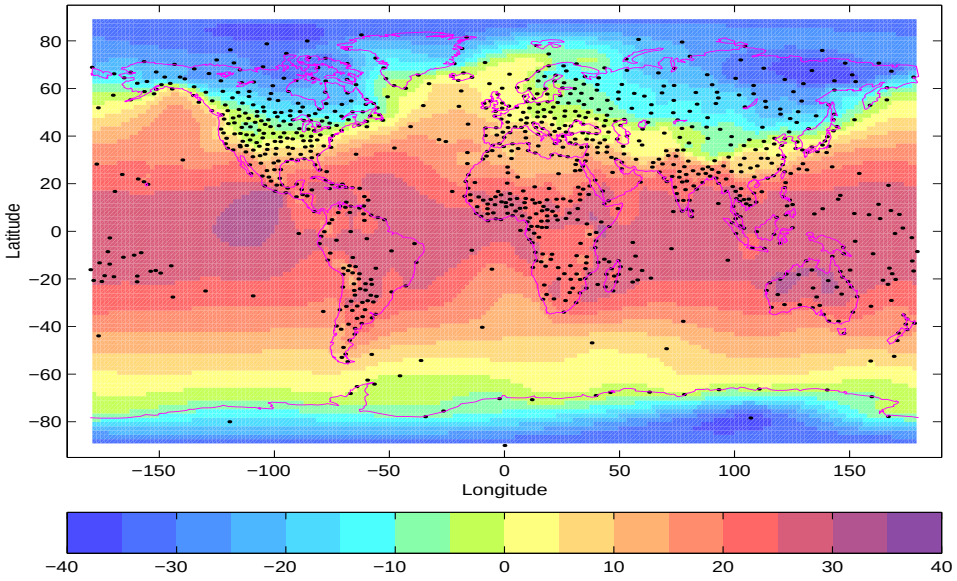
$$\begin{array}{rcccccccc}
\mathcal{H} & = & [1] & \oplus & [\phi] & \oplus & [\mathcal{H}_s^{(1)}] & \oplus & [\mathcal{H}_s^{(2)}] \\
f(x,P) & = & \mu & + & d\phi(x) & + & f_1(x) & + & f_2(P) \\
 & = & mean & + & global & + & time & + & space \\
 & & & & time & & main & & main \\
 & & & & trend & &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

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

These terms correspond to what meteorologists call anomalies, deviations from some average. Our averaging operator on time was the ordinary average and the averaging operator on the sphere was integration.

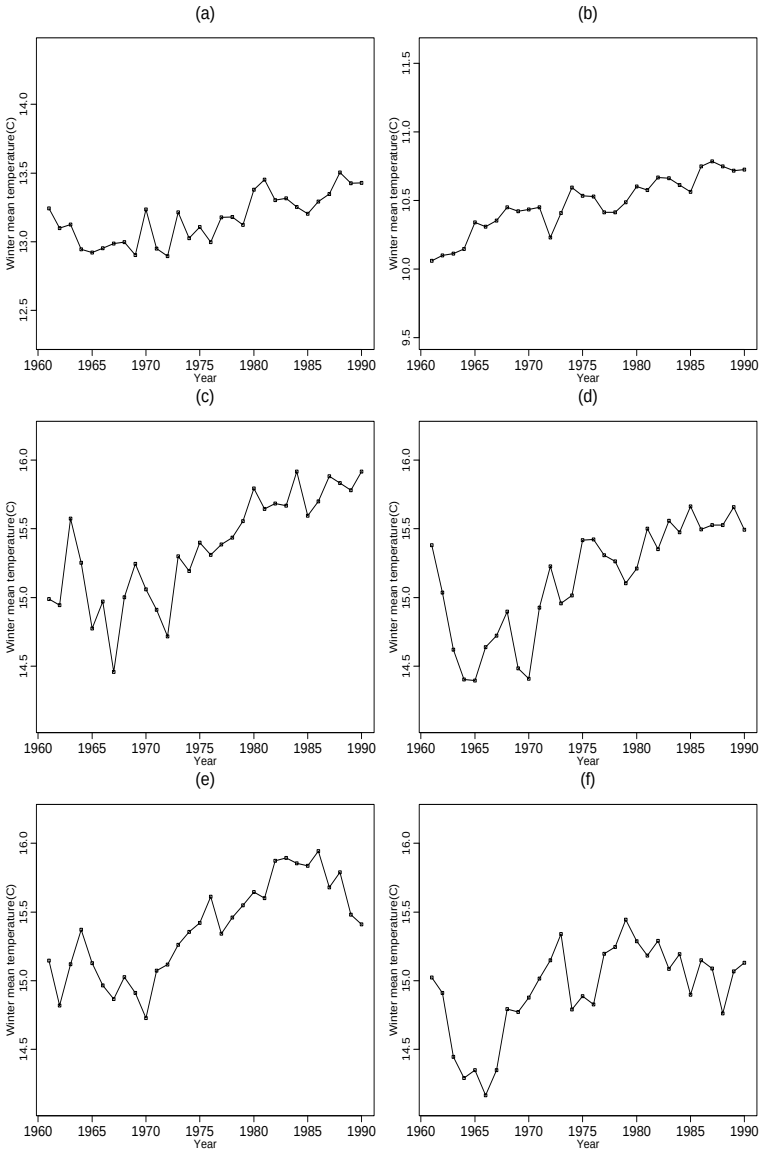
A sum of squares of second differences penalty was applied to the time variable, and a spline on the sphere penalty [W81,82] was applied to the space variable. There are four smoothing parameters for the last four terms.

We [CWJT99] fitted average winter temperatures 1961-90 from 1000 observing stations, $n = 23,119$ with missing data. Use tensor product structure to fit, with an EM-like algorithm [LWJ98] for missing data. (Figures next: Alan Chiang).

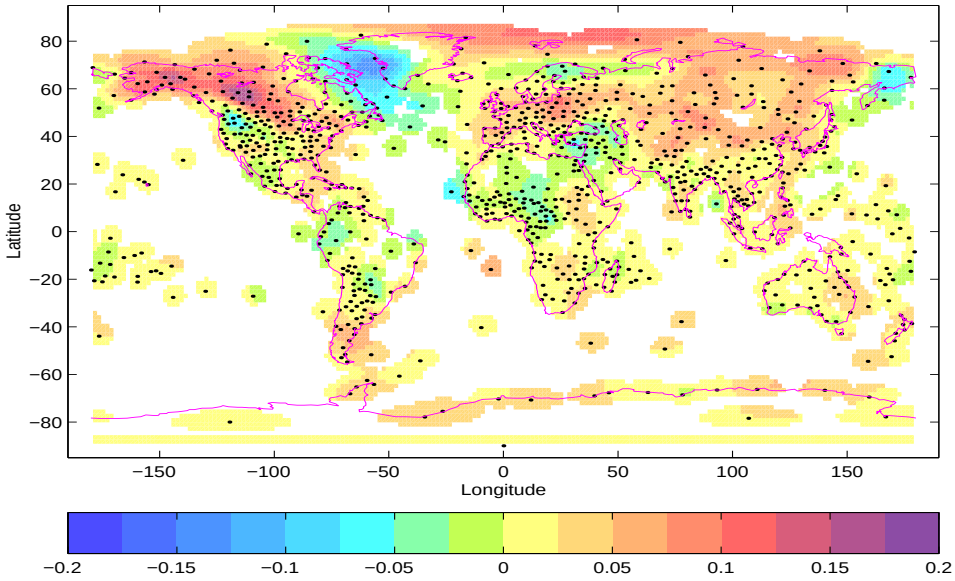


Mean of the historical average winter temperature (^{o}C), 1961-1990.

(space main effect)



Yearly average winter temperatures 1961-90 (^{O}C): (a) Historical (b)-(f) Five climate models. (mean + global time trend + time main effect)



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

(trend by space term)

Disaster for Cross-Country Skiiers in the Midwest!

♣♣ 9. Some models with complex multivariate structure.

- Multivariate correlated Gaussian observations

$$f(t) = (f_1(t), \cdots f_k(t)),$$

$$y = (y_1, \cdots, y_K)' = (f_1, \cdots, f_K)' + (\epsilon_1, \cdots, \epsilon_K)'$$

Can use

$$K(s,t) = \begin{pmatrix} K_{11}(s,t) & \dots & K_{1K}(s,t) \\ K_{21}(s,t) & \dots & K_{2K}(s,t) \\ \vdots & & \vdots \\ K_{K1}(s,t) & \dots & K_{KK}(s,t) \end{pmatrix}$$

Formulas look the same as before (!) [W92] (Meteorological and other variables which are correlated)

♣♣ Some models with complex multivariate structure (cont.).

- Multiple correlated Bernoulli observations. $y = (y_{jk}, j = 1, \cdots, J, k = 1, \cdots K_j)$ represents correlated Bernoulli outcomes for a subject with J endpoints and K_j repeated measurements on each endpoint. Then, for example [GWKK01]

$$\mathcal{C}(y, f, \alpha) =$$

$$\sum_{j=1}^{J} \sum_{k=1}^{K_j} f_{jk} y_{jk} + \sum_{j=1}^{J} \sum_{k_1 < k_2} \alpha_{jk_1, jk_2} y_{jk_1} y_{jk_2}$$

$$+ \sum_{j_1 < j_2} \sum_{k_1, k_2} \alpha_{j_1 k_1, j_2 k_2} y_{j_1 k_1} y_{j_2 k_2}$$

$$+ \ldots + \alpha_{11,12,\ldots,JK_J} y_{11} y_{12} \cdots y_{JK_J} - b(f, \alpha),$$

where the $f's$ and $\alpha's$ may depend on covariates (b detail is omitted here). ($J = 2$ eyes, $K_j = K$ diseases)

♣♣ 9. Some models with complex multivariate structure (cont.).

- Class membership-Polychotomous penalized likelihood estimates. Observe class membership of each subject and their covariates, Estimate nonparametrically, the *probability* of being in each class as a function of the covariates. [XL98]

- Class membership-Multicategory Support Vector Machines. Observe class membership and covariates and estimate the vector with 1 in the jth position if the subject is in class j, and $\frac{-1}{(k-1)}$ otherwise. [LLW02]

These two class membership models will be discussed in Lecture III.

♣♣ Bottom Line

A very broad variety of nonparametric regression and classification problems can be solved via the use of optimization problems in RKHS!

Wald II: Likelihood Basis Pursuit (LPB) for model selection.
Wald III: Polychotomous penalized likelohood and Multicategory Support Vector Machines for categorical observations.