

The Pennsylvania State University

Department of Statistics

Backfitting in Smoothing Spline ANOVA

Zhen Luo

Technical Report # 96-05

October 1996

Department of Statistics

The Penn State University

326 Thomas Building

University Park, PA 16802

Phone: 814-865-6552

Fax: 814-863-7114

Email: zhen@stat.psu.edu

<http://www.stat.psu.edu/~zhen>

Backfitting in Smoothing Spline ANOVA

By Zhen Luo¹

Department of Statistics, the Pennsylvania State University,

326 Thomas Building, University Park, PA 16802, U.S.A.

Abstract

A scheme to compute smoothing spline ANOVA estimates for large data sets with a (near) tensor-product structure is proposed. Such data sets are common in spatial-temporal analysis and image analysis. This scheme combines backfitting algorithm with iterative imputation algorithm in order to save both computational space and time. The convergence of this algorithm and various ways to further speed it up, such as collapsing component functions and successive over-relaxation, are discussed. Issues related to its application in spatial-temporal analysis are discussed too. An application to a global analysis of historical surface temperature data is described.

Key words and phrases. The Gauss-Seidel algorithm, the Gibbs sampler, spatial-temporal modeling, additive models, tensor-product data structure, global climate change.

1 Introduction

Smoothing spline ANOVA (SSANOVA) is a multivariate function estimating procedure that bases on an ANOVA type decomposition of the function to be estimated. It generalizes the decomposition of multiple factor effects in the ordinary ANOVA and the univariate smoothing spline estimating procedure. It has been applied in many areas, for example, see Gu and Wahba (1993a,b) in an environmental study and Wahba, Wang, Gu, Klein and Klein (1996) in epidemiology. In this article, an application to a climate study will be discussed.

A major difficulty with SSANOVA is a computational one. Gu (1989) carefully implemented an approach for generic smoothing spline estimation problems. It works well with small and moderate

¹This research supported in part by a startup fund provided by the Pennsylvania State University, and by NSF under Grant DMS-9121003 and NASA under Grant NAGW-2961.

size data sets. But since it does not assume any special data structure, it is inevitably slow and memory-demanding for large data sets.

A subclass of SSANOVA models is those without any kinds of interactions, i.e., smoothing spline additive models. Buja, Hastie and Tibshirani (1989) studied these models (and other additive models) and proposed a backfitting algorithm to transform a multivariate function estimating problem into many univariate function estimating problems. Since there is a sparse matrix representation of the univariate (polynomial) smoothing spline estimate (O’Sullivan, 1985), it speeds up the computation considerably and saves memory as well.

When interactions are non-negligible, other kinds of special structure in the data sets may be used to save computational space and time. In this article, a tensor product structure is the feature of large data sets that our algorithm makes use of. The key to our algorithm is also backfitting which enables us to fit a SSANOVA model with the decompositions of much smaller matrices than those without consideration of the special data structure. Because of that, we may reduce memory requirement greatly. But unlike that in additive models, correlations (in a loose sense) between component functions very often make the convergence of backfitting too slow to be feasible. In those cases, we may use some speeding-up techniques that will be a topic of this article too. Also, in many cases, the tensor-product structure is only approximately hold usually due to missing values, hence we include an iterative imputation procedure to get around with the problem.

We will use a spatial-temporal analysis of global historical surface air temperature data to illustrate, also as a primary application of, this method. To get accurate estimates of global temperature history based on scattered data, we have to avoid various biases which can be introduced through the way we collect and analyze the data. In this respect, the method proposed in this article is for correcting the biases resulting from statistical analyzing method given incomplete sampling. Techniques proposed here could be used in some other applications of spatial-temporal analysis too. Other data sets having a tensor product structure such as image data may also use the methods discussed in this article.

Section 2 introduces SS ANOVA estimates. Section 3 describes our computational strategy,

i.e., backfitting combined with iterative imputation. Various issues such as the convergence of backfitting and the validity of imputation will be addressed. Section 4 discusses several speeding-up techniques such as collapsing component functions and Successive Over-Relaxation. Section 5 describes an application to a historical global surface temperature data set. Section 6 discusses some issues related to SSANOVA procedure especially in the case of spatial-temporal analysis.

2 Smoothing spline ANOVA

In this section, we introduce smoothing spline ANOVA (SSANOVA) approach through its application to a spatial-temporal analysis of historical global surface temperature data. A general formulation of SSANOVA approach may be found in Wahba (1990), Gu and Wahba (1993a, b) and Wahba et. al. (1996).

Consider a variable, for example, winter mean surface air temperature, as a function of geographical location and year. Suppose that year is indexed by x taking values in $\{1, 2, \dots, n_1\}$ corresponding to a period of n_1 years, and location is denoted by $P = (\textit{latitude}, \textit{longitude})$ taking values on the unit sphere \mathcal{S} . The goal is to estimate function f based on scattered noisy data:

$$y_k = f(t_k) + \epsilon_k, \quad t_k \in \mathcal{T}, \quad k = 1, \dots, n \quad (1)$$

where $\mathcal{T} = \{1, 2, \dots, n_1\} \times \mathcal{S}$, and ϵ_k 's are independent noise terms. (In the case of temperature field, ϵ_k 's contain not only measurement errors but also representation errors that are associated with some high frequency signal parts of this field.)

Define two averaging operators on a function of x and P over time and space respectively by:

$$(\mathcal{E}_x f)(x, P) := \sum_{x=1}^{n_1} f(x, P)/n_1 \quad (2)$$

$$(\mathcal{E}_P f)(x, P) := \int_{\mathcal{S}} f(x, P) dP/4\pi. \quad (3)$$

A direct sum decomposition of the space of function $f(x, P)$ satisfying some minimal conditions

may hence be defined by

$$\begin{aligned}
I &= (\mathcal{E}_x + (I - \mathcal{E}_x))(\mathcal{E}_P + (I - \mathcal{E}_P)) \\
&= \mathcal{E}_x \mathcal{E}_P + (I - \mathcal{E}_x) \mathcal{E}_P + \mathcal{E}_x (I - \mathcal{E}_P) + (I - \mathcal{E}_x)(I - \mathcal{E}_P)
\end{aligned} \tag{4}$$

This decomposition singles out the average over years and the average over the globe. Suppose that we also want to single out the linear trend along the year, we may just define another averaging operator in addition to the two defined above:

$$\begin{aligned}
(\mathcal{E}'_x f)(x, P) &:= \frac{\sum_{x=1}^{n_1-1} (f(x+1, P) - f(x, P))}{n_1 - 1} \phi(x) \\
&= \frac{(f(n_1, P) - f(1, P))}{n_1 - 1} \phi(x)
\end{aligned} \tag{5}$$

where $\phi(x) = x - (n_1 + 1)/2$. Similar to (4), these three averaging operators define six unique component functions through:

$$\begin{aligned}
d_1 &:= (\mathcal{E}_x \mathcal{E}_P) f \\
d_2 \phi &:= (\mathcal{E}'_x \mathcal{E}_P) f \\
g_1 &:= (I - \mathcal{E}_x - \mathcal{E}'_x) \mathcal{E}_P f \\
g_2 &:= \mathcal{E}_x (I - \mathcal{E}_P) f \\
g_{\phi,2}(P) \phi &:= \mathcal{E}'_x (I - \mathcal{E}_P) f \\
g_{12} &:= (I - \mathcal{E}_x - \mathcal{E}'_x)(I - \mathcal{E}_P) f
\end{aligned}$$

This is equivalent to saying that f is decomposed in the following way:

$$f(x, P) = d_1 + d_2 \phi(x) + g_1(x) + g_2(P) + g_{\phi,2}(P) \phi(x) + g_{12}(x, P) \tag{6}$$

where the component functions satisfy

$$\begin{cases} \sum_{x=1}^{n_1} g_1(x) = g_1(n_1) - g_1(1) = 0 \\ \sum_{x=1}^{n_1} g_{12}(x, P) = g_{12}(n_1, P) - g_{12}(1, P) = 0 \\ \int_{\mathcal{S}} g_2(P) dP = \int_{\mathcal{S}} g_{\phi,2}(P) dP = \int_{\mathcal{S}} g_{12}(x, P) dP = 0 \end{cases} \quad (7)$$

for any x and P . This generalizes the decomposition of multiple factor effect in the ordinary ANOVA. Conditions in (7) are side conditions that make the decomposition unique, just like in the ordinary ANOVA.

These component functions are of interest because they have clearly defined practical meanings. d_1 is the grand average winter temperature over both years and the globe; d_2 is the grand linear trend of winter temperature over years (averaging over the globe); $d_1 + d_2\phi(x) + g_1(x)$ is the history of global average winter temperature. The rest three terms are the locational adjustments to the three global average terms. For example, $g_{\phi,2}(P)$ is the locational adjustment to the grand linear trend. In other words, $d_2 + g_{\phi,2}(P)$ is the linear trend at location P . A map of $d_2 + g_{\phi,2}(P)$ would give us a clear idea of where the warming areas are and where the cooling areas are over those years. Our estimating procedure is defined through this decomposition too.

In order to make inference about the function at points other than data points, we have to make some “smoothness assumption” about the function we want to estimate. We do this first by restricting our focus on a class of functions (e.g., functions with some degree of differentiability), and then penalizing extra flexibility in terms of a norm in that class.

In the class of functions of x , an inner product may be defined as

$$\begin{aligned} \langle f, g \rangle &:= \left(\sum_{x=1}^{n_1} f(x) \right) \left(\sum_{x=1}^{n_1} g(x) \right) + (f(n_1) - f(1))(g(n_1) - g(1)) \\ &+ \sum_{x=1}^{n_1-2} (f(x+2) - 2f(x+1) + f(x))(g(x+2) - 2g(x+1) + g(x)). \end{aligned} \quad (8)$$

The corresponding decomposition of the function space is denoted by:

$$\mathcal{H}^{(1)} = [1] \oplus [\phi] \oplus \mathcal{H}_a^{(1)} \quad (9)$$

with three subspaces corresponding to the three terms in (8). The first subspace consists of all constant functions, the second one consists of all linear functions summed to zero (i.e. all functions of the form $c\phi$ for some constant c), and the third one of all the functions perpendicular to the previous two.

In the class of functions of P with adequate differentiability, an inner product is defined as

$$\langle f, g \rangle := \left(\int_{\mathcal{S}} f(P) dP \right) \left(\int_{\mathcal{S}} g(P) dP \right) + \int_{\mathcal{S}} (\Delta f)(\Delta g) dP \quad (10)$$

where Δ is the Laplace-Beltrami operator, the analogue on the sphere of the Laplacian in Euclidean spaces. The corresponding decomposition of the function space is denoted by:

$$\mathcal{H}^{(2)} = [1] \oplus \mathcal{H}_a^{(2)} \quad (11)$$

where $\mathcal{H}_a^{(2)}$ contains all the functions in $\mathcal{H}^{(2)}$ which satisfies $\int_{\mathcal{S}} f(P) dP = 0$.

Now we restrict the function $f(x, P)$ into \mathcal{H} which is defined as:

$$\begin{aligned} \mathcal{H} &= \mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)} \\ &= [1] \otimes [1] \oplus [\phi] \otimes [1] \oplus \mathcal{H}_a^{(1)} \otimes [1] \oplus \\ &\quad [1] \otimes \mathcal{H}_a^{(2)} \oplus [\phi] \otimes \mathcal{H}_a^{(2)} \oplus \mathcal{H}_a^{(1)} \otimes \mathcal{H}_a^{(2)} \end{aligned} \quad (12)$$

The first two subspaces in the direct sum are combined as \mathcal{H}^0 with dimension 2. The last four are denoted by \mathcal{H}^α , for $\alpha = 1, 2, 3, 4$, in order. A corresponding decomposition of f in \mathcal{H} is

$$f = f_0 + f_1 + f_2 + f_3 + f_4 \quad (13)$$

Comparing with (6), f_0 here is the combination of the first two terms in (6), the rest have a one-to-one correspondence.

A smoothing spline estimate is then defined as the minimizer of

$$\sum_{i=1}^n (y_i - f(x_i, P_i))^2 + \frac{1}{\theta_1} J_1(g_1) + \frac{1}{\theta_2} J_2(g_2) + \frac{1}{\theta_3} J_3(g_{\phi,2}) + \frac{1}{\theta_4} J_4(g_{12}), \quad (14)$$

where θ 's are some positive numbers called smoothing parameters, J_1, J_2, J_3 and J_4 are norms in $\mathcal{H}_a^{(1)}, \mathcal{H}_a^{(2)}, \mathcal{H}_a^{(2)}$ and $\mathcal{H}_a^{(1)} \otimes \mathcal{H}_a^{(2)}$ respectively, i.e., $J_1(g_1) = \sum_{x=1}^{n_1-2} (g_1(x+2) - 2g_1(x+1) + g_1(x))^2$, $J_2(g_2) = \int_{\mathcal{S}} (\Delta f)^2 dP$, J_3 is the same as J_2 , and J_4 is derived from J_1 and J_2 as the norm of the tensor-product space.

The choices of J 's given here are not the only choices we may have. Choosing J 's and θ 's appropriate to the data will be discussed in Section 5 and 6. For now, suppose that the appropriate ones have been chosen. Let us move our attention to the computational aspect of this procedure.

3 Computing smoothing spline ANOVA estimates with backfitting

The smoothing spline estimate defined by the minimizer of (14) has a representation (see Wahba, 1990, p.12)

$$f_{\theta}(x, P) = d_0 + d_1 \phi(x) + \sum_{i=1}^n c_i \sum_{\alpha=1}^4 \theta_{\alpha} R_{\alpha}((x_i, P_i); (x, P)), \quad (15)$$

where each R_α is a nonnegative definite function decided by the choice of J_α , $d := (d_1, \dots, d_M)^T$ and $c := (c_1, \dots, c_n)^T$ are the solution of

$$\begin{cases} 0 &= S^T c \\ (Q_\theta + I)c &= (y - Sd). \end{cases} \quad (16)$$

where $y = (y_1, \dots, y_n)^T$, S is a $n \times 2$ matrix with i -th row $(1, \phi(x_i))$, $Q_\theta = \sum_{\alpha=1}^4 \theta_\alpha Q_\alpha$, and $Q_\alpha = (R_\alpha(t_i, t_j))_{i,j=1,2,\dots,n}$. R_α 's are given in Table 1.

The nonnegative definite function R_t corresponding to the norm in $\mathcal{H}_a^{(1)}$ is defined as follows.

Let L be

$$\begin{pmatrix} 1 & -2 & 1 & \cdots & 0 \\ 0 & 1 & -2 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \cdots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \quad (17)$$

Thus $J_1(f) = f^T L^T L f$. Then $R_t(j, j')$ is the jj' -th entry of $(L^T L)^\dagger$ where \dagger denotes the Moore-Penrose generalized inverse.

Table 1: *The nonnegative definite functions in the SS estimate representation (15) that correspond to the four subspaces containing the four nonparametric components in Model (6).*

α	\mathcal{H}^α	R_α
1	$\mathcal{H}_a^{(1)} \otimes [1]$	$R_1(x, P; x', P') = R_t(x, x')$
2	$[1] \otimes \mathcal{H}_a^{(2)}$	$R_2(x, p; x', P') = R_s(P, P')$
3	$[\phi] \otimes \mathcal{H}_a^{(2)}$	$R_3(x, P; x', P') = \phi(x)\phi(x')R_s(P, P')$
4	$\mathcal{H}_a^{(1)} \otimes \mathcal{H}_a^{(2)}$	$R_4(x, P; x', P') = R_t(x, x')R_s(P, P')$

The nonnegative definite function R_s corresponding to the norm in $\mathcal{H}_a^{(2)}$ is defined as:

$$R_s(P, P') = \frac{1}{2\pi} \left[\frac{1}{2} q_2(z) - \frac{1}{6} \right], \quad (18)$$

where $z = \cos(\gamma(P, P'))$, $\gamma(P, P')$ is the angle between P and P' , and

$$q_2(z) = \frac{1}{2} \left\{ \ln \left(1 + \sqrt{\frac{2}{1-z}} \right) \left[12 \left(\frac{1-z}{2} \right)^2 - 4 \left(\frac{1-z}{2} \right) \right] - 12 \left(\frac{1-z}{2} \right)^{3/2} + 6 \left(\frac{1-z}{2} \right) + 1 \right\} \quad (19)$$

(See Wahba 1981, (3.3) and (3.4). Note that this R_s does not correspond to $\int_{\mathcal{S}} (\Delta f)^2 dP$, but a norm topologically equivalent to it. In other words, we are not computing the minimizer of (14), but (14) with slight changes to J 's. The reason is computational since R_s corresponding to $\int_{\mathcal{S}} (\Delta f)^2 dP$ is too expensive to compute. By the results of Stein (1990), these changes will not make the results much different when sufficient data are available.)

When the sample size (n) is not too large, the system (16) can be solved through matrix decompositions of S and Q_θ . This approach has obvious limitations in terms of both computational speed and memory requirement. When the sample size gets larger, we have to utilize some special structures in our data set. One structure we have here is the tensor product structure of Q_α 's when the data have a tensor-product design, i.e., when we have an observation at every point (x_i, P_j) for $i = 1, 2, \dots, n_1$ and $j = 1, 2, \dots, n_2$. (Hence the sample size $n = n_1 n_2$.) Then the S and Q_α 's have the following forms:

$$\begin{aligned} S &= \mathbf{1} \otimes \tilde{S} \\ Q_1 &= \mathbf{1}\mathbf{1}^T \otimes Q_t \\ Q_2 &= Q_s \otimes \mathbf{1}\mathbf{1}^T \\ Q_3 &= Q_s \otimes \phi\phi^T \\ Q_4 &= Q_s \otimes Q_t \end{aligned}$$

where $\mathbf{1}$ is a vector of ones of appropriate length, $\phi = (\phi(1), \dots, \phi(n_1))^T$, $\tilde{S} = (\mathbf{1} \phi)_{n_1 \times 2}$, Q_s is an $n_2 \times n_2$ matrix with (i, j) -th element $R_s(P_i, P_j)$, and Q_t is an $n_1 \times n_1$ matrix with (i, j) -th element

$R_t(i, j)$.

However, even when Q_α 's have such structures, their linear combination Q_θ does not have to. Therefore, instead of solving (16), we consider an equivalent linear system derived from a representation

$$f_\theta(x, P) = d_0 + d_1\phi(x) + \sum_{\alpha=1}^4 \theta_\alpha \sum_{i=1}^n c_{i,\alpha} R_\alpha((x_i, P_i); (x, P)) \quad (20)$$

where $c_{i,\alpha}$ differs for different α . Using this representation in (14) and denoting the component functions evaluated at data points by $f_0 = Sd, f_\alpha = \theta_\alpha Q_\alpha c_\alpha$ for $\alpha = 1, 2, 3, 4$, we have:

$$f_\beta = S_\beta(y - \sum_{\alpha \neq \beta} f_\alpha), \text{ for } \beta = 0, 1, \dots, 4, \quad (21)$$

where $S_0 := S(S^T S)^{-1} S^T$ and $S_\beta := (Q_\beta + \frac{1}{\theta_\beta} I)^{-1} Q_\beta$ for $\beta = 1, 2, 3, 4$.

System (21) suggests a natural iterative method to solve itself, i.e.

$$f_\beta^{(k)} = S_\beta(y - \sum_{\alpha < \beta} f_\alpha^{(k)} - \sum_{\alpha > \beta} f_\alpha^{(k-1)}), \text{ for } \beta = 0, 1, \dots, 4. \quad (22)$$

This is exactly the backfitting algorithm studied by Buja, Hastie and Tibshirani (1989) in which additive models are fitted by backfitting where each S_β (one dimensional smoother) has a sparse matrix representation. Each S_β here has a tensor product structure because S and Q_β have. Therefore, updating in (22) can be done with the decompositions of matrices Q_s and Q_t which are much smaller than Q_β 's. In this way, we may save computational memory and time required just like what sparsity structure does to additive models.

Buja et. al. (1989) discussed the properties of backfitting in a more general setup. Chen, Gu and Wahba (1989) pointed out the possibility of applying backfitting in SSANOVA. Another interesting discussion about the convergence of backfitting is given by Ansley and Kohn (1994). In the case of smoothing spline, the backfitting algorithm is equivalent to an alternating minimization

procedure to the problem

$$\min_{f_0 \in \mathcal{L}(S), f_\alpha \in \mathcal{L}(Q_\alpha)} \|y - \sum_{\alpha=0}^p f_\alpha\|^2 + \sum_{\alpha=1}^p \frac{1}{\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha \quad (23)$$

where Q_α^\dagger is the Moore-Penrose generalized inverse of Q_α , and $\mathcal{L}(A)$ denotes the space spanned by the columns of A . Because of this equivalence, we know immediately that this iterative method converges to the solution of (21) using results in the optimization literature. (See, for example, Lunerberg (1984), Section 7.9 on pp. 227-228).

Rewrite equations (21) as

$$\begin{pmatrix} I & S_0 & \cdots & S_0 \\ S_1 & I & \cdots & S_1 \\ \cdots & & & \\ S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_p \end{pmatrix} = \begin{pmatrix} S_0 y \\ S_1 y \\ \vdots \\ S_p y \end{pmatrix}. \quad (24)$$

It is clear that the backfitting algorithm we have just described, i.e., (22), is a (block) Gauss-Seidel algorithm.

Having known $f_0 (= Sd)$, we know d immediately. By (16), $(Q_\theta + I)c = y - Sd$, hence

$$c = y - Sd - Q_\theta c = y - \sum_{\alpha=0}^p f_\alpha. \quad (25)$$

Therefore c is available after we get the f_α 's.

So far we have assumed that our data are complete in the sense that at every tensor product grid point there is an observation. But frequently in observational studies, we have data missing here and there. For example, in the climate study to be described in Section 5, nearly one third of the complete data is missing. In such cases we can still use the above computational procedure with the help of an iterative imputation procedure.

For simplicity reason, suppose that we have reordered the data in such a way that the complete

data y can be written in two parts

$$y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix}, \quad (26)$$

where $y^{(2)} = (y_{i_1}, \dots, y_{i_K})^T$ is the missing part, and $y^{(1)}$ is the observed part. The iterative imputation procedure is to impute the missing part with any initial values (of course, if we start with good ones, we will be able to converge to the results faster), then fit a smoothing spline model to the complete “data” using backfitting. After that, calculate the model’s prediction at the missing part, say, $g^{(2)}$, then impute $y^{(2)}$ with these newly predicted $g^{(2)}$ and go back to fit the same SS model again. Keep going through this cycle until the fitted values do not change anymore.

It can be shown that this iterative imputation procedure is equivalent to the EM algorithm. See Dempster, Laird and Rubin (1977) and Wu (1983) for the EM algorithm and its properties. Also see Green (1990) for its use in penalized likelihood estimation. The convergence results about EM algorithms may be used to show convergence here. In Wahba and Luo (1996), a simple necessary and sufficient condition is derived directly for the validity of this iterative imputation algorithm. Let Γ_1 be an $n \times M$ matrix of orthonormal columns which span the column space of S , partitioned after the first $n - K$ rows to match y in (26) as

$$\begin{pmatrix} \Gamma_{11} \\ \Gamma_{21} \end{pmatrix}. \quad (27)$$

A necessary and sufficient condition for the iterative imputation solution with arbitrary initial values to converge to the smoothing spline estimates based on the real data is “ $\Gamma_{21}\Gamma_{21}^T$ does not have 1 as its eigenvalue”. This condition can be checked easily for a given S . Yates (1933) used a similar idea to fit an ordinary ANOVA model to the data with a few missing values without solving

a general linear model equation. An interpretation of this condition is based on the observation

$$S(S^T S)^{-1} S^T = \Gamma_1 \Gamma_1^T = \begin{pmatrix} \Gamma_{11} \Gamma_{11}^T & \Gamma_{11} \Gamma_{21}^T \\ \Gamma_{21} \Gamma_{11}^T & \Gamma_{21} \Gamma_{21}^T \end{pmatrix}. \quad (28)$$

We see that $\Gamma_{21} \Gamma_{21}^T$ is in the same position as the diagonal elements of a “hat” matrix are in an ordinary linear regression. Hence it can be interpreted as a measure of influence of those missing data points on the SS fit. Since the largest possible eigenvalue of $\Gamma_{21} \Gamma_{21}^T$ is 1, this condition just excludes the most extreme influential case.

4 Techniques to speed up backfitting

A straight-forward implementation of the backfitting algorithm often needs many iteration steps to converge. There are many discussions about speeding up the Gauss-Seidel algorithm in the numerical analysis literature, especially about a technique called Successive Over-Relaxation. See, for example, Young (1971). Here we would like to discuss some techniques in the context of fitting a smoothing spline ANOVA model.

4.1 Orthogonality

Roughly speaking, the main reason for the slowness of convergence of the backfitting algorithm is the correlation between components. To illustrate the point, consider a trivial problem of minimizing $f(c_1, c_2) = c_1^2 + 2\rho c_1 c_2 + c_2^2$ where ρ is between -1 and 1 . The spectral radius of the updating matrix of the alternating minimization (i.e., backfitting, also Gauss-Seidel) algorithm is easily verified to be ρ^2 . Hence the larger “correlation coefficient” ρ is, the slower the backfitting algorithm converges. If ρ is zero, then the backfitting algorithm converges in one step. Therefore, if possible, we may want to formulate the original problem in such a way that as many off-diagonal elements as possible are zero and thus the problem may be reduced into smaller ones.

Recall that $Q_t = (L^T L)^\dagger$ where L is given by (17), hence $Q_t \mathbf{1} = Q_t \phi = \phi^T \mathbf{1} = 0$. Therefore, all

$Q_\alpha Q_\beta$ for $\alpha \neq \beta$ and $Q_\alpha S$ are zero except $Q_1 Q_4$, $Q_2 S$, and $Q_3 S$. Thus the minimization problem (23) can be separated into two smaller ones. For $f_0 \in \mathcal{L}(S)$, $f_\alpha \in \mathcal{L}(Q_\alpha)$, we know that $f_\alpha^T f_\beta = 0$ for any $\alpha \in \{0, 2, 3\}$ and $\beta \in \{1, 4\}$, hence

$$\begin{aligned}
& \|y - \sum_{\alpha=0}^4 f_\alpha\|^2 + \sum_{\alpha=1}^4 \frac{1}{\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha \\
= & \|y - f_0 - f_2 - f_3\|^2 + \frac{1}{\theta_2} f_2^T Q_2^\dagger f_2 + \frac{1}{\theta_3} f_3^T Q_3^\dagger f_3 + \\
& \|y - f_1 - f_4\|^2 + \frac{1}{\theta_1} f_1^T Q_1^\dagger f_1 + \frac{1}{\theta_4} f_4^T Q_4^\dagger f_4 \\
& - \|y\|^2.
\end{aligned} \tag{29}$$

Therefore, (23) is equivalent to solving the following two problems separately:

$$\min_{f_0 \in \mathcal{L}(S), f_2 \in \mathcal{L}(Q_2), f_3 \in \mathcal{L}(Q_3)} \|y - f_0 - f_2 - f_3\|^2 + \frac{1}{\theta_2} f_2^T Q_2^\dagger f_2 + \frac{1}{\theta_3} f_3^T Q_3^\dagger f_3 \tag{30}$$

and

$$\min_{f_1 \in \mathcal{L}(Q_1), f_4 \in \mathcal{L}(Q_4)} \|y - f_1 - f_4\|^2 + \frac{1}{\theta_1} f_1^T Q_1^\dagger f_1 + \frac{1}{\theta_4} f_4^T Q_4^\dagger f_4. \tag{31}$$

They correspond to solving the following two systems respectively:

$$\begin{pmatrix} I & S_0 & S_0 \\ S_2 & I & 0 \\ S_3 & 0 & I \end{pmatrix} \begin{pmatrix} f_0 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} S_0 y \\ S_2 y \\ S_3 y \end{pmatrix}, \tag{32}$$

and

$$\begin{pmatrix} I & S_1 \\ S_4 & I \end{pmatrix} \begin{pmatrix} f_1 \\ f_4 \end{pmatrix} = \begin{pmatrix} S_1 y \\ S_4 y \end{pmatrix}. \tag{33}$$

The key for such a reduction is that the x variable has an equally-spaced design. If the design

of every variable is equally-spaced, then choosing appropriate J_α 's, hence R_α 's, can make all $Q_\alpha Q_\beta$ for $\alpha \neq \beta$ and $Q_\alpha S$ zero, hence $f_\alpha = S_\alpha y$. That is to say that we only need to apply marginal smoothers to the data once in order to get all component functions including interaction terms.

4.2 Grouping and Collapsing

Consider problem (23). Instead of minimizing it with respect to one component by one component which leads to the backfitting algorithm (22), we may minimize it with respect to more than one component at a time. This is called ‘‘grouping’’. In many cases grouping will reduce the number of iterations needed in the backfitting algorithm (see Varga, 1962, for a discussion and a counter-example, p. 80). Of course, each updating step is more complicated in this case due to the higher dimension of each updating step. A compromise between the cost of updating and the number of iterations needed has to be considered.

Another way in a similar spirit of grouping in order to reduce the number of iterations is to use what we call a ‘‘collapsing’’ technique. As mentioned before, very often the slowness of backfitting algorithm is due to a strong ‘‘correlation’’ between two component functions. Suppose f_{α_1} and f_{α_2} are such two components. By manipulating

$$f_\beta = (Q_\beta + \frac{1}{\theta_\beta} I)^{-1} Q_\beta (y - \sum_{\alpha \neq \beta} f_\alpha), \quad \text{for } \beta = \alpha_1, \alpha_2,$$

we get

$$f_{\alpha_1} + f_{\alpha_2} = (Q_{\alpha_1+\alpha_2} + I)^{-1} Q_{\alpha_1+\alpha_2} (y - \sum_{\alpha \neq \alpha_1, \alpha_2} f_\alpha) \quad (34)$$

where $Q_{\alpha_1+\alpha_2}$ denotes $\theta_{\alpha_1} Q_{\alpha_1} + \theta_{\alpha_2} Q_{\alpha_2}$. When $Q_{\alpha_1+\alpha_2}$ has a tensor-product structure, this updating step can be done in a similar way as in (22). Thus f_{α_1} and f_{α_2} are collapsed into one component. In this way, the strong correlation between f_{α_1} and f_{α_2} would not affect the speed of the backfitting algorithm. $f_{\alpha_1} + f_{\alpha_2}$ can then be used in (25) to compute c and then f_{α_1} and f_{α_2} . Similar techniques

have been proposed in Markov Chain Monte Carlo literature to speed up the Gibbs sampler (Liu, 1994), but seemingly not in numerical analysis literature. The properties of this technique may be studied in a similar manner as in the work done for the Gibbs sampler.

In the case of f_1 and f_4 , due to the orthogonality,

$$f_1 + f_4 = (Q_{1+4} + I)^{-1}Q_{1+4}y, \quad (35)$$

where $Q_{1+4} = \theta_1(11^T \otimes Q_t) + \theta_4(Q_s \otimes Q_t) = (\theta_1 11^T + \theta_4 Q_s) \otimes Q_t$. In the case of f_2 and f_3 ,

$$f_2 + f_3 = (Q_{2+3} + I)^{-1}Q_{2+3}(y - f_0), \quad (36)$$

where $Q_{2+3} := \theta_2(Q_s \otimes 11^T) + \theta_3(Q_s \otimes \phi\phi^T) = Q_s \otimes (\theta_2 11^T + \theta_3 \phi\phi^T)$. Since

$$\begin{aligned} f_0 &= S_0(y - f_2 - f_3) \\ &= S_0(y - (Q_{2+3} + I)^{-1}Q_{2+3}(y - f_0)), \end{aligned}$$

Simple manipulate leads to

$$d = (S^T(I + Q_{2+3})^{-1}S)^{-1}S^T(I + Q_{2+3})^{-1}y, \quad (37)$$

which can be computed directly using the eigen-decomposition of $Q_{2+3} = Q_s \otimes (\theta_2 11^T + \theta_3 \phi\phi^T)$. Then f_2 and f_3 can be computed using $f_2 = S_2(y - f_0)$, $f_3 = S_3(y - f_0)$.

If the iteration of backfitting converges too slowly, then the extra cost of matrix decompositions and matrix products may be worthwhile if we want to save overall computing time. Note that if we apply the same argument to all four f_α 's, we will end up with (16), where Q_θ , unlike Q_{1+4} or Q_{2+3} , does not have a tensor product structure, thus it is much more difficult to invert $(Q_\theta + I)$ than to invert, say $(Q_{1+4} + I)$. Therefore the problem here is also to decide how much further we want to break down the original problem. If too much, we may end up with too many backfitting

iterations. If not enough, the updating equations may be impossible or too expensive to solve.

4.3 Successive Over-Relaxation

A very important technique to speed up the backfitting (Gauss-Seidel) algorithm is through Successive Over-Relaxation (abbreviated SOR). See, for example, Golub and Van Loan (1989), or Young (1971).

Suppose we want to solve

$$\begin{pmatrix} I & S_0 & \cdots & S_0 \\ S_1 & I & \cdots & S_1 \\ \cdots & & & \\ S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \cdots \\ f_p \end{pmatrix} = \begin{pmatrix} S_0 y \\ S_1 y \\ \cdots \\ S_p y \end{pmatrix}. \quad (38)$$

The Gauss-Seidel (backfitting) updating scheme is

$$f_\alpha^{(k+1)} = S_\alpha(y - \sum_{\beta < \alpha} f_\beta^{(k+1)} - \sum_{\beta > \alpha} f_\beta^{(k)}). \quad (39)$$

The SOR scheme is

$$f_\alpha^{(k+1)} = \omega \{ S_\alpha(y - \sum_{\beta < \alpha} f_\beta^{(k+1)} - \sum_{\beta > \alpha} f_\beta^{(k)}) \} + (1 - \omega) f_\alpha^{(k)}, \quad (40)$$

where ω is a real number known as the relaxation factor. With $\omega = 1$, we are back to the Gauss-Seidel algorithm. When $\omega < 1$ or $\omega > 1$, we have underrelaxation or overrelaxation respectively.

The trick is to find a good ω . In general a prescribed optimal ω is available only for some special kinds of matrix. Fortunately our case falls into this category. Consider system (32), and denote:

$$A := \begin{pmatrix} I & S_0 & S_0 \\ S_2 & I & 0 \\ S_3 & 0 & I \end{pmatrix}.$$

Obviously A is consistently ordered (see Young 1971, pp. 144-145). If we can show that all the eigenvalues of $B := I - (\text{diag } A)^{-1}A$ are real and have absolute values less than 1, then according to Theorem 2.2 on page 172 of Young (1971), SOR will converge for any ω in $(0, 2)$.

Since

$$\begin{aligned} |B - \lambda I| &= \left| \begin{pmatrix} -\lambda I & -S_0 & -S_0 \\ -S_2 & -\lambda I & 0 \\ -S_3 & 0 & -\lambda I \end{pmatrix} \right| \\ &= (-1)^{3n} \lambda^{2n} \left| \lambda I - (S_0 \ S_0)(\lambda I)^{-1} \begin{pmatrix} S_2 \\ S_3 \end{pmatrix} \right| \\ &= (-1)^{3n} \lambda^n |\lambda^2 I - S_0(S_2 + S_3)| \end{aligned}$$

(this is true for all nonzero λ , hence for all λ , because both sides are continuous.)

Therefore all the eigenvalues of B are

$$\{0, \pm\sqrt{\mu_i}, i = 1, \dots, n\},$$

where $\{\mu_1, \dots, \mu_n\}$ are eigenvalues of $S_0(S_2 + S_3)$ and 0 has a multiplicity n .

They are certainly real, since all S_0, S_2, S_3 are non-negative definite. We only need to show that their absolute values are less than 1. We know S_0 has eigenvalues either 0 or 1 since it is a projection matrix. So we just need to show $S_2 + S_3$ has all its eigenvalues less than 1 in absolute value.

Let $Q_s = \Gamma_s \Lambda_s \Gamma_s^T$, $Q_t = \Gamma_t \Lambda_t \Gamma_t^T$, $\Lambda_s = \text{diag}(\lambda_j^s)_{j=1}^{n_2}$, $\Lambda_t = \text{diag}(\lambda_i^t)_{i=1}^{n_1}$. Since $Q_t \mathbf{1} = Q_t \phi = 0$, and $\phi^T \mathbf{1} = 0$, we can choose Γ_t so that its first two columns are $1/\sqrt{n_1}$ and $\phi/\|\phi\|$, where $\|\phi\| = \sqrt{\sum_{x=1}^{n_1} \phi^2(x)}$. So

$$\begin{aligned} S_2 &= (Q_2 + \frac{1}{\theta_2} I)^{-1} Q_2 = (\Gamma_s \otimes \Gamma_t) ((\Lambda_s \otimes \Lambda_2 + \frac{1}{\theta_2} I)^{-1} (\Lambda_s \otimes \Lambda_2)) (\Gamma_s \otimes \Gamma_t)^T, \\ S_3 &= (Q_3 + \frac{1}{\theta_3} I)^{-1} Q_3 = (\Gamma_s \otimes \Gamma_t) ((\Lambda_s \otimes \Lambda_3 + \frac{1}{\theta_3} I)^{-1} (\Lambda_s \otimes \Lambda_3)) (\Gamma_s \otimes \Gamma_t)^T, \end{aligned}$$

where Λ_2 is a $n_1 \times n_1$ matrix with all its elements being zero except the first diagonal one which is n_1 , Λ_3 is a $n_1 \times n_1$ matrix with all its elements being zero except the second diagonal one which is $\|\phi\|^2$. Hence, we can see that the eigenvalues of $(S_2 + S_3)$ are

$$\{0, \frac{\lambda_j^s n_1}{\lambda_j^s n_1 + 1/\theta_2}, \frac{\lambda_j^s \|\phi\|^2}{\lambda_j^s \|\phi\|^2 + 1/\theta_3}, j = 1, 2, \dots, n_2\},$$

where 0 has a multiplicity $(n_1 - 2) \times n_2$. Therefore, all $(S_2 + S_3)$'s eigenvalues are in $[0, 1)$.

According to Theorem 2.2 of Young (1971, p.172), SOR converges for any choice of ω between 0 and 2. Furthermore, according to Theorem 2.3 on the same page, the best choice of ω is

$$\omega_b = \frac{2}{1 + \sqrt{1 - \bar{\mu}^2}}$$

where $\bar{\mu}$ is the spectral radius of B . It can be shown (Young 1971, Theorem 2.2, p. 142) that $\bar{\mu}^2$ is the spectral radius of the Gauss-Seidel iteration matrix which can be estimated by the power method after some Gauss-Seidel iteration steps are taken. See Young (1971, p. 206) for an explanation.

It is even easier to show that SOR for the system (33) converges too. Moreover, its optimal over-relaxation parameter also can be computed given the estimate of the spectral radius of the corresponding Gauss-Seidel iteration matrix.

Note that the cited results of Young (1971) are only for the (point) Gauss-Seidel or SOR algorithms. In our case, however, point and block versions are the same because the diagonal blocks in our linear system are all identity matrices, hence updating elements in one block one by one is the same as updating them simultaneously.

5 An application to global temperature change study

An accurate and easily accessible description of what has happened in earth climate is always of interest. It is of greater interest especially in recent years when scientists start to model the global climate and use their models to predict future climate. An important step in getting more

confidence in these models is to compare their “prediction” of the past climate with what was actually observed. This is an important reason why an accurate account of the past climate is desirable.

Temperature is certainly one of the most important variables in the climate. It is also the most intensively recorded variable so far. For a long time we only have surface station temperature records. Therefore, we have to reconstruct the whole temperature history over the sphere using these records scattered in both time and space. Various biases such as the relocation of a surface station, the change of instrument and so on exist. Another important source of bias is the incomplete time and space coverage. All these potential biases make the seemingly easy summarization job complicated. Many people have taken different approaches to correct these biases. See, e.g., Hansen and Lebedeff (1987), Jones et. al. (1986), and Vinnikov et. al. (1990). Some have also studied the effect of incomplete sampling on the estimates of the climate history. See Madden et. al. (1993) and Karl et. al. (1994).

To compare global means across time (the crudest way to look at global temperature change), we encounter biases resulting from incomplete sampling, i.e., the stations having records hence included in calculating global means are different from one year to another. If in one year the relative number of stations in a cold area is bigger than in the next year, then we do not know whether an increase in average temperature is due to a real global change, or just due to the difference between two sets of stations used in two years. The way the studies cited above choose to correct this bias is through the use of anomalies which is defined as the difference between the raw records and the average over a pre-specified reference period. While it is satisfactory in some sense, this approach has a limitation in correcting the biases resulting from incomplete sampling. It has considered the variation of mean temperatures, but not the variation of temperature change at different places. In essence, the anomaly approach has implicitly assumed that the last two terms in (6) are negligible. Figure 1 shows three estimates of the global average winter temperature history based on the averages of each year’s raw records, the averages of each year’s anomalies, and the smoothing spline estimate defined by (14). We can see that the smoothing spline ANOVA approach

can correct such biases without even using anomalies.

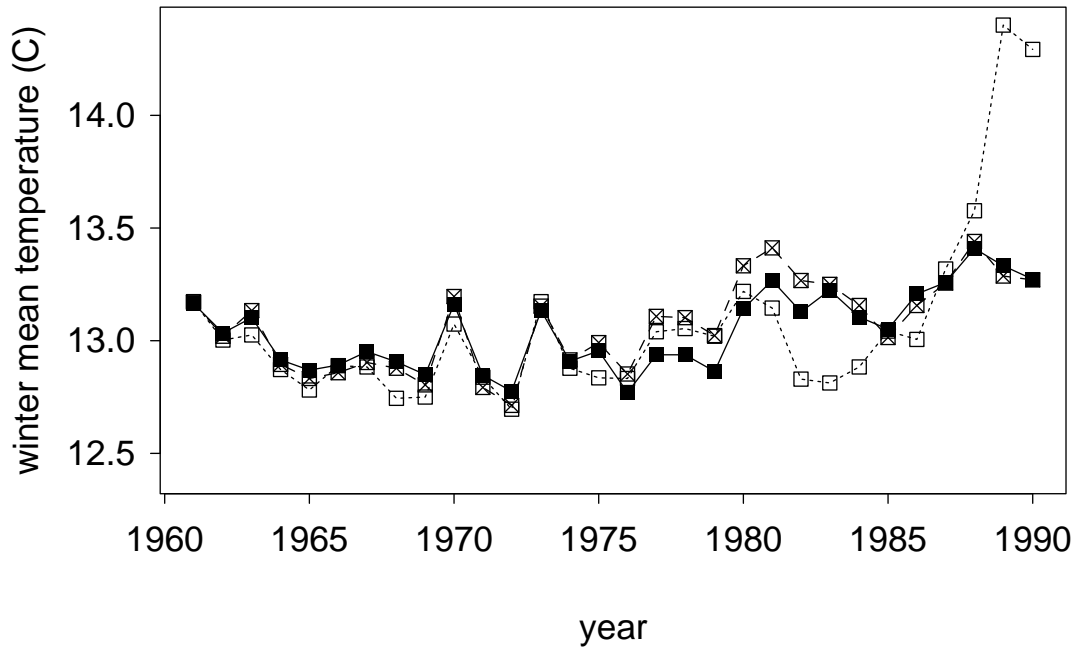


Figure 1: *Estimates of global average winter temperatures ($^{\circ}C$). Squares: yearly averages of raw data; squares with cross inside: yearly averages of anomaly; solid squares: based on the smoothing spline estimate defined by (14).*

The data set we used is so called Jones-Wigley data set (see Jones, et. al., 1991). We obtained this data set from <http://cdiac.ESD.ORNL.GOV/ftp/>. It is a combination of four files: `ndp020r1/jonesnh.dat`, `ndp020r1/jonessh.dat`, `ndp032/ndp032.tm1` and `ndp032/ndp032.tm2`. This data set is assembled from different sources of monthly temperature records at about 2000 stations distributed across the world over the period from 1851 through 1991. There are only a few stations with records dating back that far. Most of the stations started recording in this century. The stations are concentrated heavily in Europe and North America. Jones et. al. (1991) have done some cleaning and homogenizing to the original data. Only winter average temperature, defined as the average of December, January and February temperatures, is considered here. The subset of the most recent 30-year period (1961-1990) is chosen. Instead of using all the stations in this

data set, we selected 1000 stations such a way that they cover the sphere as uniformly as possible. We first choose $\theta_1 = 10^{-0.1}$ and $\theta_2 = 10^{4.5}$ which correspond to the case that little smoothing is done to g_1 and g_2 , then choose θ_3 and θ_4 by a crude grid search according to a randomized GCV criterion (see Girard, 1989), which gave us a choice of $\theta_3 = 10^{1.25}$ and $\theta_4 = 10^{4.1}$. The estimated g_2 and $g_{\phi,2}$ are plotted in Figure 2 and Figure 3 respectively. Full details about this application may be found in Luo (1996).

6 Discussion

Penalty terms in (14) are crucial elements in deciding a smoothing spline estimate. The most important factors in those terms are the smoothing parameters (θ 's). They control the overall balance between the goodness of fit and the extra complexity measured by the penalty norms. Various ways of choosing these parameters including different subjective and objective ones are discussed in Luo (1996).

Even though smoothing parameters are of the most importance in deciding the features of SS estimates, they are not the only factors in those penalty terms that matter. The choices of those norms (J_α 's) are also important. These norms, induced usually by inner-products, correspond to the covariance functions of the Gaussian processes as priors we put on the component functions, if we adopt a Bayesian point of view. The features of these covariance functions such as their range and differentiability can also affect the SS estimates significantly. The R_s given in (18) has approximately the same range as the empirical correlogram does. However, we may also consider a parsimonious parametric family that can represent a wide range of practically different random fields, and use the data to formally select the parameters in this family as we did with smoothing parameters. For the temporal component, besides the penalty defined by that the 2-nd (or m-th) order differences are independent and identically distributed normal variables, which represents a class of short-range memory dependence, long-range memory dependence modeled through fractional differencing (Hosking 1981, Haslett and Raftery 1989) may also be used in our

model (see, Taam and Yandell, 1989). For the spatial component, the one we used (18) is in a family developed by Wahba (1981). Other families such as Matern's (1986) and Vecchia's (1985) may be used too. Both families are generalizations of Whittle's (1953).

When the spatial domain is large enough, we frequently face a problem of spatial inhomogeneity. Taking this into consideration may provide us more accurate estimates for the global analysis we intend to do. Spatial adaptive regression has been a topic discussed a lot in recent years. There are three basic types of approaches. One is to select (usually stationary) basis functions adaptively according to the data. The other two approaches are to use wavelet basis functions and to use a variable smoothing parameter across space (hence a function). A variable smoothing parameter is usually very difficult to estimate from data. It is closely related to the problem of estimating a nonhomogeneous covariance function (see Sampson and Guttorp, 1992 and Smith, 1996 for some recent studies). In Luo and Wahba (1997), we have proposed a method in the first category and obtained some promising results in univariate cases (though the variable can be more than one-dimensional). While wavelet bases on the sphere have not been well studied yet, it is straightforward to apply HAS on the sphere. We may use this approach to consider the nonhomogeneous problem in this spatial-temporal modeling setup. The first issue needs to be addressed is also a computational problem, since data sizes in this kind of applications are usually too large to be handled directly by the algorithm proposed in Luo and Wahba (1997). Again we may have to take advantage of the tensor-product structure of this kind of data. The question is how to blend techniques in section 1 and adaptive selecting algorithm together. Another issue is how to deal with different spatial-adaptation requirements in different terms of (6). We note that choosing basis functions can, to some degree, automatically play the role of smoothing parameters (θ 's). That is, components with more features tend to get more basis functions.

In many cases, confidence statements about the smoothing spline estimates are needed. When the computation for fitting a model is sufficiently fast, we may always consider bootstrapping. For SS estimates, there is another choice, i.e., Bayesian confidence intervals proposed by Wahba (1983) and further studied by Nychka (1988, 1990). It is derived through a Bayesian interpretation

of (14). Computing such intervals faces similar difficulties as fitting the model does. Parallel to the computational procedure proposed in Section 1, a Monte Carlo method, the Gibbs sampler combined with data augmentation (Tanner and Wong, 1987) just like the backfitting combined with iterative imputation, may be used to sample from the posterior distribution, hence to get confidence intervals needed. Facing the same difficulty of a slow convergence, we also need to explore various speeding-up techniques like the ones we proposed for fitting the model.

Acknowledgment. This work is part of my thesis done in the department of statistics of University of Wisconsin at Madison. I would like to thank my advisor, Professor Grace Wahba, for her kind advisory and encouragement which have been very important to me.

References

- Ansley, C.F. and Kohn, R. (1994), “Convergence of the backfitting algorithm for additive models”, *J. Austral. Math. Soc. (Series A)*, Vol. 57, 316-329.
- Buja, A., Hastie, T. and Tibshirani, R. (1989), “Linear smoothers and additive models”, (with discussions) *Ann. Stat.*, Vol. 17, No. 2, 453-555.
- Chen, Z., Gu, C. and Wahba, G. (1989), Discussion to “Linear smoothers and additive models” by Buja, Hastie and Tibshirani, *Ann. Stat.*, V. 17, 515-517.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), “Maximum Likelihood from Incomplete Data via the *EM* Algorithm”, *J. Royal Stat. Soc. Ser. B*, Vol. 39, 1-38.
- Girard, D.(1989), “A fast ‘Monte Carlo cross-validation’ procedure for large least squares problems with noisy data”, *Numer. Math.*, Vol. 56, 1-23.

- Golub, G.H. and Van Loan, C.F. (1989), *Matrix Computations*, 2nd Ed., The Johns Hopkins University Press, Baltimore.
- Green, P.J. (1990) “On use of the EM Algorithm for Penalized Likelihood Estimation”, *J. Royal Stat. Soc. Ser. B*, Vol. 52, 443-452.
- Gu, C. (1989), “RKPACK and its applications: Fitting smoothing spline models”, Technical Report No. 857, Department of Statistics, University of Wisconsin-Madison.
- Gu, C. and Wahba, G. (1993a), “Semiparametric analysis of variance with tensor product thin plate splines”, *J. Royal Statistical Soc. Ser. B*, Vol. 55, 353-368.
- (1993b), “Smoothing spline ANOVA with component-wise Bayesian confidence intervals”, *Journal of Computational and Graphical Statistics*, Vol. 2, 97-117.
- Hansen, J. and Lebedeff, S. (1987), “Global trends of measured surface air temperature”, *J. Geophysical Research*, Vol. 92, No. D11, pp. 13,345-13,372.
- Haslett, J. and Raftery, A.E. (1989). Space-time Modeling with Long-memory Dependence: Assessing Ireland’s Wind Power Resource (with Discussion). *Appl. Statist.* , **38**, 1-50.
- Hegerl, G.C., Storch, H.Von., Hasselmann, K., Santer, B.D., Cubasch, U. and Jones, P.D. (1995), “Detecting Greenhouse Gas Induced Climate Change with an Optimal Fingerprint Method”, *manuscript*.
- Hosking, J.R.M. (1981), “Fractional differencing”. *Biometrika*, V. 68, 165-176.
- Jones, P.D., Raper, S.C.B., Cherry, B.S.G., Goodess, C.M., Wigley, T.M.L., Santer, B., Kelly, P.M., Bradley, R.S. and Diaz, H.F. (1991), “An Updated Global Grid Point Surface Air Temper-

- ature Anomaly Data Set: 1851-1988”, Environmental Sciences Division Publication No. 3520, U.S. Department of Energy.
- Jones, P.D., Raper, S.C.B., Bradley, R.S., Diaz, H.F., Kelly, P.M. and Wigley, T.M.L.(1986), “Northern Hemisphere Surface Air Temperature Variations: 1851-1984”, *J. Climate and Applied Meteorology*, Vol. 25, 161-179.
- Karl, T.R., Knight, R.W. and Christy, J.R. (1994), “Global and Hemispheric Temperature Trends: Uncertainties Related to Inadequate Spatial Sampling”, *J. Climate*, Vol. 7, 1144-1163.
- Liu, J.S. (1994), “The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem”, *J. Amer. Stat. Assoc.*, Vol. 89, No. 427, 958-966.
- Luenberger, D.G. (1984), *Linear and Nonlinear Programming*, 2nd Ed., Addison-Wesley, Reading, Massachusetts.
- Luo, Z. (1996), “Backfitting in smoothing spline ANOVA with application to historical global temperature data” (thesis), Technical Report No. 964, Department of Statistics, University of Wisconsin at Madison.
- Luo, Z. and Wahba, G. (1997), “Hybrid Adaptive Splines”, *J. Amer. Stat. Assoc.*, to appear.
- Madden, R.A., Shea, D.J., Branstator, G.W., Tribbia, J.J. and Weber, R.O. (1993), “The Effects of Imperfect Spatial and Temporal Sampling on Estimates of the Global Mean Temperature: Experiments with Model Data”, *J. Climate*, Vol. 6, 1057-1066.
- Matern, B. (1986), *Spatial Variation*, Lecture Notes in Statistics, No. 36, Springer-Verlag, Berlin.
- Nychka, D. (1988), “Bayesian confidence intervals for smoothing splines”, *J. Amer. Statist. Assoc.*, Vol. 83, 1134-1143.

- Nychka, D. (1990), "The average posterior variance of a smoothing spline and a consistent estimate of the average squared error", *Ann. Stat.*, Vol. 18, 415-428.
- O'Sullivan, F. (1985), Discussion to "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting" by Silverman, *J. Royal. Stat. Soc.*, Ser. B, V. 47, 39-40.
- Sampson, P.D. and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure", *J. Amer. Statist. Assoc.*, Vol. 87, 108-119.
- Smith, R. L. (1996), "Estimating Nonstationary Spatial Correlations", *manuscript*.
- Stein, M. (1990), "Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure", *Ann. Stat.*, V. 18, 850-872.
- Taam, W. and Yandell, B.S. (1989), Discussion to the paper by Haslett and Raftery (1989), *Appl. Statist.*, Vol. 38, pp. 39-40.
- Tanner, M.A. and Wong, W.H. (1987), "The Calculation of posterior distributions by data augmentation", *J. Amer. Statist. Assoc.*, Vol. 82, 528-540.
- Varga, R.S (1962), *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Vecchia, A.V. (1985), "A general class of models for stationary two-dimensional random processes", *Biometrika*, V. 72, 281-291.
- Vinnikov, K.Ya., Groisman, P.Ya. and Lugina, K.M. (1990), "Empirical Data on Contemporary Global Climate Changes (Temperature and Precipitation)", *J. Climate*, Vol. 3, 662-677.
- Wahba, G. (1981), "Spline interpolation and smoothing on the sphere", *SIAM J. Sci. Stat. Comput.*, Vol.2, No.1, 5-16; Erratum (1982), Vol.3, No.3, 385-386.

Wahba, G. (1983), “Bayesian “confidence intervals” for the cross-validated smoothing spline”, *J. Roy. Stat. Soc. Ser. B*, Vol.45, No.1, 133-150.

———(1990), *Spline Models for Observational Data* (CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59), Philadelphia: Society of Industrial and Applied Mathematics.

Wahba, G. and Luo, Z. (1996), “Smoothing spline ANOVA fits for very large, nearly regular data sets, with application to historical global climate data”, Technical Report No. 952, Department of Statistics, University of Wisconsin at Madison.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995), “Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy”, *Annals of Statistics*, Vol. 23, No. 6, 1865-1895.

Wu, C.F.J. (1983), “On the convergence properties of the EM algorithm”, *Ann. Stat.*, Vol. 11, No. 1, 95-103.

Yates, F. (1933), “The analysis of replicated experiments when the field results are incomplete”, *The Empire J. of Experimental Agriculture*, Vol. 1, No. 2, 129-142.

Young, D.M. (1971), *Iterative Solution of Large Linear Systems*, Academic Press, New York.

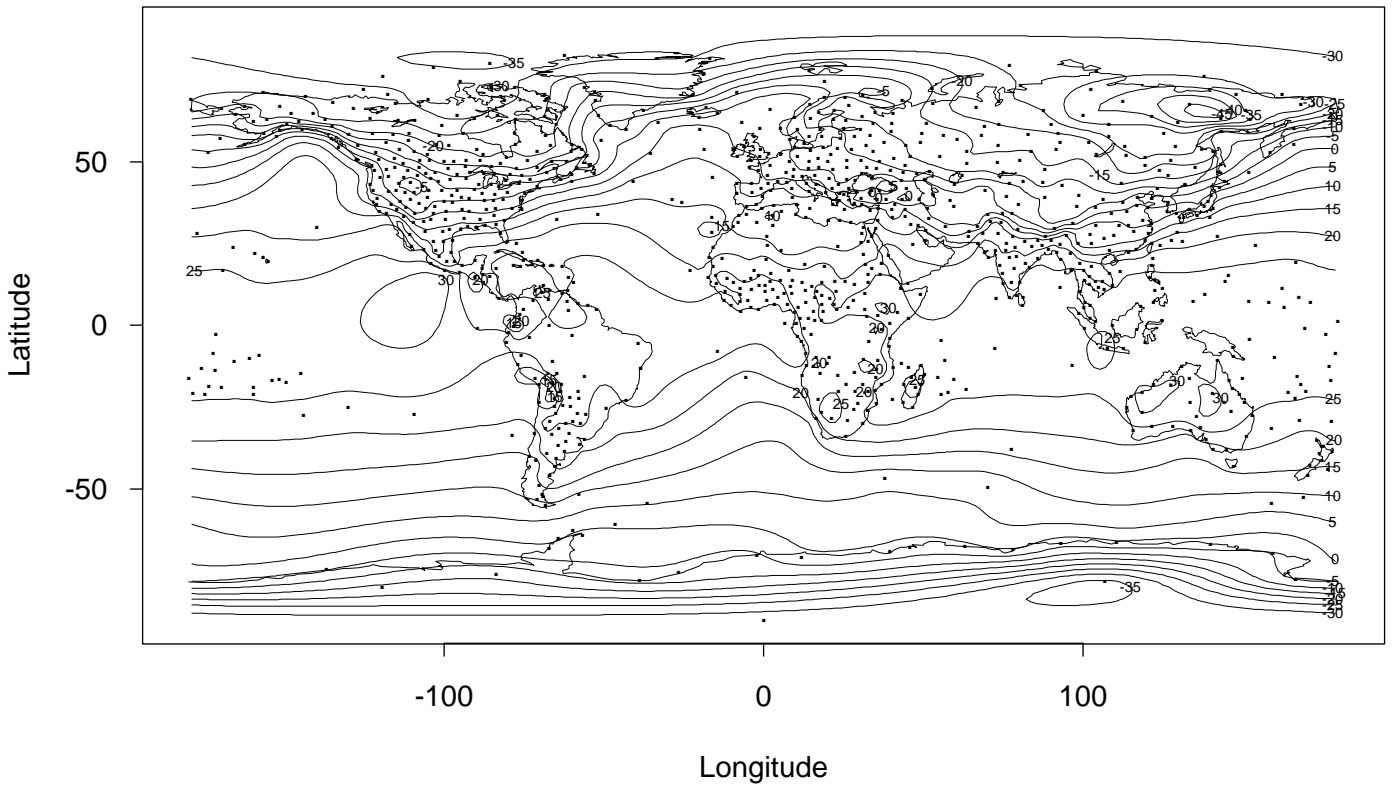


Figure 2: Winter mean temperature of 1961-1990 based on the smoothing spline estimate defined by (14).

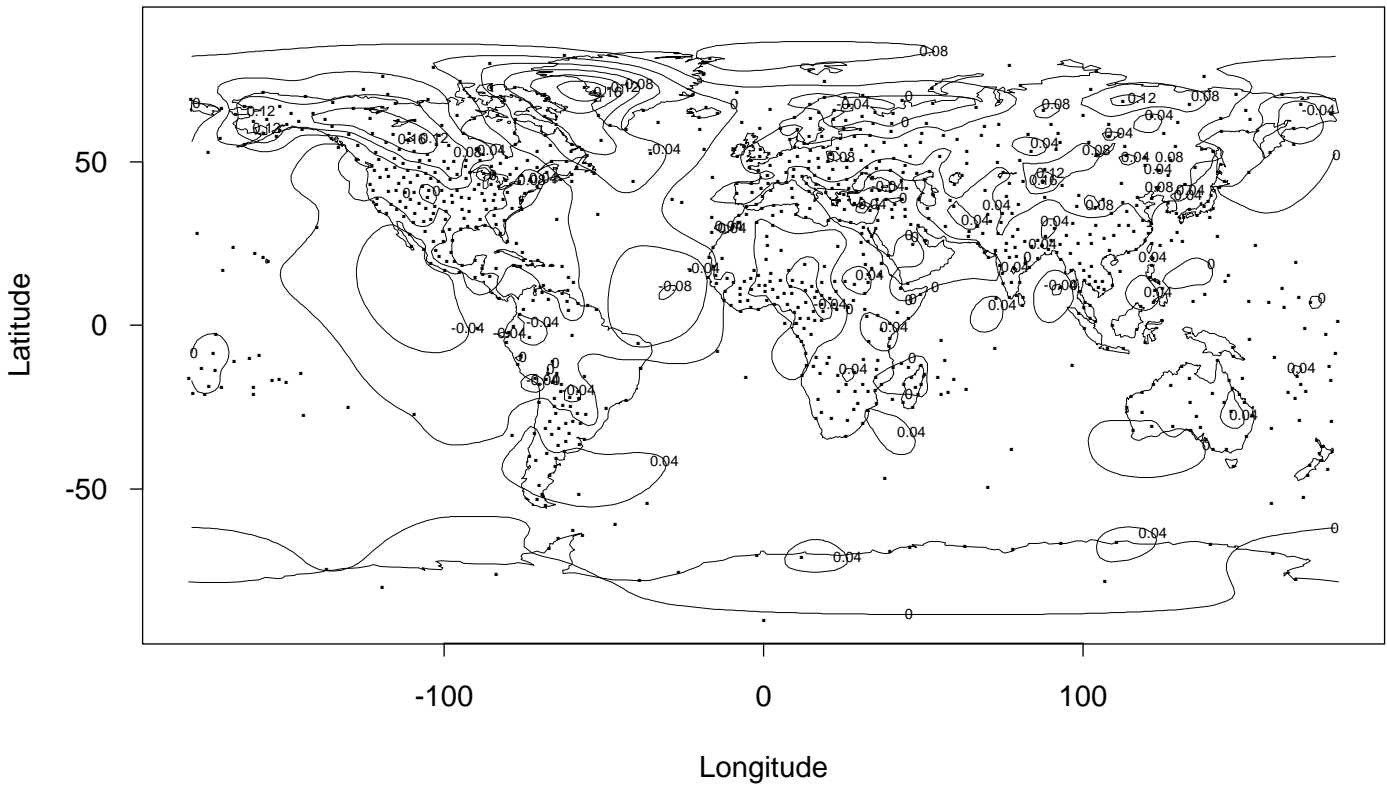


Figure 3: Linear trend coefficient of winter temperature during 1961-1990 based on the smoothing spline estimate defined by (14).