DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 952 (rev)

December 22, 1995

Smoothing Spline ANOVA Fits for Very Large, Nearly Regular Data Sets, with Application to Historical Global Climate Data [1]

by

Grace Wahba

and

Zhen Luo

[1]To appear in the Festschrift in Honor of Ted Rivlin, C. Micchelli, Ed., Baltzer Press, 1996. Research sponsored in part by NSF under Grant DMS 9121003 and NASA under Grant NAGW-2961

Smoothing Spline ANOVA Fits For Very Large, Nearly Regular Data Sets, With Application to Historical Global Climate Data †

Grace Wahba[1] and Zhen Luo[1]

[1]*Department of Statistics,*
University of Wisconsin
Madison, WI 53706
E-mail: wahba@stat.wisc.edu, zhen@stat.wisc.edu

We review smoothing spline analysis of variance (SS-ANOVA) methods for fitting a function of several variables to scattered, noisy data. The fitted function is obtained as a sum of functions of one variable (main effects) plus a sum of functions of two variables (two-factor interactions), and so forth. The terms are found as solutions to a variational problem in a reproducing kernel Hilbert space which is built up from tensor sums and products of Hilbert spaces of functions of fewer variables. These methods have found application in environmental and demographic data analysis problems, and provide an intuitively interpretable technique for examining responses as (smooth) functions of several variables. Matrix decomposition methods can be used to compute the SS-ANOVA fits while adaptively choosing multiple smoothing parameters by generalized cross validation (GCV), provided that matrix decompositions of size $n \times n$ can be carried out, where n is the sample size. We review the randomized trace technique and the backfitting algorithm, and remark that they can be combined to solve the variational problem while choosing the smoothing parameters by GCV for data sets that are much too large to use matrix decomposition methods directly. Some intermediate calculations to speed up the backfitting algorithm are given which are useful when the data has a tensor product structure. We describe an imputation procedure which can take advantage of data with a (nearly) tensor product structure. As an illustration of an application we discuss the algorithm in the context of fitting and smoothing historical global winter mean surface temperature data and examining the main effects and interactions for time and space.

Subject classification: AMS(MOS) 41A15, 62H11, 62G07, 65D07, 65D10, 65D15, 62M30, 65K10, 62F15,49J55.

Keywords: smoothing spline ANOVA, Gauss-Seidel algorithm, backfitting algo-

†This research in part by NSF under Grant DMS-9121003 and in part by NASA Grant NAGW-2961

rithm, randomized trace estimates, generalized cross validation, RKPACK, large environmental data sets.

1 The Smoothing Spline ANOVA Decomposition

Some of this section is reprised from [43]. Let $\mathcal{T}^{(\alpha)}$ be a measurable space, $t_\alpha \in \mathcal{T}^{(\alpha)}$, $\alpha = 1, \cdots, d$; $(t_1, \cdots, t_d) = t \in \mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$. For f satisfying some measurability conditions on $\mathcal{T}$ a unique ANOVA decomposition of f of the form

$$f(t_1, \cdots, t_d) = \mu + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots \tag{1}$$

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

$$(\mathcal{E}_\alpha f)(t) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \cdots, t_d) d\mu_\alpha(t_\alpha). \tag{2}$$

Then the identity is decomposed as

$$I = \prod_\alpha (\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) = \prod_\alpha \mathcal{E}_\alpha + \sum_\alpha (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta + \sum_{\alpha < \beta} (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma +$$
$$\cdots + \prod_\alpha (I - \mathcal{E}_\alpha). \tag{3}$$

The components of this decomposition generate the ANOVA decomposition of f of the form (1) by

$$\mu = (\prod_\alpha \mathcal{E}_\alpha) f, f_\alpha = ((I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta) f, f_{\alpha\beta} = ((I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma) f, \tag{4}$$

and so forth.

The idea behind SS-ANOVA is to construct a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions on $\mathcal{T}$ so that the components of the SS-ANOVA decomposition represent an orthogonal decomposition of f in $\mathcal{H}$, and, given noisy data, to estimate (some of) these components.

We suppose there are observations y_i generated according to the model

$$y_i = f(t_1(i), \cdots, t_d(i)) + \epsilon_i, \quad i = 1, \cdots, n, \tag{5}$$

where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2 I_{n \times n})$ is a 'Gaussian white noise' vector. Some (or all) of the components in the ANOVA decomposition of f will be estimated by finding f in (an appropriate subspace of) $\mathcal{H}$ to minimize

$$\sum_{i=1}^n (y_i - f(t(i)))^2 + \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots, \tag{6}$$

where the $J_\alpha, J_{\alpha\beta}$ are 'roughness penalties' and the series may be truncated at some point. This is done as follows: Let $\mathcal{H}^{(\alpha)}$ be an RKHS of (real valued) functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha = 0$ for $f_\alpha(\cdot) \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. Construct $\mathcal{H}$ as

$$\mathcal{H} = \prod_{\alpha=1}^{d} (\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\}) = [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots, \tag{7}$$

where $[1^{(\alpha)}]$ is the constant functions on $\mathcal{T}^{(\alpha)}$ and $[1]$ is the constant functions on $\mathcal{T}$. With some abuse of notation, factors of the form $[1^{(\alpha)}]$ are omitted whenever they multiply a term of a different form. Thus $\mathcal{H}^{(\alpha)}$ is a shorthand for $[1^{(1)}] \otimes \cdots \otimes [1^{(\alpha-1)}] \otimes \mathcal{H}^{(\alpha)} \otimes [1^{(\alpha+1)}] \otimes \cdots \otimes [1^{(d)}]$ (which is a subspace of $\mathcal{H}$). By letting the square norm on $[1^{(\alpha)}]$ be $(\int f d\mu_\alpha)^2$, and using the induced tensor product norm, the components of the ANOVA decomposition will be in mutually orthogonal subspaces of $\mathcal{H}$.

Next, $\mathcal{H}^{(\alpha)}$ is decomposed into a parametric part and a smooth part, by letting $\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$, where $\mathcal{H}_\pi^{(\alpha)}$ is finite dimensional (the "parametric" part) and $\mathcal{H}_s^{(\alpha)}$ (the "smooth" part) is the orthocomplement of $\mathcal{H}_\pi^{(\alpha)}$ in $\mathcal{H}^{(\alpha)}$. Elements of $\mathcal{H}_\pi^{(\alpha)}$ are not penalized through the device of letting $J_\alpha(f_\alpha) = \|P_s^{(\alpha)} f_\alpha\|^2$ where $P_s^{(\alpha)}$ is the orthogonal projector onto $\mathcal{H}_s^{(\alpha)}$. $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}]$ is now a direct sum of four orthogonal subspaces: $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] = [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}] \oplus [\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$. By convention the elements of the finite dimensional space $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}]$ will not be penalized. Continuing this way results in an orthogonal decomposition of $\mathcal{H}$ into sums of products of unpenalized finite dimensional subspaces, plus main effects 'smooth' subspaces, plus two factor interaction spaces of the form parametric $\otimes$ smooth $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$, smooth $\otimes$ parametric $[\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}]$ and smooth $\otimes$ smooth $[\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$ and similarly for the three and higher factor subspaces.

Now suppose that we have selected the model $\mathcal{M}$, that is, we have decided which subspaces will be included. Collect all of the included unpenalized subspaces into a subspace, call it $\mathcal{H}^0$, of dimension M, and relabel the other subspaces as $\mathcal{H}^\beta, \beta = 1, 2, \cdots, p$. $\mathcal{H}^\beta$ may stand for a subspace $\mathcal{H}_s^{(\alpha)}$, or one of the three subspaces in the decomposition of $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}]$ which contains at least one 'smooth' component, or, a higher order subspace with at least one 'smooth' component. Collecting these subspaces as $\mathcal{M} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$, the estimation problem becomes: Find f in $\mathcal{M} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$ to minimize

$$\sum_{i=1}^{n} (y_i - f(t(i)))^2 + \sum_{\beta=1}^{p} \theta_\beta^{-1} \|P^\beta f\|^2, \tag{8}$$

where P^β is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^\beta$, and $\theta_\beta^{-1} = \lambda_\beta$. The minimizer, call it f_λ ($\lambda = (\lambda_1, \cdots, \lambda_p)$) of (8) is known to have a representation [40], Chapter 10 in terms of a basis $\{\phi_\nu\}$ for $\mathcal{H}^0$ and the reproducing kernels (RK's) $\{R_\beta(s,t)\}$ for the $\mathcal{H}^\beta$. Letting

$$Q_\theta(s,t) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(s,t), \tag{9}$$

it is

$$f_\theta(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i Q_\theta(t(i),t) = \phi(t)'d + \xi(t)'c, \tag{10}$$

where

$$\begin{aligned} \phi'(t) &= (\phi_1(t), \cdots, \phi_M(t)), \\ \xi'(t) &= (Q_\theta(t(1),t), \cdots, Q_\theta(t(n),t)). \end{aligned}$$

$c_{n\times 1}$ and $d_{M\times 1}$ are vectors of coefficients which satisfy

$$\begin{aligned} (Q_\theta + I)c + Sd &= y \\ S'c &= 0 \end{aligned} \tag{11}$$

where here and below we are letting Q_θ be the $n \times n$ matrix with ijth entry $Q_\theta(t(i),t(j))$, and S be the $n \times M$ matrix with $i\nu$th entry $\phi_\nu(t(i))$. This system will have a unique solution for any set of positive $\{\lambda_\beta\}$ provided S is of full column rank, which we will always assume. This condition on S is equivalent to the uniqueness of least squares regression onto *span* $\{\phi_\nu\}$. Since the RK of a tensor product space is the product of the RK's of the component spaces, the computation of the R_β's is straightforward. For example, the RK corresponding to the subspace $\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}$ is (in an obvious notation), $R_{\mathcal{H}_\pi^{(\alpha)}}(s_\alpha, t_\alpha) R_{\mathcal{H}_s^{(\beta)}}(s_\beta, t_\beta)$. Of course any positive definite function may in principle play the role of a reproducing kernel here. Conditionally positive definite functions [32] as in thin plate splines [44] may also be used. The point evaluation functionals $f \to f(t(i))$ may be replaced by bounded linear functionals on $\mathcal{H}$, and other functions can be added to $\mathcal{H}^0$ subject just to the uniqueness conditions, making this class of function estimates broadly useful in many applications. See [4] [5] [6] [7] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24][25] [29] [33] [34] [35] [37] [38] [39] [40] [41] [44] .

2 Backfitting

Hastie and Tibshirani [26] Section 5.2.3, discuss the backfitting (a. k. a. Gauss-Seidel) algorithm in the context of the general setup of SS-ANOVA problems

as was described by [6]. Further discussion of the backfitting algorithm can be found in [3] and elsewhere. Referring to (8)-(11), let $f_0(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t)$ and let $f_\beta(t) = \sum_{i=1}^{n} c_i \theta_\beta R_\beta(t(i), t)$. Then $f_\lambda(\cdot) = f_0(\cdot) + \sum_{\beta=1}^{p} f_\beta(\cdot)$ with c and d satisfying (11) is the minimizer over f in $\mathcal{M}$ of

$$\sum_{i=1}^{n}(y_i - f(t(i)))^2 + \sum_{\beta=1}^{p} \lambda_\beta \|P^\beta f\|^2, \tag{12}$$

where we have set $\lambda_\beta = \theta_\beta^{-1}$. Now, define $\tilde{f}_0(\cdot) = f_0(\cdot)$ and $\tilde{f}_\beta(\cdot) = \sum_{i=1}^{n} c_{i\beta} R_\beta(t(i), \cdot)$, for arbitrary $c_{i\beta}$. In what follows it will be useful to recall that $\|\tilde{f}_\beta\|^2 \equiv \|P^\beta \tilde{f}\|^2 = c'_\beta R_\beta c_\beta$ where $c_\beta = (c_{1\beta} \cdots c_{n\beta})'$ and R_β is the $n \times n$ matrix with ijth entry $R_\beta(t(i), t(j))$, see [40]. In [26] the authors consider minimizing (12) in the span of all functions of the form $f(\cdot) = \tilde{f}_0(\cdot) + \tilde{f}_1(\cdot) + \cdots + \tilde{f}_p(\cdot)$ and they discuss finding d and $c_\beta, \beta = 1, \cdots, p$ to minimize

$$\|y - Sd - \sum_{\beta=1}^{p} R_\beta c_\beta\|^2 + \sum_{\beta=1}^{p} \lambda_\beta c'_\beta R_\beta c_\beta. \tag{13}$$

They note that the (vector) smooths corresponding to the minimizers, defined as $\tilde{\mathbf{f}}_0 \equiv S\tilde{d}$ and $\tilde{\mathbf{f}}_\beta \equiv R_\beta c_\beta, \beta = 1, \cdots, p$, satisfy the backfitting equations

$$\tilde{\mathbf{f}}_\gamma = S_\gamma(y - \sum_{\beta \neq \gamma} \tilde{\mathbf{f}}_\beta), \gamma = 0, 1, \cdots, p, \tag{14}$$

with the smoother matrix S_0 given by $S_0 = S(S'S)^{-1}S'$ and the other smoother matrices S_β given by $S_\beta = R_\beta(R_\beta + \lambda_\beta I)^{-1}, \beta = 1, \cdots p$. The backfitting algorithm solves for $\tilde{\mathbf{f}}_0$ and $\tilde{\mathbf{f}}_\beta, \beta = 1, \cdots, p$ by cycling through $\tilde{\mathbf{f}}_\gamma = S_\gamma(y - \sum_{\alpha \neq \gamma} \tilde{\mathbf{f}}_\alpha), \gamma = 0, 1, \cdots, p$. The backfitting algorithm is known to converge if the Frobenius norm of each product $S_\alpha S_\beta$ is less than 1.

Since the minimizer of (12) is in the $M + n$ dimensional space spanned by $\{\phi_\nu(\cdot), \nu = 1, \cdots, M\} \cup \{\sum_{\beta=1}^{p} \theta_\beta R_\beta(t(i), \cdot), i = 1, \cdots n\}$, minimizing (12) in the larger space spanned by the $M + np$ functions $\{\phi_\nu(\cdot), \nu = 1, \cdots, M\} \cup \{R_\beta(t(i), \cdot), \beta = 1, \cdots p; n = 1, \cdots, n\}$ will result in a solution in the $n + M$ dimensional space.

Setting $\tilde{\mathbf{f}}_\beta = R_\beta c_\beta$, the last p backfitting equations become

$$R_\beta(\lambda_\beta c_\beta + \sum_{\alpha=1}^{p} R_\alpha c_\alpha) = R_\beta(y - Sd), \ \beta = 1, \cdots, p, \tag{15}$$

although c_β will not be uniquely determined if R_β is not of full rank. However, suppose $\lambda_\alpha R_\alpha c_\alpha = R_\alpha c$ for some c, $\alpha = 1, \cdots, p$. Recalling that $\lambda_\alpha = \theta_\alpha^{-1}$, this would give

$$R_\beta(I + \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha)c = R_\beta(y - Sd), \beta = 1, \cdots, p, \tag{16}$$

so that if c satisfies $(I + \sum_{\beta=1}^{p} \theta_\beta R_\beta)c = y - Sd$, then $c_\beta = \theta_\beta c,\ \beta = 1, \cdots, p$ satisfies the backfitting equations. Thus, despite the apparently larger number $np + M$ of unknowns in (14) compared to the $n + M$ unknowns in (11), the backfitting solutions $\tilde{\mathbf{f}}_\gamma, \gamma = 0, 1, \cdots p$ are, at convergence, equivalent to solving

$$(Q_\theta + I)c + Sd = y \tag{17}$$

$$S'c = 0 \tag{18}$$

for c and d and setting $\tilde{\mathbf{f}}_0 = Sd, \tilde{\mathbf{f}}_\beta = \theta_\beta R_\beta c$. Equation (18) follows by observing that the first backfitting equation becomes $Sd = S_0(y - Q_\theta c)$. Substituting this into $(I + Q_\theta)c = (y - Sd)$ results in $c = (I - S_0)(y - Q_\theta c)$ which entails (18). Substituting back into (17) results in $c = y - \sum_{\beta=0}^{p} \tilde{\mathbf{f}}_\gamma$, which may then be used to compute $f_\beta(t)$ for general t via $f_\beta(t) = \sum_{i=1}^{n} c_i \theta_\beta R_\beta(t(i), t)$.

Ansley and Kohn [1] have a nice discussion of the backfitting algorithm in the context of von Neuman's alternating projection method.

3 Choosing the Smoothing Parameters

Probably the most popular method for choosing $\lambda = (\lambda_1, \cdots, \lambda_p) \equiv (\theta_1^{-1}, \cdots, \theta_p^{-1})$ is the GCV method,[7] [14] which chooses λ as the minimizer of

$$V(\lambda) = \frac{\|(I - A(\lambda))y\|^2}{[tr(I - A(\lambda))]^2} \tag{19}$$

where $A(\lambda)$ is the $n \times n$ matrix satisfying

$$A(\lambda)y = (f_\lambda(t(1)), \cdots f_\lambda(t(n)))'. \tag{20}$$

The matrix $I - A(\lambda)$ is known to have a representation

$$I - A(\lambda) = \Gamma_2(\Gamma_2'(\Sigma + I)\Gamma_2)^{-1}\Gamma_2' \tag{21}$$

where Γ_2 is any $n \times (n - M)$ orthogonal matrix satisfying $\Gamma_2' S = 0$ and Σ is the $n \times n$ matrix with i, jth entry $Q_\theta(t(i), t(j))$, see, for example [40], p. 13. $A(\lambda)$ is a so-called 'smoother' matrix, that is, a symmetric, non-negative definite matrix with all its eigenvalues in the interval $[0, 1]$. If the variance σ^2 in Equation (5) is given, then the unbiased risk estimate for λ is given by the minimizer of

$$U(\lambda) = \|(I - A(\lambda))y\|^2 + 2\sigma^2 trA(\lambda) \tag{22}$$

can also be used. See [14], [7]. Other estimates are discussed in [40].

The code RKPACK ([15], [20], [23]) is designed to compute $trA(\lambda)$ and compute and find the minimizer of $V(\lambda)$ and solve (11), using matrix decomposition methods.

Recently Girard [9] [10], [11] and Hutchinson [27] have proposed the randomized trace technique for estimating $trA(\lambda)$ for large n. This method is feasible for n larger than allowable with matrix decomposition methods, and can be used whenever a 'black box' is available for obtaining $\tilde{\mathbf{f}}$, the n-vector with ith entry $f_\lambda(t(i)), i = 1, \cdots, n$, given y. That is, it is assumed that some algorithm is available which produces (a good numerical approximation to) $\tilde{\mathbf{f}} = A(\lambda)y$ for any y. The the randomized trace estimate is based on the following fact: Let ξ be an n-dimensional (pseudo-)random vector with mean zero and covariance matrix I. Then the expected value of $\xi' A(\lambda)\xi = trA(\lambda)$. Furthermore, if ξ is a Gaussian random vector then the standard deviation of $\frac{1}{n}\xi' A(\lambda)\xi$ is $\sqrt{\frac{2}{n}}[\frac{1}{n}trA^2(\lambda)]^{1/2}$, [10]. Since $A(\lambda)$ is a smoother matrix then $\frac{1}{n}trA(\lambda) \in [0, 1]$ and the standard deviation of $\frac{1}{n}\xi' A(\lambda)\xi$ is no greater than $\sqrt{\frac{2}{n}}[\frac{1}{n}trA(\lambda)]^{1/2}$. In practice it is preferable to estimate $\frac{1}{n}tr(I - A(\lambda))$. Letting $\tilde{\mathbf{f}}(\xi)$ be $\tilde{\mathbf{f}}$ with the data vector y replaced by ξ, then the estimate of $\frac{1}{n}tr(I - A(\lambda))$ is $\frac{1}{n}\xi'[\xi - \tilde{\mathbf{f}}(\xi)]$. The same ξ should be used for all values of λ. This results in an estimate for $V(\lambda)$ which is a smooth function of λ and appears to have the same shape as $V(\lambda)$ computed exactly. Excellent results with n around 600 were reported in [42], for example, and have been also reported by other authors, see, for example [13]. Since the (converged) backfitting algorithm produces $\tilde{\mathbf{f}} \equiv \sum_{\beta=0}^{p} \tilde{\mathbf{f}}_\beta$ it can be used to compute the randomized trace estimate of $tr(I - A(\lambda))$ for selected values of λ.

Now, let $z_\gamma = y - \sum_{\beta \neq \gamma} \tilde{\mathbf{f}}_\beta$. Note that at convergence, when (14) is satisfied,

$$\|y - \tilde{\mathbf{f}}\|^2 = \|(y - \sum_{\beta \neq \gamma} \tilde{\mathbf{f}}_\beta) - \tilde{\mathbf{f}}_\gamma\|^2 \tag{23}$$

$$= \|z_\gamma - S_\gamma z_\gamma\|^2, \gamma = 0, 1, \cdots, p. \tag{24}$$

This suggests that $\lambda_\beta, \beta = 1, \cdots, p$ can be updated at each step as the backfitting proceeds, by considering λ_β to be fixed for $\beta \neq \gamma$. Let $A(\lambda_\gamma)$ stand for $A(\lambda)$ with all the λ_β considered fixed except λ_γ and choose λ_γ to minimize

$$V(\lambda_\gamma) = \frac{\|z_\gamma - S_\gamma z_\gamma\|^2}{tr(I - A(\lambda_\gamma))}, \gamma = 1, \cdots, p, \tag{25}$$

similarly for $U(\lambda_\gamma)$. This is the BRUTO algorithm in [26], p 262.

4 Global Climate Data

The algorithm we describe is well suited to the analysis of certain kinds of global environmental data. Rather than give a completely general description, we will motivate the discussion with respect to a particular example, of broad general interest. Generalizations to more complicated models will be fairly evident.

We have in mind monthly average surface temperature data, that has been computed from daily observations of surface temperature that have been collected at a large number of meteorological stations around the globe for varying periods of time. One of the oldest observing stations, at Trondheim, Norway, has been collecting such data since 1761. Records from around 1700 stations are available for the period 1961-90, with varying numbers of missing data. Just to give some concrete numbers, as an illustration consider data for the $n_1 = 30$ years 1961-1990, for, say $n_2 = 1500$, stations that have *mostly* complete records for that period. Considering the occasional missing data point there will then be (for a particular month) somewhat fewer than $n = n_1 \times n_2 = 45,000$ observations, indexed (after scaling) by $t_1 \in \{t_1(1), \cdots, t_1(n_1)\} \in \mathcal{T}^{(1)} \equiv [0,1]$, and $t_2 \in \mathcal{S}$, where $\mathcal{S}$ is the unit sphere. (t_2 takes on the values of $(latitude, longitude)$ of the stations).

For expository purposes and generality we let $\mu_\alpha, \alpha = 1, 2$ be Lebesgue measure on the unit interval and on the sphere, respectively. In this example, if the $t_1(j)$ are equally spaced, $\mathcal{T}^{(1)}$ can (possibly more naturally) be taken as a set of n_1 points, with μ_1 assigning mass $\frac{1}{n_1}$ to each point. In this problem, the choice of Lebesgue measure on the sphere is a natural one, since we will be interested in estimating the global average temperature. We remark parenthetically that although from a mathematical point of view this choice appears trivially obvious, among climatologists who have to deal with very irregular data of this type the choice of a commonly acceptable working definition of 'global average surface temperature' is far from obvious.

If $\mathcal{T}^{(1)} = [0,1]$, we take for $\mathcal{H}^{(1)}$ the subspace of $W_2^m = \{f, f' abs.cont., f'' \in \mathcal{L}_2[0,1]\}$ of functions which satisfy $\int_0^1 f d\mu_1 = 0$. If we endow this space with the square norm $\|f\|^2_{\mathcal{H}^{(2)}} = (f(1) - f(0))^2 + \int_0^1 (f''(u))^2 du$ then we can let $\mathcal{H}^{(1)}_\pi$ be the one dimensional space spanned by the 'trend' function $\phi(u) = (u - \frac{1}{2})/12$. Then $\mathcal{H}^{(1)}_s$ can be taken as the subspace of W_2^m satisfying $\int_0^1 f(u)du = (f(1) - f(0)) = 0$ with the square norm $\|f\|^2_{\mathcal{H}^{(1)}_s} = \int_0^1 (f''(u))^2 du$. The RK $\tilde{R}_1(u,v)$ for $\mathcal{H}^{(1)}_s$ is given in [40] as

$$\tilde{R}_1(u,u') = k_2(u)k_2(u') - k_4([u-u']) \tag{26}$$

where $k_\ell(u) = B_\ell(u)/\ell!$ with B_ℓ the ℓth Bernoulli polynomial. Elements in $\mathcal{H}^{(1)}_s$ have a natural interpretation as having been 'detrended'. If we take $\mathcal{T}^{(1)} = \{1, \cdots, n_1\}$, and let $J(f) = \sum_{i=1}^{n-2} [f(i+2) - 2f(i+1) - f(i)]^2$, then $\mathcal{H}^{(1)}$ is the subspace of E^{n_1} of vectors whose components sum to 0, the trend function is $\phi(u) = u - (n_1+1)/2$, and $\mathcal{H}^{(1)}_s$ is the $n_1 - 2$ dimensional subspace of E^{n_1} of vectors perpendicular, in the Euclidean inner product, to the constant function and the trend function. The RK for $\mathcal{H}^{(1)}_s$ can easily be shown to be the $n_1 \times n_1$ matrix obtained as follows: Let L be the $(n_1 - 2) \times n_1$ matrix with 1 down the diagonal, -2 down the super-diagonal and 1 down the next super-diagonal, i. e. $J(f) = f'L'Lf$, where $f = (f(1), \cdots f(n_1))'$. Then $\tilde{R}(j,j')$ is the jj'th entry of

$(L'L)^\dagger$ where $\dagger$ is the Moore-Penrose generalized inverse,

Splines on the sphere have been discussed in [34],[35] where $J(f) = \int_{P\in\mathcal{S}}(\Delta^{m/2}f)^2 dP$ with Δ the surface Laplacian on the sphere. Closed form expressions for RK's with a norm which is (topologically) equivalent to $J(\cdot)$ are given there for $m = 3/2, 2, 5/2, ..., 6$. The parameter m may be estimated from the data by minimizing $min_\lambda V(\lambda)$ considered as a function of m, see [44]. Other positive definite functions on the sphere may be found in [45] and references cited there. See also [8] and [36] who obtain RK's from historical meteorological data. Among sufficiently differentiable functions on the sphere for which $J(f)$ is finite, the null space of $J(f)$ is just the constant functions, so we take $\mathcal{H}_\pi^{(2)}$ as empty and $\mathcal{H}^{(2)}$ as $\mathcal{H}_s^{(2)}$. In the example below we will choose the $m = 2$ case which (from [34]) gives the following RK $\tilde{R}_2$ for $\mathcal{H}^{(2)}$: Letting P, P' be two points on the sphere, and $z = \cos\gamma(P, P')$, where $\gamma(P, P')$ is the angle between P and P'

$$\tilde{R}_2(P, P') = \frac{1}{2\pi}\left[\frac{1}{2}q_2(z) - \frac{1}{6}\right] \tag{27}$$

where

$$q_2(z) = \frac{1}{2}\left\{ln(1+\sqrt{\left(\frac{2}{1-z}\right)}[12\left(\frac{1-z}{2}\right)^2 - 4\left(\frac{1-z}{2}\right)] - 12\left(\frac{1-z}{2}\right)^{3/2} + 6\left(\frac{1-z}{2}\right) + 1\right\}. \tag{28}$$

In the remainder of this section we will relable $t = (t_1, t_2)$ as $t = (x, P)$. Our RKHS of historical global temperature functions is $\mathcal{H} = [[1^{(1)}] \oplus [\phi] \oplus \mathcal{H}_s^{(1)}] \otimes [[1^{(2)}] \oplus \mathcal{H}_s^{(2)}]$, a collection of functions $f(x, P)$, on $[0, 1] \otimes \mathcal{S}$, where $\mathcal{H}$ and f have corresponding decompositions given below:

$\mathcal{H}$	=	$[1]$	$\oplus$	$[\phi]$	$\oplus$	$[\mathcal{H}_s^{(1)}]$	$\oplus$	$[\mathcal{H}_s^{(2)}]$	$\oplus$	$[[\phi]\otimes\mathcal{H}_s^{(2)}]$	$\oplus$	$[\mathcal{H}_s^{(1)}\otimes\mathcal{H}_s^{(2)}]$
$f(x,P)$	=	C	+	$d\phi(x)$	+	$f_1(x)$	+	$f_2(P)$	+	$\phi(x)f_{\phi,2}(P)$	+	$f_{12}(x,P)$
	=	*mean*	+	*global time trend*	+	*time main effect*	+	*space main effect*	+	*trend by space effect*	+	*space–time interaction*

For the $\mathcal{T}^{(1)} = [0, 1]$ case, the components of the ANOVA decomposition satisfiy the side conditions

$$\begin{aligned} 0 &= \int_0^1 \phi(x)dx \\ 0 &= \int_0^1 f_1(x)dx = (f_1(1) - f_1(0)) = \int_0^1 f_{12}(x, P)dx = f_{12}(1, P) - f_{12}(0, P) \\ 0 &= \int_{\mathcal{S}} f_2(P)dP = \int_{\mathcal{S}} f_{\phi,2}(P)dP = \int_{\mathcal{S}} f_{12}(x, P)dP, \end{aligned}$$

the equalities involving f_{12} holding for all P and for all x. Here $M = 2$ and $\mathcal{H}_0$ is the span of the two functions $\phi_0(x, P) \equiv 1$ and $\phi_1(x, P) \equiv \phi(x)$. For the

$\mathcal{T}^{(1)} = \{1, \cdots, n_1\}$ case the integral over x is replaced by the sum over $1, \cdots n_1$. In this case,

$$\sum_{j'=1}^{n_1} \phi(j') \equiv 0, \ \sum_{j'=1}^{n_1} \tilde{R}_1(j, j') \equiv 0, \ \sum_{j'=1}^{n_1} \tilde{R}_1(j, j')\phi(j') \equiv 0, \ j = 1, \cdots, n_1. \quad (29)$$

This will lead to some algorithmic simplifications to be described, in the regular data case. There are $p = 4$ subspaces whose components will be penalized, with RK's given below:

β	*space*	*RK*		
1	$\mathcal{H}_s^{(1)}$	$R_1(x, P; x', P')$	=	$\tilde{R}_1(x, x')$
2	$\mathcal{H}_s^{(2)}$	$R_2(x, P; x', P')$	=	$\tilde{R}_2(P, P')$
3	$[\phi] \otimes \mathcal{H}_s^{(2)}$	$R_3(x, P; x', P')$	=	$\phi(x)\phi(x')\tilde{R}_2(P, P')$
4	$\mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)}$	$R_4(x, P; x', P')$	=	$\tilde{R}_1(x, x')\tilde{R}_2(P, P')$

5 Matrix decompositions with backfitting for regular data

In this section we describe an approach combining backfitting and matrix decompositions of size $n_1 \times n_1$ and $n_2 \times n_2$ which can be used for the global climate data and other examples when the data are perfectly regular, that is, there is an observation for each pair $(x_j, P_k), j = 1, \cdots, n_1, k = 1, \cdots, n_2$, where j indexes time and k indexes station. (Note: this does not mean that the data is regular on the sphere, just that each station is reporting at each time.) In the subsequent section we will show how data imputation may (safely) be used when there are (a small number of) missing data points. The results in this and the next section will appear in [31].

The results will become clear after we establish some notation. As before, let

$$z_\gamma \equiv (y - \sum_{\beta \neq \gamma} \tilde{\mathbf{f}}_\beta). \quad (30)$$

here the vectors in (30) all are of dimension $n = n_1 \times n_2$, unless otherwise noted, we consider them partitioned into n_1 blocks of dimension n_2 and let z be a generic vector of this form, with $z(j, k)$ be the kth entry in the jth block of z. Let $P_o^{(1)} z$ be the n_1-vector with $\frac{1}{n_2} \sum_{k=1}^{n_2} z(j, k)$ in the j position, $j = 1, \cdots, n_1$ and $P_o^{(2)} z$ be the n_2-vector with $\frac{1}{n_1} \sum_{j=1}^{n_1} z(j, k)$ in the k th position, $k = 1, \cdots, n_2$. Finally, let z^j be the n_2 vector in the jth block of z. In the first backfitting equation $\gamma = 0$ in (14), $\tilde{\mathbf{f}}_0$ is the result of least squares regression of z_0 onto the columns of S. To solve the other backfitting equations, write the remaining equations in (14) as

$$(R_\beta + \lambda_\beta I)\tilde{\mathbf{f}}_\beta = R_\beta z_\beta, \quad \beta = 1, \cdots p. \quad (31)$$

Let $\tilde{R}_1$ be the $n_1 \times n_1$ matrix with jj'th entry $\tilde{R}_1(x_j, x_{j'}), j, j' = 1, \cdots, n_1$, and $\tilde{R}_2$ the $n_2 \times n_2$ matrix with kk'th entry $\tilde{R}_2(P_k, P_{k'}), k, k' = 1, \cdots, n_2$. Then, examining the jth block of (14) for $\beta = 2$ gives

$$\tilde{R}_2 P_o^{(2)} \tilde{\mathbf{f}}_2 + \frac{\lambda_2}{n_1} \tilde{\mathbf{f}}_2^j = \tilde{R}_2 P_o^{(2)} z_2, \quad j = 1, \cdots n_1, \tag{32}$$

which entails that $\tilde{\mathbf{f}}_2^1 = \cdots = \tilde{\mathbf{f}}_2^{n_1} = P_o^{(2)} \tilde{\mathbf{f}}_2$, say. Defining the marginal smoother matrix $\tilde{S}_2(\lambda) = \tilde{R}_2(\tilde{R}_2 + \lambda I)^{-1}$ results in

$$P_o^{(2)} \tilde{\mathbf{f}}_2 = \tilde{S}_2(\lambda_2/n_1) P_o^{(2)} z_2. \tag{33}$$

A similar argument, interchanging the roles of j and k gives, for $\beta = 1$,

$$P_o^{(1)} \tilde{\mathbf{f}}_1 = \tilde{S}_1(\lambda_1/n_2) P_o^{(1)} z_1, \tag{34}$$

where $\tilde{S}_1(\lambda) = \tilde{R}_1(\tilde{R}_1 + \lambda I)^{-1}$

For $\beta = 3$ the jth block of (14) becomes

$$\sum_{j'=1}^{n_1} \phi(x_j)\phi(x_{j'}) \tilde{R}_2 \tilde{\mathbf{f}}_3^{j'} + \lambda_3 \tilde{\mathbf{f}}_3^j = \sum_{j'=1}^{n_1} \phi(x_j)\phi(x_{j'}) \tilde{R}_2 z_3^{j'}, \quad j = 1, \cdots n_1. \tag{35}$$

It can be seen that $\tilde{\mathbf{f}}_3^j = \phi(x_j)v, j = 1, \cdots, n_1$, for some n_2 vector v. Letting $P_1^{(2)} z$ be the n_2 vector with $\sum_{j=1}^{n_1} \phi(x_j) z(j,k)$ in the kth position, and substituting this into (35) results in

$$\phi(x_j) \sum_{j'=1}^{n_1} \phi^2(x_{j'}) \tilde{R}_2 v + \lambda_3 \phi(x_j) v = \phi(x_j) \tilde{R}_2 P_1^{(2)} z_3. \tag{36}$$

Therefore

$$(\sum_{j'=1}^{n_1} \phi^2(x_{j'}) \tilde{R}_2 + \lambda_3 I) v = \tilde{R}_2 P_1^{(2)} z_3, \tag{37}$$

which gives

$$v = (\sum_{j'=1}^{n_1} \phi^2(x_{j'}) \tilde{R}_2 + \lambda_3 I)^{-1} \tilde{R}_2 P_1^{(2)} z_3. \tag{38}$$

Letting $\sum_{j=1}^{n_1} \phi^2(x_j) = \phi^2$ finally gives

$$\tilde{\mathbf{f}}_3^j = (\phi(x_j)/\phi^2) \tilde{S}_2(\lambda_3/\phi^2) P_1^{(2)} z_3. \tag{39}$$

Considering $\beta = 4$,

$$\tilde{\mathbf{f}}_4 = S_4(\lambda_4) z_4, \tag{40}$$

where $S_4(\lambda) = R_4(R_4 + \lambda I)^{-1}$ with $R_4 = \tilde{R}_1 \otimes \tilde{R}_2$. The $n_1 \times n_1$ and $n_2 \times n_2$ smoother matrices $\tilde{S}_1(\lambda)$ and $\tilde{S}_2(\lambda)$ appear repeatedly with varying values of λ. The approach we have taken in ongoing work for max (n_1, n_2) not too large is to calculate the eigenvalue-eigenvector decompositions of $\tilde{R}_1$ and $\tilde{R}_2$. Letting $\tilde{R}_\alpha = \Gamma D \Gamma'$, $\tilde{S}_\alpha(\lambda) = \Gamma D(D + \lambda I)^{-1}\Gamma$, $\tilde{S}_\alpha(\lambda)$ can be computed for varying values of λ. The eigenvalues and eigenvectors of R_4 are obtained as the tensor products of the eigenvalues and eigenvectors of $\tilde{R}_1$ and $\tilde{R}_2$ and $S_4(\lambda)$ can then be computed similarly.

In general if some components can be combined to reduce p then the number of backfitting iterations required is likely to be smaller. In the present example, set $\tilde{\mathbf{f}}_{1+4} = \tilde{\mathbf{f}}_1 + \tilde{\mathbf{f}}_4$ and $z_{1+4} = y - (\tilde{\mathbf{f}}_2 + \tilde{\mathbf{f}}_3)$ and let $S_{1+4}(\lambda_1, \lambda_4) = R_{1+4}(R_{1+4} + I)^{-1}$, where $R_{1+4} = \theta_1 R_1 + \theta_4 R_4 = \tilde{R}_1 \otimes [\theta_1 11' + \theta_4 \tilde{R}_2]$. Then the $\beta = 1$ and $\beta = 4$ backfitting steps can be replaced by the $1 + 4$ step, $\tilde{\mathbf{f}}_{1+4} = S_{1+4}(\lambda_1, \lambda_4) z_{1+4}$. This may require repeated matrix decompositions of, say $[\frac{\theta_1}{\theta_4} 11' + \tilde{R}_2]$, say, as θ_1, θ_2 are varied but may still represent a speedup.

The speed of convergence of the backfitting algorithm generally depends on the magnitudes of the product matrices $S_\alpha S_\beta$, becoming faster as these products become 'smaller'. If the products were all 0 for $\alpha \neq \beta$, then the backfitting iteration would converge in one step. In the example with $\mathcal{T}^{(1)} = \{1, \cdots, n_1\}$, $\tilde{R}_1 = (L'L)^\dagger$, the conditions (29) lead to all of these products equal 0 except $S_0 S_2, S_0 S_3$ and $S_1 S_4$. In general the sizes of these product matrices have a dependency on the magnitute of the λ_β, decreasing as the λ_β increase. The most efficient method for computing the $\tilde{\mathbf{f}}_\gamma$ for very large, perfectly regular data sets under various circumstances is under study. See [2] [12], Chapter 10.

6 Missing Data Imputation

In practice perfectly regular observational data, at least for climate data, is the exception rather than the rule. Unfortunately, a few data points missing from a regular set $\{x_j, P_k\}, j = 1, \cdots, n_1, k = 1, \cdots, n_2$ means that both the outer and inner loop backfitting equations would not all involve the same smoother matrices $\tilde{S}_1, \tilde{S}_2$, and the ability to use a common eigenvalue- eigenvector decomposition appears to be lost. We show how to get around this with an 'imputation' loop. To demonstrate that the imputation loop is legitimate we first need a slight variation of the leaving-out-one lemma in Craven and Wahba [7].

Lemma .1 The Leaving-Out-K Lemma
Let $\mathcal{H}$ be an RKHS with subspace $\mathcal{H}^0$ of dimension M as before, and for $f \in \mathcal{H}$ let $\|Pf\|^2 = \sum_{\beta=1}^p \theta_\beta^{-1} \|P^\beta f\|^2$. Let $f^{[K]}$ be the solution to the variational problem:

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

$$\sum_{\substack{i=1 \\ i \notin \mathcal{S}_K}}^{n} (y_i - f(t(i)))^2 + \|Pf\|^2, \tag{41}$$

where $\mathcal{S}_K = \{i_1, \cdots, i_K\}$ is a subset of $1, \cdots, n$ with the property that (41) has a unique minimizer, and let $y_i^*, i \in \mathcal{S}_K$ be 'imputed' values for the 'missing' data imputed as $y_i^* = f^{[K]}(t(i)), i \in \mathcal{S}_K$. Then the solution to the problem: Find $f \in \mathcal{H}$ to minimize

$$\sum_{\substack{i=1 \\ i \notin \mathcal{S}_K}}^{n} (y_i - f(t(i))^2 + \sum_{i \in \mathcal{S}_K} (y_i^* - f(t(i)))^2 + \|Pf\|^2 \tag{42}$$

is $f^{[K]}$.

Proof
Let $h = f^{[K]}$ and let f be any element in $\mathcal{H} \neq f^{[K]}$. Then:

$$\begin{aligned}
\sum_{\substack{i=1 \\ i \notin \mathcal{S}_K}}^{n} (y_i - h(t(i))^2 + \sum_{i \in \mathcal{S}_K} (y_i^* - h(t(i)))^2 \quad &+ \quad \|Ph\|^2 \\
&= \sum_{i \notin \mathcal{S}_K} (y_i - f^{[K]}(t(i)))^2 + \|Pf^{[K]}\|^2 \\
&< \sum_{i \notin \mathcal{S}_K} (y_i - f(t(i)))^2 + \|Pf\|^2 \\
&\leq \sum_{i \notin \mathcal{S}_K} (y_i - f(t(i)))^2 + \sum_{i \in \mathcal{S}_K} (y_i^* - f(t(i)))^2 + \|Pf\|^2
\end{aligned}$$

Thus, $h = f^{[K]}$ is the minimizer of (42). □

Let y be partitioned as

$$y = \begin{pmatrix} y^{(1)} \\ \cdots \\ y^{(2)} \end{pmatrix} \tag{43}$$

where the entries have been relabeled so that $y^{(2)} = (y_{i_1}, \cdots, y_{i_K})' \equiv (y_{n-K+1}, \cdots y_n)'$, and let $A(\lambda)$ be defined as before by $\tilde{\mathbf{f}} = A(\lambda)y$. Let $A(\lambda)$ be partitioned corresponding to (43) as

$$A(\lambda) = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}. \tag{44}$$

Then, by the Leaving-Out-K Lemma,

$$\begin{pmatrix} f^{[K]}(t(i_1)) \\ \vdots \\ f^{[K]}(t(i_K)) \end{pmatrix} = A_{21} y^{(1)} + A_{22} \begin{pmatrix} f^{[K]}(t(i_1)) \\ \vdots \\ f^{[K]}(t(i_K)) \end{pmatrix}, \tag{45}$$

and, if furthermore $(I - A_{22}) \succ 0$, then

$$\begin{pmatrix} f^{[K]}(t(i_1)) \\ \vdots \\ f^{[K]}(t(i_K)) \end{pmatrix} = (I - A_{22})^{-1} A_{21} y^{(1)}. \tag{46}$$

There is an easy necessary and sufficient condition for $(I - A_{22}) \succ 0$.

Lemma .2 The Pre-Imputation Lemma
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 (43) as

$$\begin{pmatrix} \Gamma_{11} \\ \cdots \\ \Gamma_{21} \end{pmatrix}. \tag{47}$$

Then $(I - A_{22}) \succ 0$ if and only if 1 is not an eigenvalue of $\Gamma_{21}\Gamma_{21}'$.

Proof
Let Γ_2 be the $n \times n - M$ matrix in (21) and let Γ be

$$\Gamma = \left(\Gamma_1 \vdots \Gamma_2\right) = \begin{pmatrix} \Gamma_{11} & \vdots & \Gamma_{12} \\ \cdots & & \cdots \\ \Gamma_{21} & \vdots & \Gamma_{22} \end{pmatrix}, \tag{48}$$

therefore, from (21), $I - A_{22} = \Gamma_{22}(\Gamma_2'(\Sigma + I)\Gamma_2)^{-1}\Gamma_{22}'$, with $\Gamma\Gamma' = I_{n\times n}$ and $\Gamma_{21}\Gamma_{21}' + \Gamma_{22}\Gamma_{22}' = I_{K\times K}$. Let u be any K-vector, we have $u'\Gamma_{21}\Gamma_{21}'u + u'\Gamma_{22}\Gamma_{22}'u = u'u$. Thus u an eigenvector of $\Gamma_{21}\Gamma_{21}'$ with eigenvalue 1 guarantees that $u'\Gamma_{22}\Gamma_{22}'u = 0$ and so Γ_{22}' cannot be of full column rank, and hence $(I - A_{22})$ cannot be of full rank, conversely, if 1 is not an eigenvalue of $\Gamma_{21}\Gamma_{21}'$, then $\Gamma_{22}\Gamma_{22}'$ is strictly positive definite, ensuring that Γ_{22}' is of full column rank and hence $(I - A_{22}) \succ 0$. □

We have

Lemma .3 The Imputation Lemma
Let $g_{(o)}^{(2)}$ be a K-vector of initial values for an imputation of $(f^{[K]}(t(i_1)), \cdots f^{[K]}(t(i_K))'$, and suppose $0 \prec (I - A_{22})$. Let successive imputations $g_{(\ell)}^{(2)}$ for $\ell = 1, 2, \cdots$ be obtained via

$$\begin{pmatrix} g_{(\ell)}^1 \\ \cdots \\ g_{(\ell)}^2 \end{pmatrix} = A(\lambda) \begin{pmatrix} y^1 \\ \cdots \\ g_{(\ell-1)}^2 \end{pmatrix}. \tag{49}$$

Then

$$\lim_{\ell \to \infty} \begin{pmatrix} g_{(\ell)}^{(1)} \\ \cdots \\ g_{(\ell)}^{(2)} \end{pmatrix} = \begin{pmatrix} f^{[K]}(t(1)) \\ \cdots \\ f^{[K]}(t(n)) \end{pmatrix}. \tag{50}$$

Proof
By the Leaving-Out-K Lemma,

$$\begin{pmatrix} f^{[K]}(t(1)) \\ \vdots \\ f^{[K]}(t(n)) \end{pmatrix} = A(\lambda) \begin{pmatrix} y^{(1)} \\ \cdots \\ f^{[K]}(t(i_1) \\ \vdots \\ f^{[K]}(t(i_K)) \end{pmatrix}, \tag{51}$$

so we only need to show that

$$\lim_{\ell \to \infty} g_{(\ell)}^{(2)} = \begin{pmatrix} f^{[K]}(t(i_1) \\ \vdots \\ f^{[K]}(t(i_K)) \end{pmatrix}. \tag{52}$$

But

$$\begin{aligned} g_{(\ell)}^{(2)} &= A_{21} y^{(1)} + A_{22}[A_{21} y^{(1)} + A_{22} g_{(\ell-1)}^{(2)}] && (53) \\ &= \cdots && (54) \\ &= (I + A_{22} + \cdots + A_{22}^{\ell-1}) A_{21} y^{(1)} + A_{22}^{\ell} g_{(o)}^{(2)}. && (55) \end{aligned}$$

so that assuming $0 \prec (I - A_{22})$ gives

$$g_{(\ell)}^{(2)} \to (I - A_{22})^{-1} A_{21} y^{(1)}, \tag{56}$$

and the result follows. □

We remark that the randomized trace technique works perfectly well in conjunction with the imputation technique. The components of the noise vector ξ in the randomization techique are generated only where there are observations.

7 Starting Guesses, Outliers

Good starting guesses for the imputations, if any, for the λ_β's, and for the $\tilde{\mathbf{f}}_\beta$ are all required for the smoothing spline ANOVA fits to converge rapidly. Figure 1 is a contour plot of the global average winter temperature in 1981. The station winter average temperatures y_i were the averages of the December, January and February monthly average temperatures obtained from the Jones/Wigley data files obtainable from the Carbon Dioxide Information and Analysis Center (CDIAC) at Oak Ridge National Laboratory (ORNL) in the files ndp020r1/jonesnh.dat.Z and ndp020r1/jonessh.dat.Z in the pub directory at 128.219.24.36, see also [28]. The dots indicate the station locations, and we are using a subset of $n_2 = 725$ stations that had complete records for the winter of 1981. This contour plot was obtained by using RKPACK with the GCV estimate of λ to fit a smoothing spline on the sphere in the space $[1^{(2)}] \oplus \mathcal{H}_s^{(2)}$ described previously, with the RK for $\mathcal{H}_s^{(2)}$ given by $\tilde{R}_2(P, P')$. RKPACK returns the c and d of (10) (in this case $p = 1$), and the λ obtained by GCV. This could be done for each year and starting guesses for the various components of the full smoothing spline ANOVA could, for example be obtained by computing the marginal vectors (via applying $P_o^{(1)}$ and $P_o^{(2)}$ to the n_1 yearly fits at the full set of n_2 station locations). Starting guesses for λ_3 and λ_4 could be obtained by combining the λ's that come from the one year at a time global fits, and starting guesses for λ_1 and λ_2 by obtaining fits to the marginals. Alternatively, in data like this, it may be desireable to view the year main effect at several levels of smoothing, in this case it might be desireable to choose one or more values of λ_1 by 'eyeball'. A plot of the global yearly average temperature for the years 1854 thru 1993 appears in [28], and a similar plot for 1856 thru 1995 may be found in the New York Times of September 10, 1995, although technical details of the method of computing the global yearly averages are not given.

In meteorological data sets of the type we are considering, the occasional grossly erroneous data point is the rule rather than the exception, due to instrumental and human errors of various types. The residuals after a smoothing spline ANOVA fit of the kind described here may be examined and used as a screening tool for gross data errors. See, for example, Knight [30].

8 Summary

We have reviewed some theory and practice of smoothing spline ANOVA fits, and outlined an algorithm which has the potential for fitting very large environmental data sets with nearly regular structure by smoothing spline ANOVA methods and GCV estimates of smoothing parameters. Regular structure is exploited to use matrix decompositions for the marginal smoother matrices only. The Leaving-

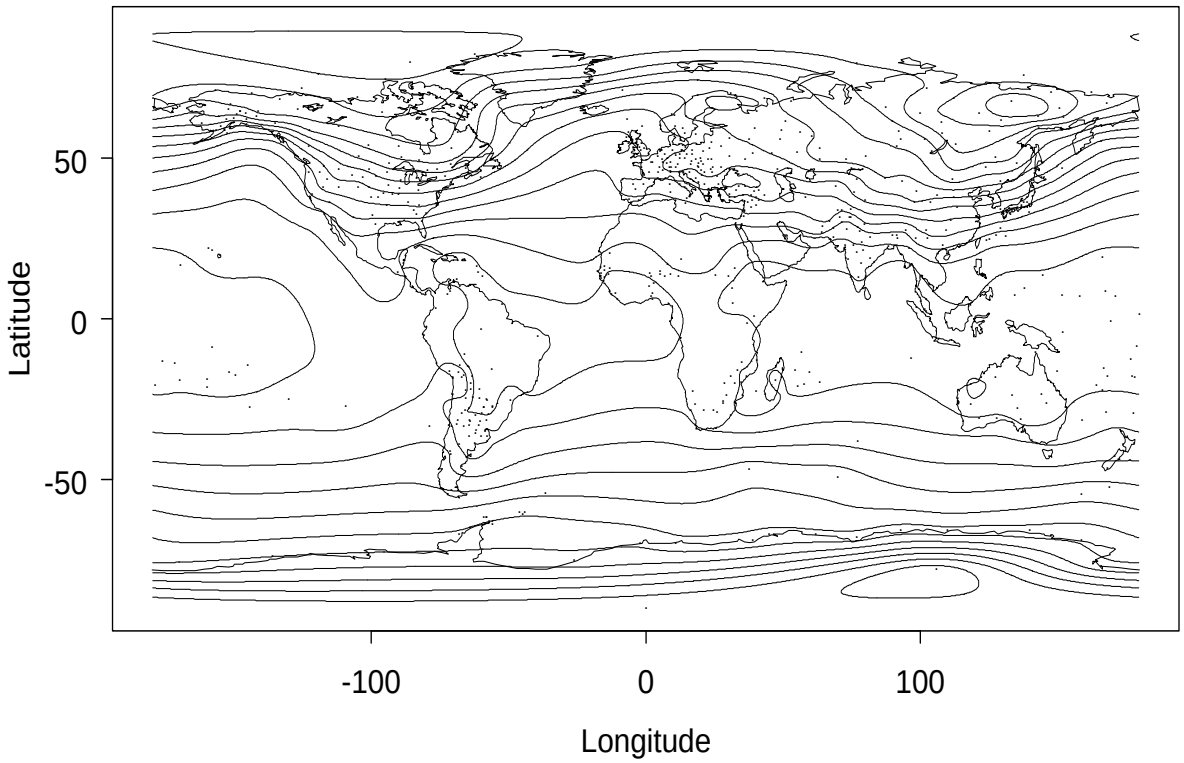

Figure 1: Contour plot of global average winter temperature, 1981

Out-K Lemma demonstrates that under a mild condition, an iterative data imputation can be used to fill in missing data in an otherwise regularly structured data set in a defensible manner - that is, the imputation converges to the imputation that would result if the smoothing spline ANOVA variational problem were solved without the imputed data and the imputation were made from the solution. The analysis of historical global winter surface temperature data has been described as a potential application. In some preliminary numerical work (in preparation) we have been able to analyze 30 years of global winter surface temperatures from 1000 stations, with about 50% of the 30,000 possible observations present.

Acknowledgements

We would like to thank Jan Helgeland who first suggested to us the fitting of a smoothing spline ANOVA model to surface temperature data, and to Donald R. Johnson and David Callan for helpful discussions.

References

[1] C. Ansley and R. Kohn. Convergence of the backfitting algorithm for additive models. *J. Austral. Math. Soc. Series A*, 57:316–329, 1994.

[2] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Ann. Statist.*, 17:453–555, 1989.

[3] J. Chambers and T. Hastie. *Statistical Models in S.* Wadsworth and Brooks, 1992.

[4] Z. Chen. Interaction spline models and their convergence rates. *Ann. Statist.*, 19:1855–1868, 1991.

[5] Z. Chen. Fitting multivariate regression functions by interaction spline models. *J. Roy. Stat. Soc. B*, 55:473–491, 1993.

[6] Z. Chen, C. Gu, and G. Wahba. Comments to 'Linear Smoothers and Additive Models', by Buja, Hastie and Tibshirani. *Ann. Statist.*, 17:515–521, 1989.

[7] 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.

[8] F. Gao. On combining data from multiple sources with unknown relative weights (thesis). Technical Report 902, Dept. of Statistics, University of Wisconsin, Madison, WI, 1993.

[9] D. Girard. A fast 'Monte Carlo cross validation' procedure for large least squares problems with noisy data. Technical Report RR 687-M, IMAG, Grenoble, France, 1987.

[10] D. Girard. A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data. *Numer. Math.*, 56:1–23, 1989.

[11] D. Girard. Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Statist.*, 19:1950–1963, 1991.

[12] G. Golub and C. VanLoan. *Matrix Computations, Second Edition.* Johns Hopkins University Press, 1989.

[13] G. Golub and Urs VonMatt. Generalized cross validation for large scale problems. Technical Report xx, Stanford University, Stanford, CA, 1995.

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

[15] C. Gu. RKPACK and its applications: fitting smoothing spline models. In *Proceedings of the Statistical Computing Section*, pages 42–51. American Statistical Association, 1989. Code available thru `netlib`.

[16] C. Gu. Adaptive spline smoothing in non-Gaussian regression models. *J. Amer. Statist. Assoc.*, 85:801–807, 1990.

[17] C. Gu. Cross-validating non-Gaussian data. *J. Comput. Graph. Stats.*, 1:169–179, 1992.

[18] C. Gu. Diagnostics for nonparametric regression models with additive terms. *J. Amer. Statist. Assoc.*, 87:1051–1057, 1992.

[19] C. Gu. Penalized likelihood regression: a Bayesian analysis. *Statistica Sinica*, 2:255–264, 1992.

[20] C. Gu, D.M. Bates, Z. Chen, and G. Wahba. The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal.*, 10:457–480, 1989.

[21] C. Gu and G. Wahba. Semiparametric ANOVA with tensor product thin plate splines. Technical Report 90-61, Dept. of Statistics, Purdue University, Lafayette, IN, 1990, to appear, *J. Roy. Stat. Soc. Ser. B.*

[22] C. Gu and G. Wahba. Comments to 'Multivariate Adaptive Regression Splines', by J. Friedman. *Ann. Statist.*, 19:115–123, 1991.

[23] C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, 12:383–398, 1991.

[24] C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *J. Royal Statistical Soc. Ser. B*, 55:353–368, 1993.

[25] C. Gu and G. Wahba. Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Computational and Graphical Statistics*, 2:97–117, 1993.

[26] T. Hastie and R. Tibshirani. *Generalized additive models.* Chapman and Hall, 1990.

[27] M. Hutchinson. A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun. Statist.-Simula.*, 18:1059–1076, 1989.

[28] P. Jones, T. Wigley, and K. Briffa. Global and hemispheric temperature anaomalies-land and marine instrumental records. In T. Boden, D. Kaiser, R. Sepanski, and F. Stoss, editors, *Trends '93: A Compendium of Data on Global Change, ORNL/CDIAC-65*, pages 603–608, Oak Ridge, TN, 1994. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory.

[29] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[30] R. Knight. A comparison of some methods for flagging erroneous observations in certain types of meteorological data. Technical Report 610, Dept. of Statistics, University of Wisconsin, Madison, WI 53706, 1980.

[31] Z. Luo. *Ph.D thesis.* PhD thesis, Dept. of Statistics, University of Wisconsin, Madison, WI 53706, 1996.

[32] C. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[33] G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Stat. Soc. Ser. B*, 40:364–372, 1978.

[34] G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comput.*, 2:5–16, 1981.

[35] G. Wahba. Erratum: Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comput.*, 3:385–386, 1982.

[36] G. Wahba. Vector splines on the sphere, with application to the estimation of vorticity and divergence from discrete, noisy data. In W. Schempp and K. Zeller, editors, *Multivariate*

Approximation Theory, Vol.2, pages 407–429. Birkhauser Verlag, 1982b.

[37] G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Stat. Soc. Ser. B*, 45:133–150, 1983.

[38] 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.

[39] G. Wahba. Partial and interaction splines for the semiparametric estimation of functions of several variables. In T. Boardman, editor, *Computer Science and Statistics: Proceedings of the 18th Symposium*, pages 75–80. American Statistical Association, Washington, DC, 1986.

[40] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

[41] G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII*, pages 95–112. Addison-Wesley, 1992.

[42] G. Wahba, D. Johnson, F. Gao, and J. Gong. Adaptive tuning of nummerical weather prediction models: Part I: randomized GCV and related methods in three and four dimensional data assimilation. Technical Report 920, Dept. of Statistics, University of Wisconsin, Madison, WI, to appear *Monthly Weather Review*, 1994.

[43] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. Technical Report 940, Department of Statistics, University of Wisconsin, Madison, WI, to appear, *Ann. Statist.*, 1994.

[44] G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, 108:1122–1145, 1980.

[45] R. Weber and P. Talkner. Some remarks on spatial correlation function models. *Mon. Wea. Rev.*, 121:2611–2617, 1993.