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

TECHNICAL REPORT NO. 964

Backfitting in smoothing spline ANOVA, with application to historical global temperature data[1]

by

Zhen Luo

July 22, 1996

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
(Statistics)

at the
UNIVERSITY OF WISCONSIN – MADISON
1996

[1]This research was supported in part by National Science Foundation Grant DMS 9121003 and National Aeronautics and Space Administration Grant NAGW-2961. E-mail zhen@stat.wisc.edu

Abstract

In the attempt to estimate the temperature history of the earth using the surface observations, various biases can exist. An important source of bias is the incompleteness of sampling over both time and space. There have been a few methods proposed to deal with this problem. Although they can correct some biases resulting from incomplete sampling, they have ignored some other significant biases.

In this dissertation, a smoothing spline ANOVA approach which is a multivariate function estimation method is proposed to deal simultaneously with various biases resulting from incomplete sampling. Besides that, an advantage of this method is that we can get various components of the estimated temperature history with a limited amount of information stored. This method can also be used for detecting erroneous observations in the data base. The method is illustrated through an example of modeling winter surface air temperature as a function of year and location. Extension to more complicated models are discussed.

The linear system associated with the smoothing spline ANOVA estimates is too large to be solved by full matrix decomposition methods. A computational procedure combining the backfitting (Gauss-Seidel) algorithm and the iterative imputation algorithm is proposed. This procedure takes advantage of the tensor product structure in the data to make the computation feasible in an environment of limited memory. Various related issues are discussed, e.g., the computation of confidence intervals and the techniques to speed up the convergence of the backfitting algorithm such as collapsing and successive over-relaxation.

Acknowledgements

It has been a great pleasure for me to work with Professor Grace Wahba, whose dedication to statistics and science in general has been a tremendous inspiration to me. She has always encouraged me to think hard and deep statistically, and at the same time keep a scientific perspective in mind. Professor Wahba initiated the research described in this dissertation. During the course of the study she had many useful discussions with me and provided several constructive suggestions and comments which significantly improved the scope and depth of this work.

Professor Donald Johnson of the Space Science and Engineering Center has given me many scientific insights about the application described here. Professor Michael Newton and Professor Brian Yandell have discussed with me several issues in this study, including its Markov chain Monte Carlo aspect, and presented some insightful comments. Professor Wei-Yin Loh and Professor Michael Kosorok reviewed this dissertation and provided me with many refinement suggestions. They, together with Professor Johnson and Professor Newton, as members of my defense committee, have made quite a few excellent suggestions for further work.

During my four-year-and-half study at the University of Wisconsin at Madison, I have also benefited from the discussions with Professors Richard Chappell, Tom Kurtz, Jun Shao and Bin Yu in and out of the classroom.

Fellow graduate students Patrick Gaffney, Jianjian Gong, Yuedong Wang and Dong Xiang have helped me in various ways in the course of this study and the actual writing of the thesis. These and other fellow graduate students, especially Yinzhong Chen, Claes Ekman, Anna Grimby, Zhengqing Li, Bob Mau, Wei Pan, Peng Qu, Jaya Satagopan, Wen-Hsiang Wei, Yonghong Yang and Yunlei Zhang have discussed statistics with me at different times and made

my life in Madison an enjoyable one.

My family in China and my wife, Aiping, have always held enormous confidence in me. Without their support and love, I would not have been able to come this far. Aiping read an earlier draft of this thesis and helped me a lot with my writing.

This research was supported in part by NSF under Grant DMS-9121003 and in part by NASA Grant NAGW-2961.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter the motivation for the research documented in this dissertation is discussed first. It is then followed by an outline of the contents of the subsequent chapters.

1.1 Motivation

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 are starting to model the global climate and to use their models to predict future climate. From "climate models which include as their central components atmospheric and oceanic General Circulation Models (GCMs)" to "climate system models which include all aspects of the climate system: the atmosphere, the ocean, the cryosphere, the biosphere and terrestrial ecosystems, other land surface processes, and additional parts of the hydrosphere including rivers, and all the complex interactions between these components" (Trenberth (1992)), the models get more and more complicated. A more complicated model is supposed to be closer to the true 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. Another reason for an accurate and easily accessible description is for the purpose of getting more information, especially graphical patterns, out of the data.

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 have only 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 exist such as the relocation of a surface station, the change of instrument, etc. 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 used different approaches to avoid 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). See also Hurrell and Trenberth (1996) for a comparison of monthly mean surface temperatures with those of global Microwave Sounding Unit (MSU) 2R temperatures for the period of 1979-1995.

Despite the many advantages previous studies have, there are also some common inherited biases existing in them, due to the fact that while all of them have considered the variation of mean temperatures, none of them has taken into consideration the variation of temperature change at different places when correcting the bias resulting from the spatial sampling difference,

We propose a smoothing spline method to deal simultaneously with these biases resulting from incomplete data coverage. Computational demand is huge if we want to solve the linear system associated with the smoothing spline method using decomposition methods for full matrices. There are many ways to save computing time and space in different contexts. Approximate computation is one way that may save computing space or time in many cases. See Luo and Wahba (1997) for such an example. Another way is to make use of the special structures of the specific data at hand. An example is the backfitting algorithm used in fitting additive models (Buja, Hastie and Tibshirani (1989)). One purpose of our study here is to explore the possibility of making use of the (space-time) tensor product structure in our data. This kind of structure exists in many climate, environmental and other studies, hence the methods described in this dissertation may be of wider interest than just in the study of global surface temperature data.

1.2 Outline

In Chapter 2, we discuss a computational procedure for fitting the smoothing spline ANOVA models to data sets of a tensor product structure, with an example of modeling global winter mean surface temperature. This example is the primary model considered in Chapter 3. The computational procedure combines the backfitting (Gauss-Seidel) algorithm with an iterative imputation procedure in order to take care of the situations of incomplete tensor product structure.

We introduce smoothing spline ANOVA models first in Section 2.1. It is followed by the derivation of the backfitting algorithm for a perfect tensor product structured data set and the discussion of its convergence in Section 2.2. Then we discuss some issues in speeding up this algorithm in Section 2.3. The issues we discuss include orthogonality, grouping, collapsing, and successive over-relaxation (SOR). Those techniques are often necessary to insure the backfitting algorithm to converge in real time. Then the justification for the use of an iterative imputation procedure is given in Section 2.4. Finally some empirical studies of the convergence of this computational procedure are discussed in Section 2.5.

In Chapter 3, we apply the computational procedure introduced in Chapter 2 to a smoothing spline ANOVA model of global winter mean surface air temperature. We summarize other people's approaches first in Section 3.1. Then in Section 3.2, we discuss the relationship between the smoothing spline estimates and the "statistical optimal averaging" estimates in Vinnikov et. al. (1990), and the use of anomalies to correct the biases resulting from spatial sampling difference as well as the limitation of the anomaly approach in correcting all such biases. In Section 3.3, a smoothing spline ANOVA model is fitted to the global winter mean temperature data. Various issues in fitting such a model, such as choosing smoothing parameters and diagnostics, are discussed. We also discuss an extension to a more complicated model. A small simulation is used in Section 3.4 to get some confidence statements about the estimates in Section 3.3.

In Chapter 4, we discuss a correspondence between Monte Carlo methods and optimization methods. This chapter is an introduction to our on-going research. Section 4.1 introduces a Bayesian model. The posterior mode under this model is exactly the smoothing spline estimate in Section 3.3. The posterior

mean (same as the mode in this case) and variance may be used to construct confidence intervals for the smoothing spline estimate. The computation of the posterior variance has the similar difficulties encountered in computing the posterior mode. Monte Carlo methods may be used to compute them. This leads to the discussion in Section 4.2 about a correspondence between the backfitting algorithm and the Gibbs sampler. Their parallel speeding-up techniques are discussed too. Section 4.3 describes other analogous Monte Carlo and optimization algorithms.

Chapter 2

Smoothing spline estimates and the backfitting algorithm

In this chapter we will describe a computational procedure for fitting a smoothing spline ANOVA model when data have a tensor product design and are too large in size to use direct matrix decomposition methods.

The basic idea of this algorithm is that the backfitting (block Gauss-Seidel) algorithm enables us to take advantage of a tensor product design when we solve the linear system associated with smoothing spline estimates. To speed up the convergence of this iterative method, various techniques such as SOR, collapsing components, etc., are used. When the data do not have a perfect tensor product design, which is often the case, an iterative imputation method is used to impute the data into the desired form.

We will discuss in a general SS-ANOVA setup whenever it is convenient. More often, we will use a model useful in a climate study to illustrate our points.

2.1 Multivariate smoothing spline estimates

Our central problem here is to estimate a multivariate function f based on some noisy data

$$y_i = f(t_i) + \epsilon_i,\ t_i \in \mathcal{T},\ i = 1, ..., n \tag{2.1.1}$$

where $\mathcal{T}$, the domain of the function, is of more than one dimension. For various reasons such as the convenience of interpretation or building a model, we may be interested in a decomposition of f into some component functions besides f itself, for example, an ANOVA type decomposition. In order to make these component functions well defined, we assume that $\mathcal{F}$, a linear space of functions of t which we assume contains f, can be decomposed as a direct sum of its subspaces

$$\mathcal{F} = \mathcal{F}^0 + \mathcal{F}^1 + \cdots + \mathcal{F}^p$$

i.e., the decomposition of any f in $\mathcal{F}$ into component functions in these subspaces is unique. Usually $\mathcal{F}^0$ is of finite dimension and we denote its dimension by M.

Example In some climate studies, we are interested in a meteorology variable, for example, winter mean surface air temperature, as a function of year and geographical location (a time-space model). The year index x takes values in $\{1, 2, \cdots, n_1\}$ corresponding to a period of time. The location $P = (latitude, longitude)$ takes values on the unit sphere $\mathcal{S}$. Hence here $\mathcal{T} = \{1, 2, \cdots, n_1\} \times \mathcal{S}$.

Define averaging operators:

$$(\mathcal{E}_x f)(x, P) := \sum_{x=1}^{n_1} f(x, P)/n_1 \tag{2.1.2}$$

$$(\mathcal{E}_P f)(x, P) := \int_{\mathcal{S}} f(x, P) dP/4\pi \tag{2.1.3}$$

where the integral is an integration over the sphere. Then

$$\begin{aligned} I &= (\mathcal{E}_x + (I - \mathcal{E}_x))(\mathcal{E}_P + (I - \mathcal{E}_P)) \\ &= \mathcal{E}_x\mathcal{E}_P + (I - \mathcal{E}_x)\mathcal{E}_P + \mathcal{E}_x(I - \mathcal{E}_P) + (I - \mathcal{E}_x)(I - \mathcal{E}_P) \end{aligned} \tag{2.1.4}$$

defines a direct sum decomposition of the space of function $f(x, P)$ satisfying some integrability conditions. This decomposition singles out the year average and the global average. It corresponds to a decomposition of f:

$$f(x, P) = d_1 + g_1(x) + g_2(P) + g_{12}(x, P)$$

where these component functions satisfy

$$\mathcal{E}_x g_1 = \mathcal{E}_x g_{12} = \mathcal{E}_P g_2 = \mathcal{E}_P g_{12} = 0$$

Suppose we want to single out the linear trend along year too. We can just define another averaging operator in addition to the two defined above:

$$\begin{aligned}(\mathcal{E}'_x f)(x,P) &:= \frac{\sum_{x=1}^{n_1-1}(f(x+1,P)-f(x,P))}{n_1-1}\phi(x) \\ &= \frac{(f(n_1,P)-f(1,P))}{\phi(n_1)-\phi(1)}\phi(x) \qquad (2.1.5)\end{aligned}$$

where $\phi(x) = x - \frac{n_1+1}{2}$.

Similar to (2.1.4), these three averaging operators define six component functions through:

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

It is equivalent to say that f is decomposed in the following way:

$$f(x,P) = d_1 + d_2\phi(x) + g_1(x) + g_2(P) + g_{\phi,2}(P)\phi(x) + g_{12}(x,P) \qquad (2.1.6)$$

where the component functions satisfy

$$\begin{cases} \sum_{x=1}^{n_1} g_1(x) = g_1(n_1) - g_1(1) = 0 \\ \sum_{x=1}^{n_1} g_{12}(x,P) = g_{12}(n_1,P) - g_{12}(1,P) = 0 \\ \int_{\mathcal{S}} g_2(P)dP = \int_{\mathcal{S}} g_{\phi,2}(P)dP = \int_{\mathcal{S}} g_{12}(x,P)dP = 0 \end{cases} \qquad (2.1.7)$$

for any x and P. See Wahba (1990, Chapter 10), Gu and Wahba (1993a,b) for more about the way of formulating such ANOVA models. ■

Now we need to make some smoothness assumptions about the functions we want to estimate based on our finite noisy data. Without such assumptions to relate function values at different points, it is an impossible task to estimate true function values from a single copy of the function's noisy version, let alone to estimate function values at points other than data points. Suppose each $\mathcal{F}^\alpha$ has a subspace $\mathcal{H}^\alpha$ which is a reproducing kernel Hilbert space with an inner

product $< ., . >_{\mathcal{H}^\alpha}$ and the corresponding reproducing kernel $R_\alpha(t', t)$. That is, R_α is a positive definite function satisfying:

$$\begin{cases} R_\alpha(., t) \in \mathcal{H}^\alpha, \ \forall t \in \mathcal{T} \\ < f, R_\alpha(., t) >_{\mathcal{H}^\alpha} = f(t), \ \forall f \in \mathcal{H}^\alpha, t \in \mathcal{T} \end{cases} \tag{2.1.8}$$

This is equivalent to say that all the point evaluation functionals, $L_t(f) := f(t)$, on $\mathcal{H}^\alpha$ are continuous. See Aronszajn (1950) or Wahba (1990) for more about reproducing kernel Hilbert spaces.

We assume that the function f to be estimated is in

$$\mathcal{H} := \mathcal{H}^0 + \mathcal{H}^1 + \cdots + \mathcal{H}^p$$

It is easy to see that $\mathcal{H}$ is also a reproducing kernel Hilbert space when it is endowed with an inner product

$$< f, g >_{\mathcal{H}} := < f_0, g_0 >_{\mathcal{H}^0} + \sum_{\alpha=1}^{p} \frac{1}{\theta_\alpha} < f_\alpha, g_\alpha >_{\mathcal{H}^\alpha} \tag{2.1.9}$$

where $f = \sum_{\alpha=0}^{p} f_\alpha, g = \sum_{\alpha=0}^{p} g_\alpha, \ f_\alpha, g_\alpha \in \mathcal{H}^\alpha$, for any given positive numbers $\theta_\alpha, \alpha = 1, ..., p$. The corresponding reproducing kernel is

$$R(t', t) = R_0(t', t) + \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha(t', t) \tag{2.1.10}$$

It is quite clear that with such an inner product in $\mathcal{H}$, $\mathcal{H}^\alpha \perp \mathcal{H}^\beta$ for any $\alpha \neq \beta$. Therefore

$$\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \cdots \oplus \mathcal{H}^p \tag{2.1.11}$$

i.e. $\mathcal{H}$ is an orthogonal sum of $\{\mathcal{H}^\alpha\}_{\alpha=0}^{p}$.

Example (continued) Define an inner product in the space of functions of x, denoted by $\mathcal{H}^{(1)}$, as

$$\begin{aligned} < f, g > := & \ (\sum_{x=1}^{n_1} f(x))(\sum_{x=1}^{n_1} g(x)) + (f(n_1) - f(1))(g(n_1) - g(1)) \\ + & \sum_{x=1}^{n_1-2} (f(x+2) - 2f(x+1) + f(x))(g(x+2) - 2g(x+1) + g(x)) \end{aligned}$$

where three terms correspond to three subspaces. The first subspace consists of all constant functions, the second one 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. Hence we have a decomposition of the space of functions of x:

$$\mathcal{H}^{(1)} = [1] \oplus [\phi] \oplus \mathcal{H}_s^{(1)} \tag{2.1.12}$$

with obvious notations. It is apparent that $\mathcal{H}^{(1)}$ and its subspaces are all reproducing kernel Hilbert spaces.

In the space of functions of P, an inner product is defined as

$$< f, g >:= \frac{1}{4\pi}(\int_{\mathcal{S}} f(P)dP)(\int_{\mathcal{S}} g(P)dP) + \int_{\mathcal{S}} (\Delta f)(\Delta g)dP \tag{2.1.13}$$

where Δ is the Laplace-Beltrami operator, the analogue on the sphere of the Laplacian in Euclidean space. Hence a decomposition of the space of functions of P:

$$\mathcal{H}^{(2)} = [1] \oplus \mathcal{H}_s^{(2)} \tag{2.1.14}$$

where $\mathcal{H}_s^{(2)}$ contains all the functions in $\mathcal{H}^{(2)}$ such that $\int_{\mathcal{S}} f(P)dP = 0$. $\mathcal{H}^{(2)}$ and its subspaces are also reproducing kernel Hilbert spaces (see Wahba (1981)).

Now the decomposition of $\mathcal{H}$ is obtained through the tensor product of (2.1.12) and (2.1.14):

$$\begin{aligned} \mathcal{H} &= \mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)} \\ &= ([1] \oplus [\phi] \oplus \mathcal{H}_s^{(1)}) \otimes ([1] \oplus \mathcal{H}_s^{(2)}) \\ &= [1] \otimes [1] \ \oplus \ [\phi] \otimes [1] \ \oplus \ \mathcal{H}_s^{(1)} \otimes [1] \ \oplus \\ &\quad [1] \otimes \mathcal{H}_s^{(2)} \ \oplus \ [\phi] \otimes \mathcal{H}_s^{(2)} \ \oplus \ \mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)} \end{aligned} \tag{2.1.15}$$

The first two are combined as $\mathcal{H}^0$ with dimension 2 (i.e., $M = 2$). The last four are denoted by $\mathcal{H}^\alpha$, for $\alpha = 1, 2, 3, 4$, respectively.

It can be shown that all these spaces with corresponding inner-products are reproducing kernel Hilbert spaces too. It turns out that a closed form for the reproducing kernel of $\mathcal{H}^{(2)}$ is not available. Consequently, the evaluation of such a kernel can be expensive. We change the inner product to a topologically equivalent one so that the corresponding reproducing kernel has a simple closed

form easier for computation. See Wahba (1981, Section 3) for details. See also Wahba (1990, Chapter 3) for the reasons why such a change is reasonable in practice. ■

A smoothing spline (SS) estimate is a minimizer over $\mathcal{H}$ of

$$\sum_{i=1}^{n}(y_i - f(t_i))^2 + \sum_{\alpha=1}^{p}\frac{1}{\theta_\alpha}\|f_\alpha\|^2_{\mathcal{H}^\alpha} \tag{2.1.16}$$

$$= \sum_{i=1}^{n}(y_i - < \eta_i, f >_{\mathcal{H}})^2 + \|P_1 f\|^2_{\mathcal{H}} \tag{2.1.17}$$

where η_i is the representer of the evaluation functional at t_i, i.e., $< \eta_i, f >_{\mathcal{H}} = f(t_i)$, for any $f \in \mathcal{H}$, and P_1 is the projection operator of $\mathcal{H}$ into $\mathcal{H}_1 := \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \cdots + \mathcal{H}^p$. Note that the θ's, called "smoothing parameters", control the smoothness of each f_α. We will discuss the issue of choosing them in Section 3.3.1. Here, and everywhere in this chapter, we assume that smoothing parameters have been chosen.

Let $\xi_i = P_1\eta_i$, and $\{\phi_\nu,\ \nu = 1, 2, \cdots, M\}$ span $\mathcal{H}^0$. By the argument in Wahba (1990, p. 12) the SS estimate has a representation

$$f_\theta = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{i=1}^{n} c_i \xi_i \tag{2.1.18}$$

The argument is very simple. Since any f in $\mathcal{H}$ can be represented as $\sum_{\nu=1}^{M} d_\nu\phi_\nu + \sum_{i=1}^{n} c_i\xi_i + \rho$ where ρ is orthogonal to all ϕ_ν and ξ_i. Hence

$$\begin{aligned} < \eta_j, f > &= \sum_{\nu=1}^{M} d_\nu < \eta_j, \phi_\nu > + \sum_{i=1}^{n} c_i < \eta_j, \xi_i > + < \eta_j, \rho > \\ &= \sum_{\nu=1}^{M} d_\nu < \eta_j, \phi_\nu > + \sum_{i=1}^{n} c_i < \eta_j, \xi_i > \end{aligned} \tag{2.1.19}$$

because $< \eta_j, \rho >=< \eta_j, P_1\rho >=< P_1\eta_j, \rho >=< \xi_j, \rho >= 0$ (the first equality is due to the fact that $\rho \in \mathcal{H}_1$). Therefore the first part of (2.1.17) does not depend on ρ, while the second part is

$$\begin{aligned} \|P_1 f\|^2_{\mathcal{H}} &= \|\sum_{i=1}^{n} c_i\xi_i + \rho\|^2_{\mathcal{H}} \\ &= \sum_{i,j} c_i c_j < \xi_i, \xi_j >_{\mathcal{H}} + \|\rho\|^2_{\mathcal{H}} \end{aligned} \tag{2.1.20}$$

As a result, in order to minimize (2.1.17), ρ must be 0, that is f_θ must be of the form (2.1.18).

The representer ξ_i can be expressed in terms of these reproducing kernels:

$$\begin{aligned}\xi_i(t) &= <\xi_i, R(.,t)> = <P_1\eta_i, R(.,t)> \\ &= <\eta_i, P_1R(.,t)> = <\eta_i, \sum_{\alpha=1}^{p}\theta_\alpha R_\alpha(.,t)> \\ &= \sum_{\alpha=1}^{p}\theta_\alpha <\eta_i, R_\alpha(.,t)> = \sum_{\alpha=1}^{p}\theta_\alpha R_\alpha(t_i,t). \qquad (2.1.21)\end{aligned}$$

Considering (2.1.17-21) and

$$<\eta_j, \phi_\nu> = \phi_\nu(t_j), \qquad (2.1.22)$$

$$<\eta_j, \xi_i> = \xi_i(t_j) = \sum_{\alpha=1}^{p}\theta_\alpha R_\alpha(t_i,t_j), \qquad (2.1.23)$$

$$<\xi_i, \xi_j> = <P_1\eta_i, \xi_j> = <\eta_i, P_1\xi_j> = <\eta_i, \xi_j>, \qquad (2.1.24)$$

the SS estimate can be expressed as

$$f_\theta(t) = \sum_{\nu=1}^{M} d_\nu\phi_\nu(t) + \sum_{i=1}^{n} c_i \sum_{\alpha=1}^{p}\theta_\alpha R_\alpha(t_i,t), \qquad (2.1.25)$$

where $\{d := (d_1, \cdots, d_M)^T, c := (c_1, \cdots, c_n)^T\}$ is a minimizer of

$$\|y - Sd - Q_\theta c\|^2 + c^T Q_\theta c, \qquad (2.1.26)$$

where

$$S = (\phi_\nu(t_i))_{n\times M}, \qquad (2.1.27)$$

$$Q_\theta = (\sum_{\alpha=1}^{p}\theta_\alpha R_\alpha(t_i,t_j))_{n\times n}. \qquad (2.1.28)$$

The component functions corresponding to (2.1.11) are

$$f_0(t) = \sum_{\nu=1}^{M} d_\nu\phi_\nu(t), \qquad (2.1.29)$$

$$f_\alpha(t) = \theta_\alpha \sum_{i=1}^{n} c_i R_\alpha(t_i,t), \qquad (2.1.30)$$

for $\alpha = 1, 2, \cdots, p$.

The stationary equations for (2.1.26) are

$$\begin{cases} (S^T S)d & = & S^T(y - Q_\theta c) \\ (Q_\theta + I)Q_\theta c & = & Q_\theta(y - Sd). \end{cases} \tag{2.1.31}$$

Since (2.1.26), as a quadratic function in d and c, is non-negative, all its stationary points, i.e., solutions to (2.1.31), are minimizers. Even though these stationary points may not be the same, the functions they correspond to through the representation (2.1.18) are the same, thus the minimizer of (2.1.17) is unique, as long as S is of full rank. The reason is simple. Since by the second equation of (2.1.31),

$$Q_\theta c = (Q_\theta + I)^{-1} Q_\theta (y - Sd), \tag{2.1.32}$$

hence by the first equation of (2.1.31),

$$(S^T S)d = S^T(y - (Q_\theta + I)^{-1} Q_\theta (y - Sd)). \tag{2.1.33}$$

Rearrange terms on both sides,

$$S^T(I - (Q_\theta + I)^{-1} Q_\theta) Sd = S^T(I - (Q_\theta + I)^{-1} Q_\theta) y. \tag{2.1.34}$$

That is,

$$S^T(Q_\theta + I)^{-1} Sd = S^T(Q_\theta + I)^{-1} y. \tag{2.1.35}$$

Therefore,

$$d = (S^T(Q_\theta + I)^{-1} S)^{-1} S^T(Q_\theta + I)^{-1} y. \tag{2.1.36}$$

Hence d is uniquely decided by the stationary equations when S is of full rank. For any two different c's satisfying (2.1.31), their difference δ must satisfy $Q_\theta \delta = 0$. From

$$0 = \delta^T Q_\theta \delta = \sum_{i,j} \delta_i \delta_j < \xi_i, \xi_j > = < \sum_i \xi_i \delta_i, \sum_j \xi_j \delta_j >$$

we know that $\sum_i \xi_i \delta_i = 0$, therefore by the representation (2.1.18), the corresponding f_θ's are the same.

Since it does not matter which solution $\{d, c\}$ to (2.1.31) we pick to compute f_θ, we can just pick the solution to the following equations:

$$\begin{cases} 0 &= S^T c \\ (Q_\theta + I)c &= (y - Sd). \end{cases} \tag{2.1.37}$$

These are equations usually used to compute SS estimates through direct matrix decompositions. See equations (1.3.16) and (1.3.17) of Wahba (1990, pp. 12-13).

Example (continued) Corresponding to the space decomposition (2.1.15), a decomposition of f is:

$$f = f_0 + f_1 + f_2 + f_3 + f_4 \tag{2.1.38}$$

where $f_0(x, P) = d_1 + d_2\phi(x), f_1(x, P) = g_1(x), f_2(x, P) = g_2(P), f_3(x, P) = g_{\phi,2}(P)\phi(x), f_4(x, P) = g_{12}(x, P)$.

A smoothing spline estimate is defined as the minimizer of

$$\sum_{i=1}^{n}(y_i - f(x_i, P_i))^2 + \frac{1}{\theta_1}J_1(g_1) + \frac{1}{\theta_2}J_2(g_2) + \frac{1}{\theta_3}J_3(g_{\phi,2}) + \frac{1}{\theta_4}J_4(g_{12}), \tag{2.1.39}$$

where $J_1(g_1) = \sum_{x=1}^{n_1-2}(g_1(x+2) - 2g_1(x+1) + g_1(x))^2$, J_2 and J_3 are the same and topologically equivalent to $\int_{\mathcal{S}}(\Delta f)^2 dP$, and J_4 is derived from J_1 and J_2 as the norm of the tensor-product space. (J_1 and J_2 are norms in $\mathcal{H}_s^{(1)}$ and $\mathcal{H}_s^{(2)}$ respectively, J_4 is the corresponding tensor-product norm in $\mathcal{H}_4 = \mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)}$.)

The reproducing kernel for $\mathcal{H}_s^{(1)}$ is defined as follows. Let L be

$$\begin{pmatrix} 1 & -2 & 1 & \cdots & 0 \\ 0 & 1 & -2 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{2.1.40}$$

Thus $J_1(f) = f^T L^T L f$. Then $R_t(j, j')$, the reproducing kernel for $\mathcal{H}_s^{(1)}$, is the jj'-th entry of $(L^T L)^\dagger$ where $\dagger$ denotes the Moore-Penrose generalized inverse.

The reproducing kernel for $\mathcal{H}_s^{(2)}$ is defined as

$$R_s(P, P') = \frac{1}{2\pi}[\frac{1}{2}q_2(z) - \frac{1}{6}], \tag{2.1.41}$$

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

$$\begin{aligned} q_2(z) &= \frac{1}{2}\{ln\left(1+\sqrt{\frac{2}{1-z}}\right)[12\left(\frac{1-z}{2}\right)^2 - 4\left(\frac{1-z}{2}\right)] \\ &\quad -12\left(\frac{1-z}{2}\right)^{3/2} + 6\left(\frac{1-z}{2}\right) + 1\} \end{aligned} \tag{2.1.42}$$

(From Wahba (1981), (3.3) and (3.4)).

The reproducing kernels for $\mathcal{H}_\alpha$, $\alpha = 1, 2, 3, 4$, are therefore defined as in Table 1. Solving (2.1.37) with these kernels, we get a SS estimate through

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

Table 1: *The reproducing kernels of the four subspaces containing the four non-parametric components in Model (2.1.38).*

(2.1.25). ■

When the sample size, n, is not too large, the equations (2.1.37) can be solved through direct matrix decompositions. RKPACK developed by Chong Gu (1989) implements an approach in this direction. For a Dec Alpha 3000/400 machine with 188M memory, the largest data size we can handle using RKPACK is about 2000. When the sample size gets larger, for example, in the climate study of Chapter 3 the data size is easily larger than 10000, unless we make use of some special structures in our specific data set, it is difficult to imagine that those equations can be solved using full-matrix methods on the computers of the present or near future. One special structure which has been widely studied is the sparsity of matrices. Sparse matrices often result from the problems of numerical solution to partial differential equations. The special structure we will use here is tensor product structures.

2.2 The backfitting algorithm

The representation (2.1.25) can certainly be written as

$$f_\theta(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t) + \sum_{\alpha=1}^{p} \theta_\alpha \sum_{i=1}^{n} c_{i,\alpha} R_\alpha(t_i, t) \tag{2.2.1}$$

too, where $c_{i,\alpha}$ differs for different α. Since the minimizer of (2.1.17) is unique (assuming as usual that S is of full rank), we can minimize (2.1.17) within the class of functions of form (2.2.1) and get the same SS estimates as before. This leads to a problem of minimizing:

$$\|y - Sd - \sum_{\alpha=1}^{p} \theta_\alpha Q_\alpha c_\alpha\|^2 + \sum_{\alpha=1}^{p} \theta_\alpha c_\alpha^T Q_\alpha c_\alpha \tag{2.2.2}$$

over d and c_α, for $\alpha = 1, 2, \cdots, p$, where $Q_\alpha := (R_\alpha(t_i, t_j))_{n\times n}$.

The corresponding stationary equations are:

$$\begin{cases} (S^T S)d &= S^T(y - \sum_{\alpha=1}^{p} \theta_\alpha Q_\alpha c_\alpha) \\ (\theta_\beta Q_\beta + I)Q_\beta c_\beta &= Q_\alpha(y - Sd - \sum_{\alpha\neq\beta} \theta_\alpha Q_\alpha c_\alpha), \text{ for } \beta = 1, 2, ..., p \end{cases} \tag{2.2.3}$$

With an argument similar to the one used in the last section, any solution to the above equations will result in the uniquely defined smoothing spline estimate f_θ and its components. Without confusion within their context, we denote the component functions of SS estimate f_θ evaluated at data points as $f_0, f_1, \cdots, f_p$ also. That is,

$$\begin{aligned} f_0 &= Sd, \\ f_\alpha &= \theta_\alpha Q_\alpha c_\alpha, \end{aligned}$$

for $\alpha = 1, 2, \cdots, p$.

They must satisfy

$$\begin{cases} f_0 &= S_0(y - \sum_{\alpha=1}^{p} f_\alpha) \\ f_\beta &= S_\beta(y - \sum_{\alpha\neq\beta} f_\alpha), \text{ for } \beta = 1, 2, ..., p, \end{cases} \tag{2.2.4}$$

where $S_0 := S(S^T S)^{-1} S^T$ and $S_\beta := (Q_\beta + \frac{1}{\theta_\beta} I)^{-1} Q_\beta$ for $\beta = 1, 2, \cdots, p$. These S matrices are called "smoother matrices" (S_0, a projection matrix, is an extreme case of smoother matrices.)

This suggests an iterative method to solve the above equations, i.e.

$$\begin{cases} f_0^{(k)} &= S_0(y - \sum_{\alpha=1}^{p} f_\alpha^{(k-1)}) \\ f_\beta^{(k)} &= S_\beta(y - \sum_{\alpha<\beta} f_\alpha^{(k)} - \sum_{\alpha>\beta} f_\alpha^{(k-1)}), \text{ for } \beta = 1, 2, ..., p. \end{cases} \tag{2.2.5}$$

This is exactly the backfitting algorithm studied in Buja, Hastie and Tibshirani (1989).

It can be seen that this iterative method is equivalent to an alternating minimization scheme to the problem

$$\min_{f_0 \in \mathcal{L}(S), f_\alpha \in \mathcal{L}(Q_\alpha)} \|y - \sum_{\alpha=0}^{p} f_\alpha\|^2 + \sum_{\alpha=1}^{p} \frac{1}{\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha \tag{2.2.6}$$

where $Q_\alpha^\dagger$ 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 (2.2.4) using results in the optimization literature. (See, for example, Lunerberg (1984), Section 7.9 on pp. 227-228). See Ansley and Kohn (1994) for an interesting discussion of convergence issue.

Rewrite the equations (2.2.4) as

$$\begin{pmatrix} I & S_0 & \cdots & S_0 \\ S_1 & I & \cdots & S_1 \\ \cdots & & & \\ S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_p \end{pmatrix} = \begin{pmatrix} S_0 y \\ S_1 y \\ \vdots \\ S_p y \end{pmatrix}. \tag{2.2.7}$$

It is clear that the backfitting algorithm we have just described, (2.2.5), is a (block) Gauss-Seidel algorithm.

Having known f_0 $(= Sd)$, we know d immediately. By (2.1.37), $(Q_\theta + I)c = y - Sd$, hence

$$c = y - Sd - Q_\theta c = y - \sum_{\alpha=0}^{p} f_\alpha. \tag{2.2.8}$$

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

One advantage of the backfitting algorithm is that it enables us to take advantage of some special structures of Q_α in some specific applications. In Buja et. al. (1989), additive models are fitted by backfitting where each marginal

smoother is a one-dimensional smoother which has a sparse matrix representation due to O'Sullivan. Here marginal smoothers are full matrices, but they have a tensor product structure if the data have a tensor-product design. This structure is what we want to make use of.

Example (continued) Suppose we have data at every point (x_i, P_j) for $i = 1, 2, ..., n_1$ and $j = 1, 2, ..., n_2$. That is, the data have a tensor product design. Hence the sample size $n = n_1 n_2$. Then the S and Q_α's have the following forms:

$$\begin{aligned} S &= 1 \otimes \tilde{S} \\ Q_1 &= 11^T \otimes Q_t \\ Q_2 &= Q_s \otimes 11^T \\ Q_3 &= Q_s \otimes \phi\phi^T \\ Q_4 &= Q_s \otimes Q_t \end{aligned}$$

where 1 is a vector of ones of appropriate length, $\phi = (\phi(1), ..., \phi(n_1))^T$, $\tilde{S} = (1\ \phi)_{n_1 \times 2}$, 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 $R_t(i,j)$.

Given such tensor product structures, in order to get the eigen-decomposition of matrices $\{Q_\alpha\}$, we only need to decompose Q_s and Q_t which are much smaller in size compared with $\{Q_\alpha\}$. Note that we cannot take advantage of this structure in (2.1.37), because $Q_\theta = \sum_{\alpha=1}^4 \theta_\alpha Q_\alpha$ does not have a tensor-product structure even though every single Q_α does. This is exactly the reason why we want to use the backfitting algorithm. Now with the eigen-decompositions of $\{Q_\alpha\}$, hence $\{S_\alpha\}$, updating (2.2.5) involves just a few matrix multiplications. ■

2.3 Issues in speeding up backfitting

In many cases, a straight-forward implementation of the backfitting algorithm converges very slowly. There are many discussions about speeding up the Gauss-Seidel algorithm in the numerical analysis literature, especially about an algorithm called successive over-relaxation (GS is its special case). See, for example, Young (1971). Here we would like to discuss some of these issues in the context of fitting a smoothing spline model.

2.3.1 Orthogonality

Roughly speaking, the main reason for the slowness of the backfitting (Gauss-Seidel) algorithm is the correlation between components. For the purpose of illustration, consider a trivial problem of minimizing $f(c) := c^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} c$ where ρ is between -1 and 1. The spectral radius of the updating matrix of the alternating minimization (i.e., GS) algorithm applied here is easy to be verified to be ρ^2. Hence the larger "correlation coefficient" ρ is, the slower the GS algorithm converges. If ρ is zero, then the GS 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 can be reduced into some smaller problems. Besides the possible gain in the computational speed, the benefit of such an approach is that the smaller problems are also easier to be analyzed in general. It is because of this, i.e., the reduction of the original problem into smaller ones, we are able to analyze SOR in our application analytically in Section 2.3.3.

Example (continued) Recall that $Q_t = (L^T L)^\dagger$ where L is given by (2.1.40), hence $Q_t 1 = Q_t \phi = \phi^T 1 = 0$. 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 $Q_3 S$. Hence the minimization problem (2.2.6) 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

$$\begin{aligned} & \|y - \sum_{\alpha=0}^{4} f_\alpha\|^2 + \sum_{\alpha=1}^{4} \frac{1}{\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha \\ = \; & \|y - f_0 - f_2 - f_3\|^2 + \frac{1}{\theta_2} f_2^T Q_2^\dagger f_2 + \frac{1}{\theta_3} f_3^T Q_3^\dagger f_3 + \\ & \|y - f_1 - f_4\|^2 + \frac{1}{\theta_1} f_1^T Q_1^\dagger f_1 + \frac{1}{\theta_4} f_4^T Q_4^\dagger f_4 \\ & - \|y\|^2. \end{aligned} \tag{2.3.1}$$

Therefore, (2.2.6) is equivalent to solving the following two problems separately:

$$\min_{f_0 \in \mathcal{L}(S), f_2 \in \mathcal{L}(Q_2), f_3 \in \mathcal{L}(Q_3)} \|y - f_0 - f_2 - f_3\|^2 + \frac{1}{\theta_2} f_2^T Q_2^\dagger f_2 + \frac{1}{\theta_3} f_3^T Q_3^\dagger f_3 \tag{2.3.2}$$

and

$$\min_{f_1\in\mathcal{L}(Q_1),f_4\in\mathcal{L}(Q_4)} \|y-f_1-f_4\|^2+\frac{1}{\theta_1}f_1^TQ_1^\dagger f_1+\frac{1}{\theta_4}f_4^TQ_4^\dagger f_4. \tag{2.3.3}$$

They correspond to solving the following two systems:

$$\begin{pmatrix} I & S_0 & S_0 \\ S_2 & I & 0 \\ S_3 & 0 & I \end{pmatrix}\begin{pmatrix} f_0 \\ f_2 \\ f_3 \end{pmatrix}=\begin{pmatrix} S_0y \\ S_2y \\ S_3y \end{pmatrix} \tag{2.3.4}$$

and

$$\begin{pmatrix} I & S_1 \\ S_4 & I \end{pmatrix}\begin{pmatrix} f_1 \\ f_4 \end{pmatrix}=\begin{pmatrix} S_1y \\ S_4y \end{pmatrix}, \tag{2.3.5}$$

respectively.

The key reason for such a reduction is that the x variable has an equally-spaced design. But even with the same design, if we choose to treat x after normalization as a continuous variable in $[0,1]$ and to use the same reproducing kernel as that used in Gu and Wahba (1993b), then Q_t1 and $Q_t\phi$ will not be zero anymore even though they may be very small.

If the design of every variable is equally-spaced, then choosing appropriate reproducing kernels 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, we only need to apply marginal smoothers to the data once to get all component functions. ■

2.3.2 Grouping and Collapsing

Consider the problem (2.2.6). Instead of minimizing it with respect to one component by one component which leads to the backfitting algorithm (2.2.5), we can minimize it with respect to more than one component at a time. Of course, each updating step is more complicated due to the higher dimension of the problem. In many cases, however,this will reduce the number of iterations needed in the backfitting algorithm. (See Varga (1962), p. 80, for a counter-example.) A compromise between the cost of updating and the number of iterations needed has to be considered.

Another possible way, in a similar spirit to save computing time, is through what we call a "collapsing" technique. We will now illustrate this method, again using the same example used before.

Example (continued) It may be the case that the backfitting applied to (2.3.5) is so slow that we would like to avoid any iteration completely.

Rewrite equations in (2.3.5) as

$$\begin{aligned} f_1 &= (\frac{1}{\theta_1}I + Q_1)^{-1}Q_1(y - f_4) \\ f_4 &= (\frac{1}{\theta_4}I + Q_4)^{-1}Q_4(y - f_1). \end{aligned}$$

Hence

$$\begin{aligned} (\frac{1}{\theta_1}I + Q_1)f_1 &= Q_1(y - f_4) \\ (\frac{1}{\theta_4}I + Q_4)f_4 &= Q_4(y - f_1). \end{aligned}$$

Rearrange terms on both sides:

$$\begin{aligned} f_1 &= \theta_1 Q_1(y - f_1 - f_4) \\ f_4 &= \theta_4 Q_4(y - f_1 - f_4). \end{aligned}$$

Add these two equations together:

$$f_1 + f_4 = (\theta_1 Q_1 + \theta_4 Q_4)(y - f_1 - f_4).$$

Denote $\theta_1 Q_1 + \theta_4 Q_4$ by Q_{1+4}, we have:

$$f_1 + f_4 = (Q_{1+4} + I)^{-1}Q_{1+4}y. \tag{2.3.6}$$

Therefore, we do not need any iteration to compute $f_1 + f_4$ if we can easily invert $(Q_{1+4} + I)$. $f_1 + f_4$ can then be used in (2.2.8) to get c and hence f_1 and f_4 afterwards.

We certainly do not want to decompose Q_{1+4} directly. In this case, fortunately, Q_{1+4} has a tensor product structure too:

$$Q_{1+4} = \theta_1(11^T \otimes Q_t) + \theta_4(Q_s \otimes Q_t) = (\theta_1 11^T + \theta_4 Q_s) \otimes Q_t =: \tilde{Q}_{1+4} \otimes Q_t.$$

We can eigen-decompose $\tilde{Q}_{1+4}$ which is of the same size as Q_s, then we get the eigen-decomposition of Q_{1+4} through the tensor product of the eigen-decompositions of $\tilde{Q}_{1+4}$ and Q_t.

Another application of collapsing is in solving (2.3.4). By the similar argument as that for (2.3.5),

$$f_2 + f_3 = (Q_{2+3} + I)^{-1} Q_{2+3}(y - f_0), \tag{2.3.7}$$

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

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

we get

$$(I - S_0(I + Q_{2+3})^{-1} Q_{2+3}) f_0 = S_0(I - (Q_{2+3} + I)^{-1} Q_{2+3}) y.$$

Since $f_0 = S_0 f_0$, it is equivalent to

$$S_0(I + Q_{2+3})^{-1} f_0 = S_0(I + Q_{2+3})^{-1} y.$$

Hence

$$S(S^T S)^{-1} S^T (I + Q_{2+3})^{-1} S d = S(S^T S)^{-1} S^T (I + Q_{2+3})^{-1} y.$$

Therefore,

$$d = (S^T (I + Q_{2+3})^{-1} S)^{-1} S^T (I + Q_{2+3})^{-1} y, \tag{2.3.8}$$

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 $f_2 = S_2(y - f_0), f_3 = S_3(y - f_0)$. Again no iteration is needed.

If the iteration of backfitting applied to (2.3.4) or (2.3.5) converges too slowly, then the extra cost of matrix decompositions (actually for (2.3.4), no extra decomposition is needed besides those of Q_s and Q_t) and matrix products may be worth taking in order to save overall computing time.

Note that if we apply the same argument to all four f_α's, we would end up with (2.1.37), where Q_θ does not have a tensor product structure such as Q_{2+3} in (2.3.7) has, thus it is much more difficult to invert $(I + Q_\theta)$ than to invert $(I + Q_{2+3})$. Therefore the problem here is 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. ■

2.3.3 SOR

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

Suppose we want to solve

$$\begin{pmatrix} I & S_0 & \cdots & S_0 \\ S_1 & I & \cdots & S_1 \\ \cdots & & & \\ S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \cdots \\ f_p \end{pmatrix} = \begin{pmatrix} S_0 y \\ S_1 y \\ \cdots \\ S_p y \end{pmatrix}. \tag{2.3.9}$$

The Gauss-Seidel updating scheme is

$$f_\alpha^{(k+1)} = S_\alpha (y - \sum_{\beta<\alpha} f_\beta^{(k+1)} - \sum_{\beta>\alpha} f_\beta^{(k)}). \tag{2.3.10}$$

The SOR scheme is

$$f_\alpha^{(k+1)} = \omega\{S_\alpha (y - \sum_{\beta<\alpha} f_\beta^{(k+1)} - \sum_{\beta>\alpha} f_\beta^{(k)})\} + (1-\omega) f_\alpha^{(k)}, \tag{2.3.11}$$

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.

The trick is to find a good ω. In general, only for some special kinds of matrix a prescribed optimal ω is available. Fortunately our case falls into this kind of situation.

Example (continued) Consider system (2.3.4), 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 $B := I - (diag\ A)^{-1}A$ 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

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

where $\{\mu_1, ..., \mu_n\}$ 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 $Q_s = \Gamma_s\Lambda_s\Gamma_s^T$, $Q_t = \Gamma_t\Lambda_t\Gamma_t^T$, $\Lambda_s = diag(\lambda_j^s)_{j=1}^{n_2}, \Lambda_t = diag(\lambda_i^t)_{i=1}^{n_1}$. Since $Q_t 1 = Q_t\phi = 0$, and $\phi^T 1 = 0$, 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

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

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

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

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

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

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

It is even easier to show that SOR for the system (2.3.5) converges too, also its optimal over-relaxation parameter can be computed given the estimate of the spectral radius of the corresponding GS 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 of the special structure of our linear system. The diagonal blocks are all identity matrices. Hence updating elements in one block one by one is the same as updating them simultaneously. ■

2.4 Iterative Imputation

So far we have assumed that our data are complete in the sense that every tensor product grid point has one observation. But frequently in observational studies, we have data missing here and there. For example, in the climate study to be described in Chapter 3, many stations have interrupted records due to various reasons. In such cases we can still make use of previously discussed computational procedures through the aid of an imputation technique.

For simplicity reason, suppose that we have reordered the data in such a way that the complete data y can be written as two parts

$$y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix}, \tag{2.4.1}$$

where $y^{(2)} = (y_{i_1}, \cdots, y_{i_K})^T$ 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",

then to calculate its predicted values at the missing part, e.g., $g^{(2)}$. After that, we impute $y^{(2)}$ with these newly predicted $g^{(2)}$ and go back to fit the same SS model again. We keep going through this cycle until the fitted values do not change anymore.

In the step of fitting a SS model, we can use backfitting and all other techniques discussed before.

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 more about the EM algorithm and its properties. See also Green (1990) for its use in penalized likelihood estimation.

The following lemmas are taken from Wahba and Luo (1996). It is shown in these lemmas that this iterative imputation procedure does converge to the SS estimate we want.

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 $\|P_1 f\|^2 = \sum_{\beta=1}^{p} \frac{1}{\theta_\beta} \|P^\beta f\|^2$. Let $f^{[K]}$ be the solution to the variational problem: Find $f \in \mathcal{H}$ to minimize

$$\sum_{i=1, i \notin S_K}^{n} (y_i - f(t(i)))^2 + \|P_1 f\|^2, \tag{2.4.2}$$

where $S_K = \{i_1, \cdots, i_K\}$ is a subset of $1, \cdots, n$ with the property that (2.1.17) has a unique minimizer, and let $y_i^*, i \in S_K$ be "imputed" values for the "missing" data imputed as $y_i^* = f^{[K]}(t(i)), i \in S_K$. Then the solution to the problem: Find $f \in \mathcal{H}$ to minimize

$$\sum_{i=1, i \notin S_K}^{n} (y_i - f(t(i)))^2 + \sum_{i \in S_K} (y_i^* - f(t(i)))^2 + \|P_1 f\|^2 \tag{2.4.3}$$

is $f^{[K]}$.

Yates (1933) uses a similar idea to fit an ordinary ANOVA model to the data with a few missing values, without solving a general linear model equation.

Let $A(\lambda)$ be defined by $\tilde{f} := (f^{[K]}(t(i)))_{i=1}^n = A(\lambda) y$, and $A(\lambda)$ be partitioned, corresponding to (2.4.1), as

$$A(\lambda) = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}. \tag{2.4.4}$$

Lemma 2 (The Imputation Lemma)
Let $g_{(0)}^{(2)}$ *be a* K*-vector of initial values for an imputation of* $(f^{[K]}(t(i_1)), \cdots, f^{[K]}(t(i_K)))^T$, *and suppose* $(I - A_{22}) \succ 0$ *(i.e. positive definite). Let successive imputations* $g_{(l)}^{(2)}$ *for* $l = 1, 2, \cdots$, *be obtained via*

$$\begin{pmatrix} g_{(l)}^{(1)} \\ g_{(l)}^{(2)} \end{pmatrix} = A(\lambda) \begin{pmatrix} y^{(1)} \\ g_{(l-1)}^{(2)} \end{pmatrix}. \tag{2.4.5}$$

Then

$$\lim_{l \to \infty} \begin{pmatrix} g_{(l)}^{(1)} \\ g_{(l)}^{(2)} \end{pmatrix} = \begin{pmatrix} f^{[K]}(t(1)) \\ \vdots \\ f^{[K]}(t(n)) \end{pmatrix}. \tag{2.4.6}$$

There is a simple sufficient and necessary condition for $(I - A_{22})$ to be positive definite.

Lemma 3 (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 (2.4.1) as*

$$\begin{pmatrix} \Gamma_{11} \\ \Gamma_{21} \end{pmatrix}. \tag{2.4.7}$$

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

An interpretation of this condition is based on the observation

$$S(S^TS)^{-1}S^T = \Gamma_1\Gamma_1^T = \begin{pmatrix} \Gamma_{11}\Gamma_{11}^T & \Gamma_{11}\Gamma_{21}^T \\ \Gamma_{21}\Gamma_{11}^T & \Gamma_{21}\Gamma_{21}^T \end{pmatrix}. \tag{2.4.8}$$

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, the condition in the lemma is a condition to exclude the most extreme influential case.

2.5 Convergence criteria and the verification of computation

In order to verify the above method of combining the backfitting and the EM algorithm to compute SS estimates, we compare the results using this method with the results using RKPACK of Gu (1989) which solves (2.1.37) using direct matrix decomposition methods.

Consider the model in the example of previous sections. From the data set used in Chapter 3, we choose a subset of 100 stations, distributed as uniformly as possible, and their 30 years' records. There are 2046 observations available. About one third of the total observations is missing. 2000 is about the largest data size for the model in our example which RKPACK can handle on our current computer with 192M memory.

Smoothing parameters (θ_α's) are chosen in a way that they are comparable to the results for 1000 stations shown in the next chapter, i.e. $\theta_1 = 10^{0.5}, \theta_2 = 10^3, \theta_3 = 1, \theta_4 = 10^{1.5}$.

For the backfitting iteration, we choose the relative differences of f_α's:

$$\max_{\alpha=0,1,\cdots,4} \frac{\|f_\alpha^{(k+1)} - f_\alpha^{(k)}\|^2}{\|f_\alpha^{(k)}\|^2} \tag{2.5.1}$$

as the convergence index. The convergence criterion is that the maximum relative difference is smaller than a pre-specified number δ_1.

For the EM iteration, (2.1.26) is chosen as the convergence index. Note that by (2.2.8), (2.1.26) equals

$$\begin{aligned} & \|y - f_0 - \sum_{\alpha=1}^{p} \theta_\alpha Q_\alpha c\|^2 + c^T \sum_{\alpha=1}^{p} \theta_\alpha Q_\alpha c \\ = \quad & \|y - \sum_{\alpha=0}^{p} f_\alpha\|^2 + (y - \sum_{\alpha=0}^{p} f_\alpha)^T \sum_{\alpha=1}^{p} f_\alpha, \end{aligned} \tag{2.5.2}$$

hence it can be computed easily after the f_α's are computed. The convergence criterion for the EM iteration is that (2.5.2) is smaller than a pre-specified number δ_2.

We compare three different levels of convergence. The first set of results are for $\delta_1 = 5. \times 10^{-4}$ and $\delta_2 = 1. \times 10^{-4}$. The second set of results are for $\delta_1 = 5. \times 10^{-5}$ and $\delta_2 = 1. \times 10^{-5}$. The third set of results are for $\delta_1 = 5. \times 10^{-6}$

and $\delta_2 = 1. \times 10^{-6}$. From the figure of estimated $f_0 + f_1$ by RKPACK and three backfitting fits (Figure 1), we see that even a relatively loose convergence criterion still gives a solution close to that given by RKPACK. From Figure 2 of the plots of backfitting results with $\delta_1 = 5. \times 10^{-6}$ and $\delta_2 = 1. \times 10^{-6}$ against RKPACK's results, we see that there are still some discrepancy between these two sets of computational results, especially in estimating f_4. We can certainly make our convergence criteria stricter so as to make the results even closer to those of RKPACK, but in practice, this may not be necessary for large size problems. The extra computing time may not be worthwhile since after all even the exactly computed results may not be the best estimates. If we start with some smooth values (for example, imputing missing values by the station means), we know our results are a little bit smoother than the exact SS estimates.

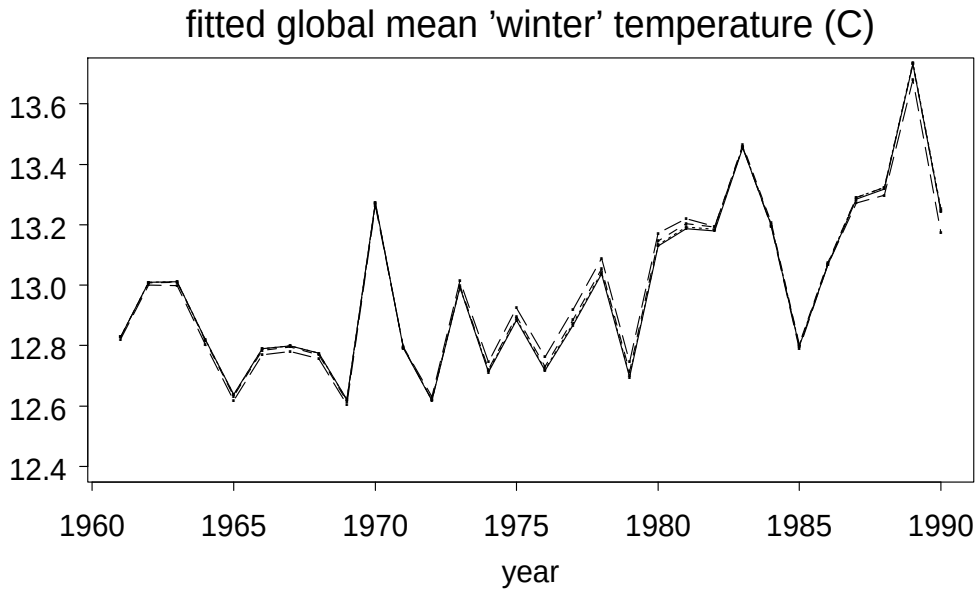

Figure 1: *Compare results computed by backfitting with different convergence criteria with those by RKPACK. 100 stations. Solid lines are the RKPACK results, broken lines are the backfitting results. Longer broken lines correspond to a cruder criterion.*

While the theory described in the previous sections requires that in every

backfitting step the iteration has to converge before the next imputation step starts, in practice there is a flexibility in varying how close the backfitting iteration is to its convergence. Our experience suggests that we can just choose an adequately strict criterion comparable to the criterion for the EM iteration which controls the closeness of our results to those of RKPACK eventually. An extreme choice is to do only one iteration in backfitting, hence form a big iteration including backfitting and EM updatings simultaneously. This reminds us of what many Bayesians usually do with missing values: to treat them as unknown parameters and update them together with "real" parameters. The problem is that it may be very slow to converge, and the theory described before does not apply anymore. Therefore, even though we may still use SOR or other speeding-up techniques, we do not know in theory when it does and when it does not converge.

Finally, as a precaution, the component function values computed from representation (2.1.29-30) using d and c in (2.2.8) should be checked against those obtained directly from backfitting. Any discrepancy may indicate that the convergence criterion is not strict enough. In general, results directly from backfitting are more reliable than those computed using d and c whose computations are more ill-conditioned than those of function values. It is important to check this point, especially when some smoothing parameters are extremely small compared with the others.

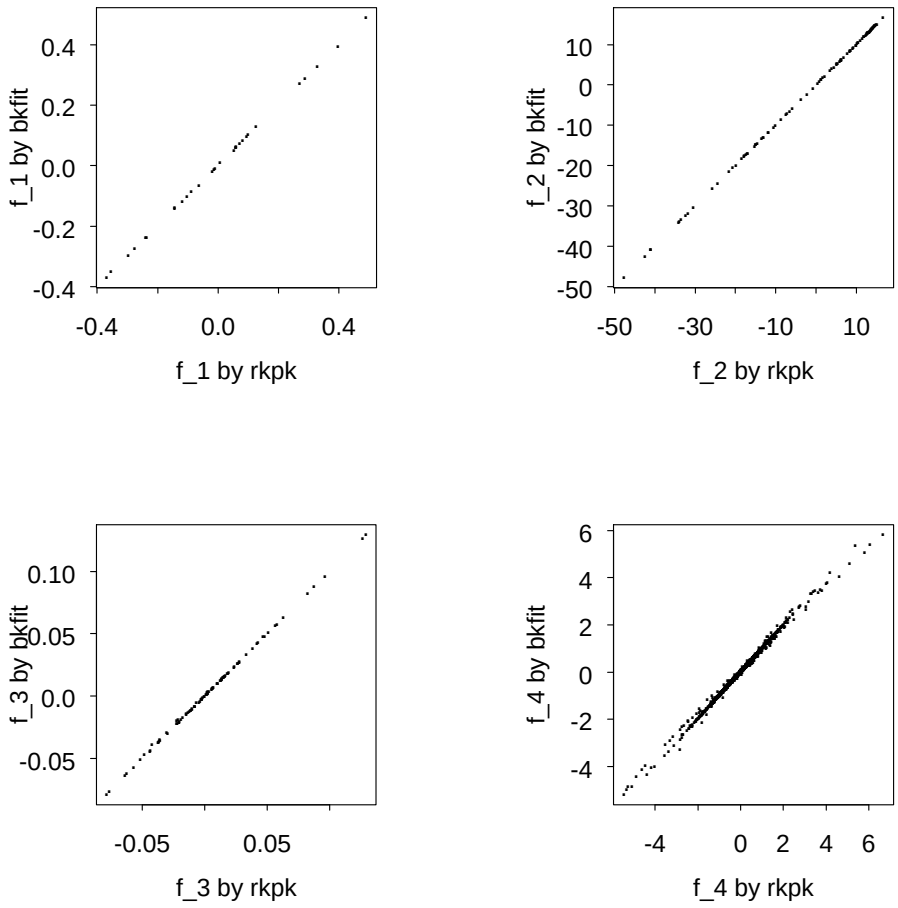

Figure 2: *Compare four component functions computed by backfitting (SOR's criterion is 5.d-6 while EM's criterion is 1.d-6) with those by RKPACK. 100 stations.*

Chapter 3

An application of smoothing spline ANOVA to global historical temperature data

In this chapter, a smoothing spline ANOVA model for global surface air temperature history is described and fitted using the computational procedures described in Chapter 2. Some important issues in practice such as choosing smoothing parameters and diagnostics are discussed too.

3.1 Introduction

In recent years a lot of attention has been paid to the climate change of the earth due to its tremendous impact on human life and the possibility of anthropogenic influences on the climate. There have been many studies on different aspects of this issue. An initial and yet important step towards a thorough understanding of this issue is to have an accurate picture of what has happened so far in the climate. Here we consider only one aspect of the climate, i.e. surface air temperature. However the method used here can also be applied to precipitation or other climate variables. Our goal is to give an accurate description of the global temperature history. We would like to know the history of global mean temperature, as well as local temperature history and its variation across the globe. Essentially, we want to calculate a whole bunch of summary statistics, e.g., different kinds of averages of the past temperature records. This seemingly easy

job is complicated by the non-uniform distribution of the weather stations taking these records and by the incomplete time coverage of the records available. Stations are concentrated more in Europe and North America. Some stations have a long history of records, others have very short ones, and some have interrupted records due to various reasons. The newly available satellite data may help to solve the problem resulting from the nonuniform spatial coverage, but its history is too short. For many historical studies of climate, surface station data are still the main source of information. Also, see Hurrell and Trenberth (1996) for a comparison of estimates based on satellite and surface data.

To calculate a global mean, the simple average of available station records is obviously biased towards the area concentrated with more stations. More sophisticated methods are needed. Vinnikov et. al. (1980) subjectively contoured the station data to get grid point estimates, then averaged them with cosine weighting to account for the change of grid density along latitude. Jones et. al. (1982) did a similar computation except with an objective method (nearest neighbor (with 6 neighbors) inverse distance weighting) to get grid point estimates. Later in Jones et.al. (1986), they divided the globe into 36 by 36 boxes and within each box the inverse distance weighted average was used to estimate the grid point value corresponding to that box. Hansen and Lebedeff (1987) divided the globe into a number of equal-area small boxes and computed a mean value within each box. Then a hierarchical average of box mean values (from small boxes to bigger boxes, then to latitude bands, to hemispheres, with different weighting schemes at different levels) is used as an estimate of the global mean. Vinnikov et.al. (1990) used an "optimal statistical averaging" method to compute different area mean values directly without computing grid point values.

To compare global means across time (the crudest way to look at global temperature change), there is another bias due to the incompleteness of time sampling, i.e. the stations having records are different from one year to another. The temperature change in time is confounded with the change in the location of stations. 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 the change in the average temperature is due to a real global change, or just due to the sampling difference between these two different sets of stations. The way most studies choose to correct this bias is through the use of anomalies which is defined as the difference

of raw records and the average over a pre-specified reference period. We will see in Section 3.2.2, while this approach is satisfactory in general, there are some significant biases it cannot correct. We will also show that our smoothing spline ANOVA approach can correct such biases without even using anomalies.

In Section 3.2, we will show our SS approach to the problem of spatial averaging of one-time data. In Section 3.3, we will describe our approach to data with both time and space dimensions.

The data set we choose to apply our approach is Jones et. al. (1991)'s data. 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 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.

A subset of this data set is chosen to illustrate our method. Only winter average temperature, defined as the average of December, January and February temperatures, is considered. The most recent 30-year period (1961-1990) is chosen. Instead of using all the stations in this data set, we selected 1000 stations due to the limit of our computing capacity. These 1000 stations are chosen deliberately so that they cover the sphere as uniformly as possible. Hence most stations left out are those in Europe and North America while almost all the stations in other regions are included. Note that this selection of stations only mitigates the problem of non-uniformness of station distribution, it does not eliminate the problem. The distribution of these 1000 stations is plotted in Figure 3.

To have a graphical idea of the incomplete time coverage, a plot of (year, latitude) for the records in our data set is given in Figure 4. The year variable is blurred by a small uniform random variable in order to get a better idea of the density of data.

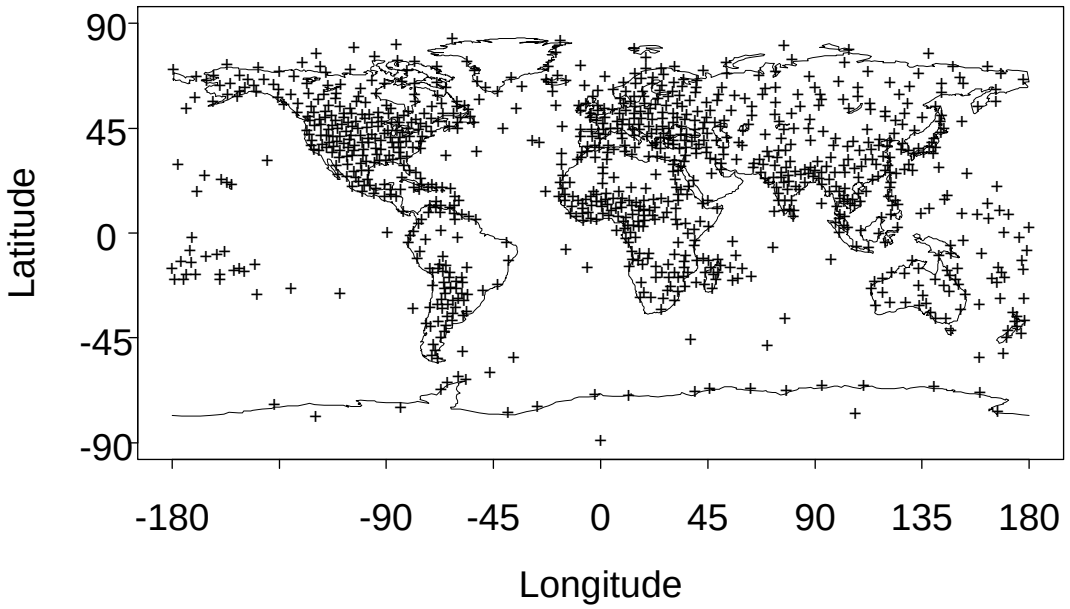

Figure 3: *The distribution of the 1000 stations used in our analysis.*

3.2 Smoothing spline estimates for a single year

3.2.1 SS estimates and BLUP estimates

Consider a variable defined over the sphere, for example, winter mean temperature. Suppose we have noisy data at some locations,

$$y_i = f(P_i) + \epsilon_i,\ i = 1, 2, ..., n \tag{3.2.1}$$

where $P_i \in \mathcal{S}$, the sphere. ϵ_i represents a "noise" term which contains not only the measurement error in record taking but also the representation error which is in connection with the density of data points. Hence f represents a smoothed version of the actual temperature field. Its smoothness depends on the resolution of data points. In other words, the denser those data points are, the smaller the area represented by f's value at one point is.

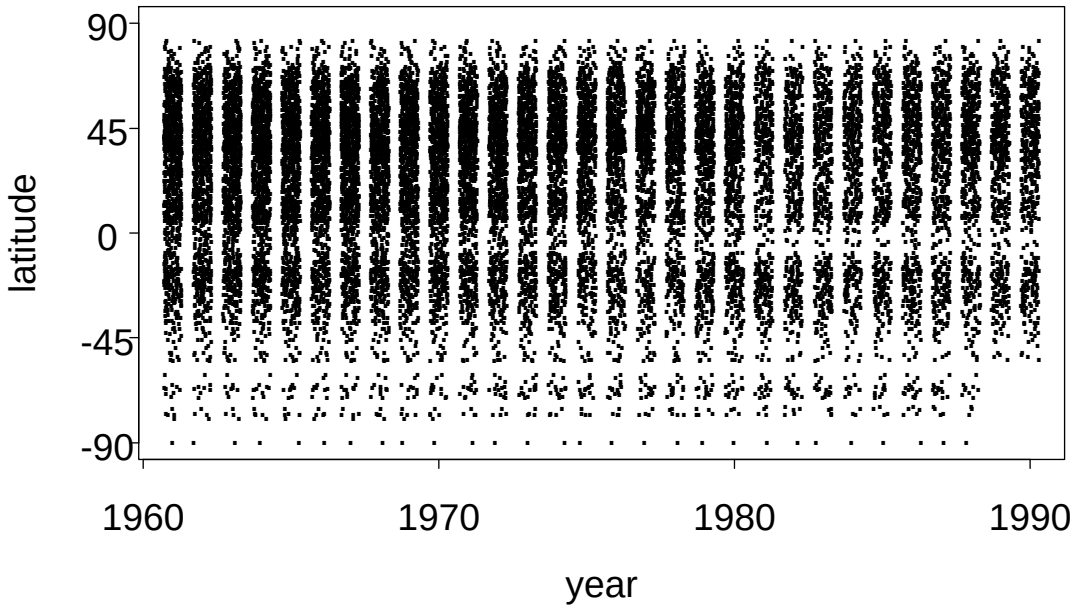

Figure 4: *The missing pattern in the 1000 stations. The year variable is blurred by a small uniform random variable.*

A smoothing spline estimate of f, denoted by f_θ to emphasize its dependence on a smoothing parameter θ, is defined as the minimizer of

$$\sum_{i=1}^{n}(y_i - f(P_i))^2 + \frac{1}{\theta}J(f) \tag{3.2.2}$$

over a reproducing kernel Hilbert space with $1 + R$ as its reproducing kernel, where R is defined in (2.1.41). J is the semi-norm corresponding to R in this space. J is topologically equivalent to the integrated squared Laplacian on the sphere, $\int_{\mathcal{S}}(\Delta f)^2 dP$, as defined in Section 2.1. The smoothing parameter, θ, controls the smoothness of f_θ besides R.

As in chapter 2, it can be proved that the solution to the above problem has a representation

$$f_\theta(P) = d + \theta\sum_{i=1}^{n} c_i R(P, P_i), \tag{3.2.3}$$

where d and c are the solutions to the following linear system:

$$\begin{cases} 1^T c &= 0 \\ (\theta Q + I)c &= y - d1. \end{cases} \tag{3.2.4}$$

It can be easily derived from (3.2.4) (see (2.1.36)) that

$$\begin{cases} d &= 1^T(\theta Q + I)^{-1} y / 1^T(\theta Q + I)^{-1} 1 \\ c &= (\theta Q + I)^{-1}(y - d1). \end{cases} \tag{3.2.5}$$

It is not difficult to verify that

$$\int_{\mathcal{S}} R(P, P')dP = 0, \text{ for any } P' \in \mathcal{S}, \tag{3.2.6}$$

hence $\int_{\mathcal{S}} f_\theta(P)dP / \int_{\mathcal{S}} 1dP = d$. That is, d is the global mean of f_θ.

We can also integrate f_θ over a region, say $\mathcal{K} \subset \mathcal{S}$, to get an estimate of the average temperature in that region. It turns out that this is the same as what Vinnikov et. al. (1990) called the "statistical optimal averaging" estimate (also called "Best Linear Unbiased Prediction" (BLUP) in many statistical references), even though these two estimates result from two different approaches. Vinnikov et. al. (1990) assume that f is a random field over the sphere with a constant mean , say C, and a covariance function $R(P, P')$. They also assume that $\{\epsilon_i\}$ are independent random variables such that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. $\{\epsilon_i\}$ are assumed to be independent of f as well. Then the mean squared error of predicting the mean of f over a region $\mathcal{K} \subset \mathcal{S}$, $\int_{\mathcal{K}} fdP/b$ where $b := \int_{\mathcal{K}} 1dP$, by a linear combination of observed data is

$$\begin{aligned} MSE &= E(\int_{\mathcal{K}} fdP/b - \sum_{i=1}^n p_i y_i)^2 \\ &= E(\int_{\mathcal{K}} fdP/b - \sum_{i=1}^n p_i f_i - \sum_{i=1}^n p_i \epsilon_i)^2 \\ &= Var(\int_{\mathcal{K}} fdP/b) + Var(\sum_{i=1}^n p_i f_i) + Var(\sum_{i=1}^n p_i \epsilon_i) \\ &\quad -2Cov(\int_{\mathcal{K}} fdP/b, \sum_{i=1}^n p_i f_i) + (E\int_{\mathcal{K}} fdP/b - \sum_{i=1}^n p_i Ef_i)^2 \\ &= [Var(\int_{\mathcal{K}} fdP/b) + \sum_i \sum_j p_i p_j R(P_i, P_j) + \sum_{i=1}^n p_i^2 \sigma^2 \end{aligned}$$

$$
\begin{aligned}
& -2\sum_{i=1}^{n} p_i\Omega_i] + [C^2(1-\sum_{i=1}^{n} p_i)^2] \\
= \; & [variance] + [bias], \qquad (3.2.7)
\end{aligned}
$$

where

$$\Omega_i = Cov(\int_{\mathcal{S}} f dP/b, f_i) = \int_{\mathcal{S}} R(P, P_i) dP/b.$$

They restrict estimators to unbiased ones, i.e. require that the "bias" term in (3.2.7) equals to zero. Hence coefficients $\{p_i\}$ must satisfy:

$$\sum_{i=1}^{n} p_i = 1. \qquad (3.2.8)$$

Under the condition (3.2.8), the minimizer of the MSE which is the same as the "variance" term now is

$$p = (Q+\sigma^2 I)^{-1}\Omega + (Q+\sigma^2 I)^{-1}1\frac{1-1^T(Q+\sigma^2 I)^{-1}\Omega}{1^T(Q+\sigma^2 I)^{-1}1}, \qquad (3.2.9)$$

hence the estimate of f's mean over region $\mathcal{K}$ is

$$
\begin{aligned}
y^T p &= y^T(Q+\delta^2 I)^{-1}\Omega + y^T(Q+\delta^2 I)^{-1}1\frac{1-1^T(Q+\delta^2 I)^{-1}\Omega}{1^T(Q+\delta^2 I)^{-1}1} \\
&= \frac{y^T(Q+\delta^2 I)^{-1}1}{1^T(Q+\delta^2 I)^{-1}1} \\
&\quad +(y - \frac{y^T(Q+\delta^2 I)^{-1}1}{1^T(Q+\delta^2 I)^{-1}1}1)^T(Q+\delta^2 I)^{-1}\Omega \qquad (3.2.10)
\end{aligned}
$$

which is exactly $\int_{\mathcal{K}} f_\theta dP / \int_{\mathcal{K}} 1 dP$ with $\theta = 1/\sigma^2$.

In Vinnikov et. al. (1990), they use an empirically estimated $R(P, P')$. As a matter of fact, what they have used is not exactly a covariance function, since it is not positive semi-definite. The discontinuity in its derivative is also a undesirable property. Estimating covariance functions, especially non-homogeneous ones, from empirical data is a very difficult problem. See Sampson and Guttorp (1992) for an example in this direction. For the data set we chose to use, we concluded, after a few attempts that a reasonable homogeneous covariance function is better than or at least as good as a very crudely estimated non-homogeneous one. We note that given any covariance function, homogeneous or not, our method is still applicable.

3.2.2 Spatial sampling difference and anomalies

Applying the SS estimate technique to each year's records in the period of 1961-1990, we get a sequence of global averages of winter temperature. A plot of these averages is shown in Figure 5. An easily seen feature of this sequence

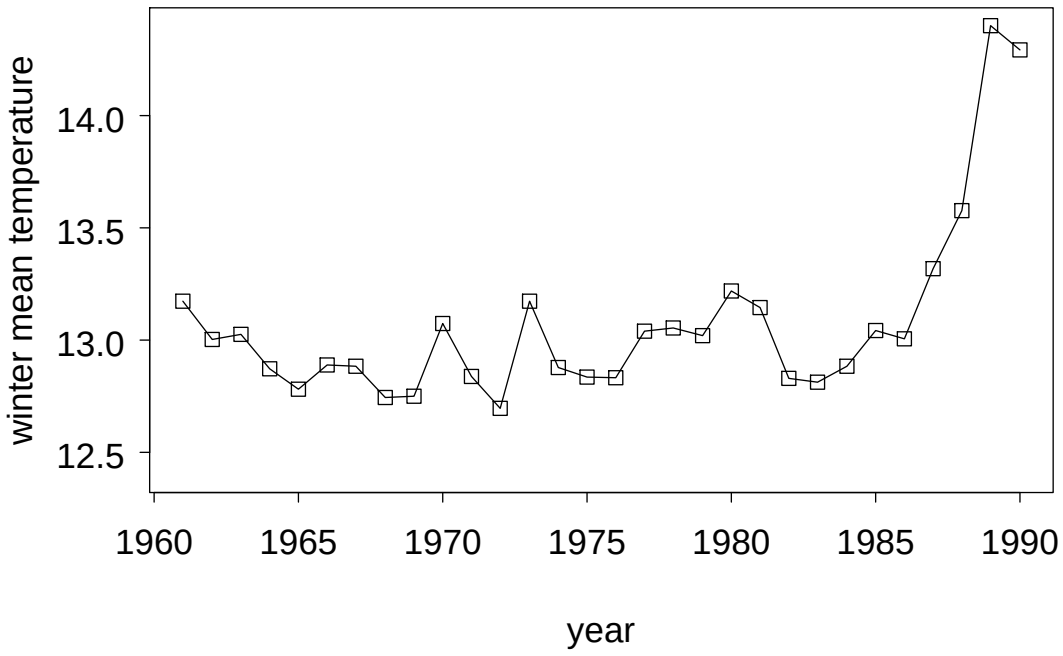

Figure 5: *Global average winter temperatures (^oC) based on yearly fits to raw data. Grand mean temperature is* $13.07(^oC)$*, the linear trend coefficient over a 30 year period is* $.025(^oC)/year$.

is the outstanding high values of the last two years. If we hence conclude that we have seen a dramatic increase of winter temperature in the last two years of 80's, then we have been misled by the bias resulting from the spatial sampling difference (or, equivalently, unbalanced time coverage). The records for the Antarctic region end in 1988 (see Figure 4). Obviously this abrupt increase of winter temperature in the last two years is mainly because of the lack of data in the Antarctic region where it is much colder than most other regions of the world.

In order to correct the bias resulting from the spatial sampling difference, many previous studies have chosen anomalies, instead of raw temperature records, as input. An anomaly is defined as a difference between a temperature record and the average temperature over a specified reference period. Choosing the average over 1961-1990 as the reference period, we get a sequence of average temperature anomalies shown in Figure 6. It is quite clear that the outlying feature of the last two years in Figure 5 disappeared.

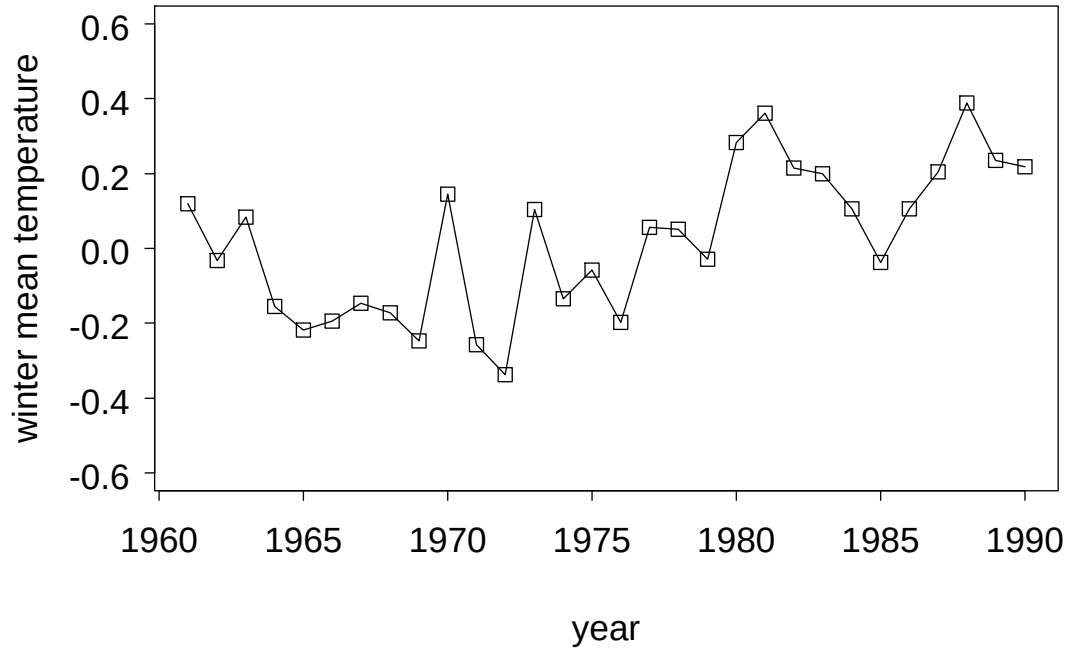

Figure 6: *Global average winter temperature anomalies (^oC) based on yearly fits to anomalies. The grand mean anomaly is* $.02(^oC)$ *and the linear trend coefficient over the 30 year period is* $.014(^oC)$*/year.*

The reason for the effectiveness of using anomalies to correct the bias resulting from the spatial sampling difference can be easily explained by the data decomposition of (2.1.1) and (2.1.6).

Since an observation is

$$y(x,P) = d_1 + d_2\phi(x) + g_1(x) + g_2(P) + g_{\phi,2}(P)\phi(x) + g_{12}(x,P) + \epsilon,$$

considering (2.1.7), the station mean over the same period is

$$\overline{y}(P) := \frac{1}{n_1}\sum_{x=1}^{n_1} y(x,P) \simeq d_1 + g_2(P). \tag{3.2.11}$$

The approximate equality in (3.2.11) is because that the records of some years may be missing when calculating $\sum y(x,P)$, and $\sum_{x=1}^{n_1}\epsilon_x$ is only approximately zero. Therefore the anomaly is

$$y(x,P) - \overline{y}(P) \simeq d_2\phi(x) + g_1(x) + g_{\phi,2}(P)\phi(x) + g_{12}(x,P). \tag{3.2.12}$$

Now it is clear that the locational difference across years in $g_2(P)$ does not affect the anomaly. However, the locational differences in the last two terms of (3.2.12) still do. The only case in which using anomalies will eliminate any bias resulting from spatial sampling difference is when we are certain that the last two terms in (3.2.12) are not significant, i.e. we know in advance that an additive model:

$$y(x,P) = d_1 + d_2\phi(x) + g_1(x) + g_2(P) + \epsilon, \tag{3.2.13}$$

is adequate. In our application here, we know that not only the average temperatures in different locations ($g_2(P)$) can be different, the change trends across years at different locations ($g_{\phi,2}(P)$, linear change trend coefficient, and $g_{12}(x,P)$, the other change) can be significantly different too. Some locations may have an increase, others may have a smaller increase or even a decrease. See Hergel et. al. (1995)'s Figure 2. This makes the last two terms un-negligible when considering the bias resulting from spatial sampling differences.

Having pointed this out, we would like to make clear that it is true that the locational difference in $g_2(P)$ is the most prominent one among the three terms in (2.1.6) involving P. The locational difference in the average temperatures (in a range of $(-40^oC, 40^oC)$) is much larger than the locational difference in the changes of temperature (a few degrees (^oC)). Therefore the anomaly approach has eliminated most bias resulting from locational differences. This is probably one of the reasons for its satisfactory use so far.

In our approach described in the following section, we fit raw temperature records instead of anomalies directly. By choosing appropriate averaging, we can correct the bias resulting from the locational difference in both $g_2(P)$ and

the other two terms $g_{\phi,2}(P)$ and $g_{12}(x,P)$. A global average temperature history delineated using this approach is shown in Figure 8. We can see that the pattern shown there is very similar to the one shown in Figure 6 which is obtained using anomalies. This is obviously an evidence that our approach has a similar ability to correct the bias resulting from spatial sampling difference as the anomaly approach does. Since we consider simultaneously all three terms in (2.1.6) involving P, it is reasonable to expect that our approach will correct the bias resulting from the other two terms too.

3.3 Smoothing spline model for multiple years

Now consider a more complicated model than (3.2.1). Suppose we have some winter mean temperature data at certain combinations of year and location,

$$y_i = f(x_i, P_i) + \epsilon_i,\ i = 1, 2, ..., n, \tag{3.3.1}$$

where $x_i \in \{1, 2, ..., n_1\}$ and $P_i \in \mathcal{S}$, the sphere. We are not only interested in the "signal" f itself (see the interpretation after (3.2.1)) but also its certain component functions representing certain marginal signals.

Adopt the model discussed in the example of Chapter 2, and write f as a sum of its component functions:

$$f(x,P) = d_1 + d_2\phi(x) + g_1(x) + g_2(P) + g_{\phi,2}(P)\phi(x) + g_{12}(x,P), \tag{3.3.2}$$

where $x \in \{1, 2, ..., n_1\}$ and $P = (latitude, longitude) \in \mathcal{S}$, and ϕ is a known linear function $\phi(x) = x - (n_1+1)/2$. Condition (2.1.7) guarantees that representation (3.3.2) is unique.

Note that these component functions and their combinations are often of clear climatology interest. For example, d_1 is the grand mean temperature over both year and location; d_2 is the linear trend coefficient of global means; $d_1 + d_2\phi + g_1$ is the global mean temperature history; g_2, $g_{\phi,2}$ and g_{12} are locational adjustments to d_1, d_2 and g_1, respectively; $d_1 + g_2(P)$ is the average winter temperature at location P; and $d_2 + g_{\phi,2}(P)$ is the linear trend coefficient of winter temperatures at location P.

A smoothing spline estimate, f_θ, is defined as the minimizer of

$$\sum_{i=1}^{n}(y_i - f(x_i, P_i))^2 + \frac{1}{\theta_1}J_1(g_1) + \frac{1}{\theta_2}J_2(g_2) + \frac{1}{\theta_3}J_3(g_{\phi,2}) + \frac{1}{\theta_4}J_4(g_{12}) \tag{3.3.3}$$

See the example in Section 2.1 for the meanings of J's.

In Chapter 2, we have discussed a computational procedure for getting such a SS estimate given the smoothing parameters, the θ's. In the next subsection, we will discuss different ways to choose these smoothing parameters.

3.3.1 Choosing smoothing parameters

How to choose smoothing parameters (the θ's in (3.3.3)) is a very crucial issue here, because the choice affects the smoothing spline estimate to a great extent. For example, if we choose θ_3 and θ_4 to be very small, we will penalize $g_{\phi,2}\phi$ and g_{12} heavily when their corresponding semi-norm values are not zero. Therefore we will essentially make these two terms disappear in our model and adopt an additive model (3.2.13).

There are basically two types of techniques for choosing smoothing parameters. One consists of the so-called "objective" or "data-driven" methods such as cross-validation (CV), generalized cross-validation (GCV), and generalized maximum likelihood estimation (GMLE) (See Wahba (1990) Chapter 4). The other consists of "subjective" methods. This category includes actually quite different types of techniques. For example, we could examine estimates corresponding to different choices of smoothing parameters to see which one is more consistent with our prior (subject) knowledge about what the fit should look like. We may also compute for each choice of smoothing parameters an estimate of the standard deviation of the observation, then compare it with our prior knowledge about the size of such observation "error". We may also use the past data to estimate these parameters. This is exactly Vinnikov et. al. (1990)'s approach for deciding both their smoothing parameter and covariance function. Finally, we may just want to make a choice basing on our subjective decision about how much smoothing we want, e.g. for visual enhancement. In general, these subjective criteria rarely give us a precise choice of smoothing parameters, but still they are very important in guiding us, and are even sufficient for our needs in many applications. It is also important to keep these criteria in mind even when we use "data-driven" criteria since so-called "objective" methods may give us misleading results also, not to mention that some important information is very hard to be formulated into "objective" criteria.

In our particular application here, we decide to use a "subjective" method

to choose θ_1 and θ_2, and an "objective" method to choose θ_3 and θ_4. The main reason is because of the large computational demand of choosing all four θ's by an "objective" method. Another reason is that we have a relatively clearer idea about how much smoothing should be done to g_1 and g_2. As a matter of fact, we want little smoothing done to them. A way to relate this information to a smoothing parameter is through the "degrees of freedom" of the corresponding marginal smoother. The usual definition of the "degrees of freedom" in smoothing spline estimates (see Wahba (1990)) is $tr(A(\theta))$, where $A(\theta)$ is the influence matrix defined by $(f_\theta(t_1), \cdots, f_\theta(t_n))^T = A(\theta)y$. This concept can be readily generalized to marginal smoothers, i.e. the "degrees of freedom" for a marginal smoother $S_\alpha(\theta_\alpha) = (Q_\alpha + \frac{1}{\theta_\alpha} I)^{-1} Q_\alpha$ (see (2.2.4)) is defined as $tr(S_\alpha)$. It is not difficult to see that the maximum degrees of freedom for $S_1(\theta_1)$ is $(n_1 - 2)$, n_2 for $S_2(\theta_2)$, n_2 for $S_3(\theta_3)$, and $(n_1 - 2)n_2$ for $S_4(\theta_4)$. (It is natural that they, together with 2 degrees of freedom for the parametric part, do not add up to the maximum overall degrees of freedom, $n_1 n_2$, because they are not independent.) To choose θ_1 and θ_2 in such a way that little smoothing is done to g_1 and g_2, we just choose them so that their corresponding degrees of freedom are close to their maximum values. The degrees of freedom for S_3 and S_4 are also useful to help us set a preliminary searching range of θ_3 and θ_4 when we use an "objective" method to choose them.

A commonly used "data-driven" method is to choose θ's minimizing GCV score which is defined as

$$V(\theta) = \frac{\|y - \hat{f}\|^2}{(tr(I - A(\theta)))^2}, \tag{3.3.4}$$

where $\hat{f} = (f_\theta(t_1), \cdots, f_\theta(t_n))^T = A(\theta)y$. The numerator $\|y - \hat{f}\|^2$, the residual sum of squares, can be easily computed after we get the estimate of the function. But the denominator $(tr(I - A(\theta)))^2$ is much more difficult to compute. Usually when the data size, hence the size of the matrix $A(\theta)$, is not very large, we can compute this $V(\theta)$ for any θ easily after a sequence of matrix decompositions are done (See (4.6.2) of Wahba (1990)). When the data size is very large as in our case here, we cannot use this matrix decomposition method anymore. Instead of computing $tr(I - A(\theta))$ exactly, we use an approximation called "randomized

GCV" (RGCV)

$$RGCV(\theta) := \frac{\|y - \hat{f}\|^2}{[\xi^T(\xi - \hat{f}(\xi))]^2}, \tag{3.3.5}$$

where ξ is a standard multivariate normal random vector with the same length as the data vector, and $\hat{f}(\xi)$ is the smoothing spline estimate when the data vector y is substituted by ξ. (See Girard (1989, 1991) and Hutchinson (1989)). The reason behind this approximate GCV criterion is that $E[\xi^T(\xi - \hat{f}(\xi))] = tr(I - A(\theta))$, i.e. $[\xi^T(\xi - \hat{f}(\xi))]$ is an unbiased estimate of $tr(I - A(\theta))$. In order to minimize the variation induced by ξ, it is better to use the same ξ for all choices of θ. Both $\hat{f}$ and $\hat{f}(\xi)$ can be computed using the same procedure discussed in Chapter 2.

Considering that $y - \hat{f} = (I - A(\theta))y$, $\xi - \hat{f}(\xi) = (I - A(\theta))\xi$, and a representation of $I - A(\theta)$ in Wahba (1990, (1.3.23)), it is straightforward to verify that

$$\frac{\partial RGCV(\theta)}{\partial \theta_\beta} = \frac{2(u^T u)(v^T Q_\beta v) - 2(w^T Q_\beta u)(\xi^T v)}{(\xi^T v)^3}, \tag{3.3.6}$$

for any $\beta = 1, 2, 3, 4$, where

$$u = (I - A(\theta))y, \tag{3.3.7}$$

$$v = (I - A(\theta))\xi, \tag{3.3.8}$$

$$w = (I - A(\theta))u, \tag{3.3.9}$$

and $Q_\beta = (R_\beta(x_i, P_i; x_j, P_j))_{i,j=1}^n$,

We need u and v to compute $RGCV$ anyway; with one more fit with y replaced by u, we can get all partial derivatives of $RGCV$. This information may be used in minimizing $RGCV$.

3.3.2 Results

Taking the approach described in Section 3.3.1, we choose $\theta_1 = 10^{-0.1}$ and $\theta_2 = 10^{4.5}$ which correspond to 27.8 degrees of freedom for $S_1(\theta_1)$, 989.8 for $S_2(\theta_2)$. With little smoothing done to g_1, our results should be comparable to those of other studies where only single year data are used to calculate a global average in any particular year. We choose θ_2 this large because the smaller θ_2 is,

the smoother g_2 is, and the closer our estimated global mean history is to what was obtained by the naive single year means of raw data (Figure 5). The reason for this is the following. If θ_2 is so small that g_2 is practically constant, then we essentially disregard the potential bias resulting from the locational difference in g_2 totally. The smaller θ_2 is, the more bias we disregard.

With θ_1 and θ_2 chosen as above, we choose θ_3 and θ_4 by a crude grid search, We first set some preliminary limits for them by the tool of the degrees of freedom of their corresponding marginal smoothers. For θ_3, the limits are $10^{.5}$ and $10^{1.5}$ corresponding to 565.4 and 890.5 degrees of freedom (the maximum is 1000) respectively. For θ_4, the limits are $10^{2.8}$ and $10^{4.4}$ corresponding to 7052.6 and 17138.2 degrees of freedom (the maximum is 28000, but the total number of observations is 20910) respectively. Part of the search results are given in Table 2. A (local) minimum in RGCV gives us a choice of $\theta_3 = 10^{1.25}$ and $\theta_4 = 10^{4.1}$ which correspond to 831.1 degrees of freedom for S_3 and 14860.5 degrees of freedom for S_4 respectively.

	$log10(\theta_3)$				
$log10(\theta_4)$	1.5	1.25	1	.75	.5
4.4		.63452	.63617		.64697
4.1	.92752	.62737(*)	.62747	.62909	.63298
3.8		.63958	.63905		.64201

Table 2: *RGCV for the 1000 station data set.* $log10(\theta_1)$ *and* $log10(\theta_2)$ *are fixed at* $-.1$ *and* 4.5 *respectively. (*) indicates a local minimum.*

Some results based on a fit with smoothing parameters chosen as above are shown in Figures 7-10. The estimated standard deviation of ϵ, $\hat{\sigma}$, by the formula of Wahba (1990, Section 4.7):

$$\hat{\sigma}^2 = \frac{\|y - \hat{f}\|^2}{tr(I - A(\theta))} \simeq \frac{\|y - \hat{f}\|^2}{\xi(\xi - \hat{f}(\xi))} \tag{3.3.10}$$

is $.49^oC$ which is a little bit larger than what a typical measurement error of mean temperature is expected to be. This is reasonable considering the fact that here ϵ contains not just the measurement error. If σ is too large, e.g. larger than 1^oC, then we may suspect too much smoothing has been done to the data. This is how a subjective criterion is used. A comparison of fitted values and observations at two arbitrarily selected stations is plotted in Figure 7.

From Figure 8, we see that in the global mean winter temperatures, there exists an overall cooling trend in the early sixties and an overall warming trend from the seventies on. The overall linear trend over these 30 years is about $.011^oC/year$.

In Figure 9, we see a familiar pattern of winter mean temperature across the world. In Figure 10, we see that most of the European area has a warming trend (positive coefficient) except the eastern Mediterranean region and a large area of the North Atlantic, including Greenland. A cooling trend has been observed in part of Africa and America also. Strong warming trends have been noticed in parts of Siberia and North America. Such a local trend pattern gives us more information about what has happened to the climate in the past, and it may be used in a comparison with the predications of climate models in order to verify these models.

The whole history of these 30 year winter temperature anomaly based on our SS fit is made into a movie which can be accessed at `ftp://ftp.stat.wisc.edu/pub/wahba/theses/luo.movie`. Viewing such a movie may help climatologists identify important patterns observed in the climate.

In order to see the effect of the number of stations used on the results, we did a similar analysis based on 500 stations. θ's are chosen in the same way as for the 1000 stations. θ_1 is chosen as 1 and θ_2 as $10^{3.8}$ which correspond to 27.7 degrees of freedom for $S_1(\theta_1)$ and 491.8 degrees of freedom for $S_2(\theta_2)$ (the maximum for S_2 here is 500) respectively. Then a crude grid search in RGCV gives us $\theta_3 = 10^{.5}$ and $\theta_4 = 10^{3.8}$ which correspond to a local minimum of RGCV. A part of the search results is given in Table 3. These θ's correspond to 369.2 degrees of freedom for the marginal smoother $S_3(\theta_3)$ and 7864.2 for $S_4(\theta_4)$, respectively.

	$log10(\theta_3)$					
$log10(\theta_4)$	2	1.5	1	.5	0	-.5
4.2	1.0191	1.0211	1.0208	1.0260	1.0500	1.0923
3.8	0.9957	0.9950	0.9916	0.9900(*)	0.9979	1.0157
3.4	1.0407	1.0392	1.0346	1.0291	1.0300	1.0389

Table 3: *RGCV for the 500 station data set.* $log10(\theta_1)$ *and* $log10(\theta_2)$ *are fixed at* 0 *and* 3.8 *respectively. (*) indicates a local minimum.*

Some results based on this fit of the 500 station data set are shown in Figures 11-12. We see that even though there exist some discrepencies between them and the 1000 station data set results (Figure 8 and Figure 10), the general patterns are quite similar. The plot of linear trend coefficient of winter temperature using the 1000 stations has more details than its 500 station counterpart, but they agree in large patterns. A similar fit using only 250 stations results in much more different plots (compare Figure 13 with Figure 8 or 11). This suggests that a few hundred stations are probably the minimal number of stations for calculating reliable global mean temperatures. Of course these stations still have to be distributed as uniformly over the sphere as possible. As a matter of fact, with 500 stations chosen randomly from the original stations, hence more stations concentrated in Europe and North Ameria, a fit based on these stations is much more different than the results based on the 1000 stations.

3.3.3 Outliers and other diagnostics

The results shown in Section 3.3.2 are based on a corrected version of the original data from CDIAC. There are six places where we have found some possible typos and corrected them for the purposes of this study.[1] Our purpose here was to demonstrate the power of the method, not to criticize the data base. The corrections we made are documented below. These "typos" were found as a by-product of fitting smoothing spline models to the data (see Knight (1980) for a comparison of such a approach with others). For example, when a SS model (3.3.1-3) is fitted to a 500 station subset of the original version of the data, some residual plots resulting from this fit are shown in Figure 14.

From the QQ plot of residuals, we see that one observation has an extremely large residual. That observation turns out to belong to station (72.0N, 102.5E) in Hatanga/Khatanga of the former USSR. Its December temperature of 1980 is 28.8^oC as shown in the original database, while all other years' December temperatures during the 1951-1991 period range from -38.3^oC to -19.3^oC. Also the November and January records in the same year do not show any extreme pattern. Therefore we strongly suspect that 28.8 should be -28.8 and the record in the original data base results from a missing minus sign.

[1] As of June 17, 1996, the last time we visited this data base at CDIAC. However, this is not the latest data base and not the one used in the latest IPCC Report, see Nicholls at al (1996).

This observation's extreme outlying feature is so strong that it makes two of its neighboring observations in time (year 1979 and 1981), and another in location (station (68.5N, 112.4E), December) look like outliers too. They are the three observations with largest negative residuals. With Hatanga's December record corrected, however, they look just "normal".

From the plot of residual vs latitude, we notice two outliers in the southern hemisphere. Station (29.9S, 31.0E) in Durban of South Africa has a February temperature of 4.0^oC in 1983 in the original database, while all other February temperatures in the period of 1885-1991 are in the range $(22.0^oC, 25.9^oC)$. Again we suspect that this is a record with a typo. It should probably be 24.0 rather than 4.0. Station (39.0S, 68.0W) in Neuquen Aero of Argentina has a January temperature record of 2.3^oC in 1977, but all other January temperatures in the period of 1957-1991 are in the range $(20.7^oC, 25.4^oC)$. Hence we suspect that 2.3 should be 22.3.

Following are three other possible typos we have found through various smoothing spline fits to different subsets of the original data. Station (22.0S, 60.7W) in Mariscal Estigar of Paraguay has a December temperature of 38.1^oC in 1972 which might be 28.1^oC since all other December temperatures in the period of 1951-1991 range from 25.8^oC to 30.7^oC. Station (42.8N, 73.8W) in Albany of USA has a January temperature record of 9.6^oC in 1968 which might be -9.6^oC since all other January temperatures in the period of 1820-1991 range from -12.4^oC to 1.8^oC. Station (38.4N, 27.3E) in Izmir of Turkey has a February temperature of -7.0^oC in 1976 which might be 7.0^oC since all other February temperatures in the period of 1843-1991 range from 4.8^oC to 14.0^oC.

Of course these are only suspicions, no matter how strong they might be. Further examinations of original station records or comparisons with other records are needed to confirm that these records are really incorrect in the database. But it is certainly helpful to have these extreme observations pointed out for further examination. Plotting residuals from smoothing spline models in a QQ plot and against year, latitude, or longitude etc., has proved to be a useful tool in identifying these extreme cases.

Looking at the QQ plots resulting from various smoothing spline fits, we found that all of them are S-shaped, which indicates in the distribution of the residuals a heavier tail than that of a Gaussian distribution. In general, since seasonal temperatures such as winter temperatures considered here are

calculated by a series of averaging steps, we would expect a Gaussian or at least approximate Gaussian distribution in the residuals. The reason for the heavier tail is that these residuals are a mixture of more than one zero-mean Gaussian distributions of different variances. It can be proved easily that any mixture of this kind will have a larger kurtosis than a Gaussian distribution's kurtosis. That is to say that it will have a heavier tail. In our data set, the variation of temperature in different locations can be very different. For example the variations at stations in the central continental regions may be quite different from those at stations in the coastal regions. The difference in the variation of temperature may be also due to the difference in the altitude, or the difference in the latitude, and so forth. This suggests that we should not treat the stochastic model used in deriving BLUP estimates too literally. The variance of ϵ in Model (3.3.1-3) is not constant. However smoothing spline estimates can still be justified through penalized least square estimates instead of penalized likelihood estimates. In practice, as long as there exists little positive dependence among ϵ's, smoothing spline estimates with smoothing parameters chosen by a GCV kind criterion work just fine even when the variance of ϵ is not constant.

3.3.4 Extension to more variables

If, besides year and location, we want to include other variables, e.g. season, into our model, the computational procedures discussed in Chapter 2 can be easily extended to deal with such cases. Even though the model may get very complicated, it will work fine as long as the newly added variables have a uniform design.

We illustrate such an extension through a model of monthly temperature. First, define averaging operators in each of three variables: year(x), location(P), month (m) (actually two averaging operators for variable year since we want to single out its linear trend as well as its mean):

$$(\mathcal{E}_x f)(x,P,m) = \frac{1}{n_1}\sum_{x=1}^{n_1} f(x), \tag{3.3.11}$$

$$(\mathcal{E}_P f)(x,P,m) = \int_{\mathcal{S}} f(P)dP/4\pi, \tag{3.3.12}$$

$$(\mathcal{E}_m f)(x,P,m) = \frac{1}{12}\sum_{m=1}^{12} f(m), \tag{3.3.13}$$

$$(\mathcal{E}'_x f)(x,P,m) \;=\; \frac{f(n_1,P,m)-f(1,P,m)}{\phi(n_1)-\phi(1)}\phi(x). \tag{3.3.14}$$

Then we have a unique decomposition of f:

$$\begin{aligned}
f &= [\mathcal{E}_x+\mathcal{E}'_x+(I-\mathcal{E}_x-\mathcal{E}'_x)][\mathcal{E}_P+(I-\mathcal{E}_P)][\mathcal{E}_m+(I-\mathcal{E}_m)]f\\
&= \mathcal{E}_x\mathcal{E}_P\mathcal{E}_m f+\mathcal{E}'_x(I-\mathcal{E}_P)\mathcal{E}_m f+(I-\mathcal{E}_x-\mathcal{E}'_x)(I-\mathcal{E}_P)\mathcal{E}_m f\\
&\quad+\mathcal{E}_x(I-\mathcal{E}_P)\mathcal{E}_m f+\mathcal{E}'_x(I-\mathcal{E}_P)\mathcal{E}_m f+(I-\mathcal{E}_x-\mathcal{E}'_x)(I-\mathcal{E}_P)\mathcal{E}_m f\\
&\quad+\mathcal{E}_x\mathcal{E}_P(I-\mathcal{E}_m)f+\mathcal{E}'_x\mathcal{E}_P(I-\mathcal{E}_m)f+(I-\mathcal{E}_x-\mathcal{E}'_x)\mathcal{E}_P(I-\mathcal{E}_m)f\\
&\quad+\mathcal{E}_x(I-\mathcal{E}_P)(I-\mathcal{E}_m)f+\mathcal{E}'_x(I-\mathcal{E}_P)(I-\mathcal{E}_m)f\\
&\qquad+(I-\mathcal{E}_x-\mathcal{E}'_x)(I-\mathcal{E}_P)(I-\mathcal{E}_m)f\\
&= d_1+d_2\phi(x)+g_1(x)\\
&\quad+g_2(P)+g_{\phi,2}(P)\phi(x)+g_{12}(x,P)\\
&\quad+g_3(m)+g_{\phi,3}(m)\phi(x)+g_{13}(x,m)\\
&\quad+g_{23}(P,m)+g_{\phi,23}(P,m)\phi(x)+g_{123}(x,P,m),
\end{aligned} \tag{3.3.15}$$

where these components satisfy some side conditions similar to (2.1.7). These side conditions are also sufficient to make the decomposition uniquely defined.

A smoothing spline estimate is defined as the minimizer of

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i-f(x_i,P_i,m_i))^2+\frac{1}{\theta_1}J_1(g_1)\\
&+\frac{1}{\theta_2}J_2(g_2)+\frac{1}{\theta_3}J_3(g_{\phi,2})+\frac{1}{\theta_4}J_4(g_{12})\\
&+\frac{1}{\theta_5}J_5(g_3)+\frac{1}{\theta_6}J_6(g_{\phi,3})+\frac{1}{\theta_7}J_7(g_{13})\\
&+\frac{1}{\theta_8}J_8(g_{23})+\frac{1}{\theta_9}J_9(g_{\phi,23})+\frac{1}{\theta_{10}}J_{10}(g_{123}),
\end{aligned} \tag{3.3.16}$$

where J_5 and J_6 are the same and defined as

$$J(f) := \sum_{m=1}^{12}(f(m+1)-f(m))^2, \tag{3.3.17}$$

with $f(13) := f(1)$. This form of penalty is chosen because of the periodic nature of the variable month. Other J's are defined through the tensor-product structure of their corresponding Hilbert spaces.

Define

$$L := \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \cdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 1 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix}_{12\times 12} \tag{3.3.18}$$

Then $J(f) = f^T L^T L f$. The reproducing kernel for the variable month is $Q_m := (\tilde{R}_m(i,j))_{i,j=1}^{12} := (L^T L)^\dagger$, where $\dagger$ means the Moore-Penrose generalized inverse. Since $(L^T L)1 = 0$,

$$Q_m 1 = 0 \tag{3.3.19}$$

Table 4 shows the reproducing kernel matrices for the subspaces containing those component functions defined in (3.3.15). The projection operator for the parametric component is: $S_0 = S(S^T S)^{-1} S^T$, where $S = 1 \otimes (1\ \phi) \otimes 1$, and the marginal smoothing matrices are $S_\alpha = (Q_\alpha + \frac{1}{\theta_\alpha} I)^{-1} Q_\alpha$, for $\alpha = 1, 2, \cdots, 10$.

smoother S_α	component	r.k. matrix Q_α
S_1	g_1	$11^T \otimes Q_x \otimes 11^T$
S_2	g_2	$Q_P \otimes 11^T \otimes 11^T$
S_3	$g_{\phi,2}\phi$	$Q_P \otimes \phi\phi^T \otimes 11^T$
S_4	g_{12}	$Q_P \otimes Q_x \otimes 11^T$
S_5	g_3	$11^T \otimes 11^T \otimes Q_m$
S_6	$g_{\phi,3}\phi$	$11^T \otimes \phi\phi^T \otimes Q_m$
S_7	g_{13}	$11^T \otimes Q_x \otimes Q_m$
S_8	g_{23}	$Q_P \otimes 11^T \otimes Q_m$
S_9	$g_{\phi,23}$	$Q_P \otimes \phi\phi^T \otimes Q_m$
S_{10}	g_{123}	$Q_P \otimes Q_x \otimes Q_m$

Table 4: *The reproducing kernel matrices of the ten subspaces containing the ten nonparametric components in Model (3.3.15).*

Because of (3.3.19), all but a few products of $S_\alpha S_\beta$ for $\alpha \neq \beta$ are zero. Hence, the stationary equations which lead to the backfitting algorithm are

(after rearranging the order):

$$\begin{pmatrix} I & S_0 & S_0 & & & & & & & & \\ S_2 & I & 0 & & & & & & & & \\ S_3 & 0 & I & & & & & & & & \\ & & & I & S_1 & & & & & & \\ & & & S_4 & I & & & & & & \\ & & & & & I & S_5 & & & & \\ & & & & & S_8 & I & & & & \\ & & & & & & & I & S_6 & & \\ & & & & & & & S_9 & I & & \\ & & & & & & & & & I & S_7 \\ & & & & & & & & & S_{10} & I \end{pmatrix} \begin{pmatrix} f_0 \\ f_2 \\ f_3 \\ f_1 \\ f_4 \\ f_5 \\ f_8 \\ f_6 \\ f_9 \\ f_7 \\ f_{10} \end{pmatrix} = \begin{pmatrix} S_0 y \\ S_2 y \\ S_3 y \\ S_1 y \\ S_4 y \\ S_5 y \\ S_8 y \\ S_6 y \\ S_9 y \\ S_7 y \\ S_{10} y \end{pmatrix},$$

where blank spaces mean zero.

Therefore the backfitting algorithm is reduced into five smaller groups, each of which can be handled by the techniques discussed in Section 2.3.

3.4 Confidence intervals and simulation

In order to get some idea about the accuracy of our estimates in Section 3.3.2, we conduct a small simulation study. Due to computing time limitations, the 500 station subset of the data is chosen. Pretending the fitted functions in Section 3.3.2 to be the truth, generate 10 copies of bootstrap samples from Model (3.3.1), with ϵ_i's generated from a zero-mean normal pseudo random variable with a standard deviation .61, an estimation based on Formula (3.3.10). Then the same SS model with the same smoothing parameters as those used in Section 3.3.2 is fitted to each of these copies.

All 10 estimates of the global average winter temperature history are superimposed in Figure 15. This plot can be viewed as a confidence statement about the estimate in Figure 11. The width of the bundle of 10 estimates at one point can be treated as a measure of variation of the SS estimate at that point. 10 estimated grand global winter mean temperatures range in $(12.90^oC, 12.94^oC)$ with a mean 12.92^oC. 10 estimated linear trend coefficients range in $(.013^oC/year, .017^oC/year)$ with a mean $.015^oC/year$.

Similarly we can use the range of 10 estimates of the linear trend coefficient at any geographical point as a measure of the variation of the estimate in Figure 12. In Figure 16, the white areas are where there is a consistently estimated trend (either consistent warming trend or consistent cooling trend) in all 10 estimates. Therefore these are areas where the estimated trend in Figure 12 is more reliable. For example, the cooling trend in the North Atlantic and the warming trend in most of the European region are relatively more trustworthy. Of course the black areas include also those regions of the world where there was no linear trend at all over the period of 1961-1990. The black regions in Figure 16 also serve as a division of the world into warming areas and cooling areas.

We also generated a similar pseudo data copy for the 1000 station subset based on the results in Section 3.3.2. The fitted results using this pseudo data are shown in Figure 17-19. They can be compared with their counterparts in Figure 8-10 to get a rough idea about the accuracy of the estimates in Section 3.3.2 for the 1000 station subset of the data. The pattern observed in Figure 8 is relatively reliable, while the large features in Figure 9-10 are reliable too.

Of course the confidence statements above are in general underestimates of the variation inherited in the estimates of Section 3.3.2. We did not apply RGCV to the pseudo data, instead we just used the same smoothing parameters used in Section 3.3.2. The extra variation resulting from choosing smoothing parameters is not considered here. Nevertheless, these simulation results give us some idea about the accuracy of our estimates. See also the discussion in Wang (1994) for an interpretation of these bootstrap confidence intervals. Another way to formulate confidence statements is through "Bayesian" confidence intervals which will be discussed in Chapter 4. Unfortunately their computation is also quite demanding. We do not have any numerical results here to compare with these bootstrap results.

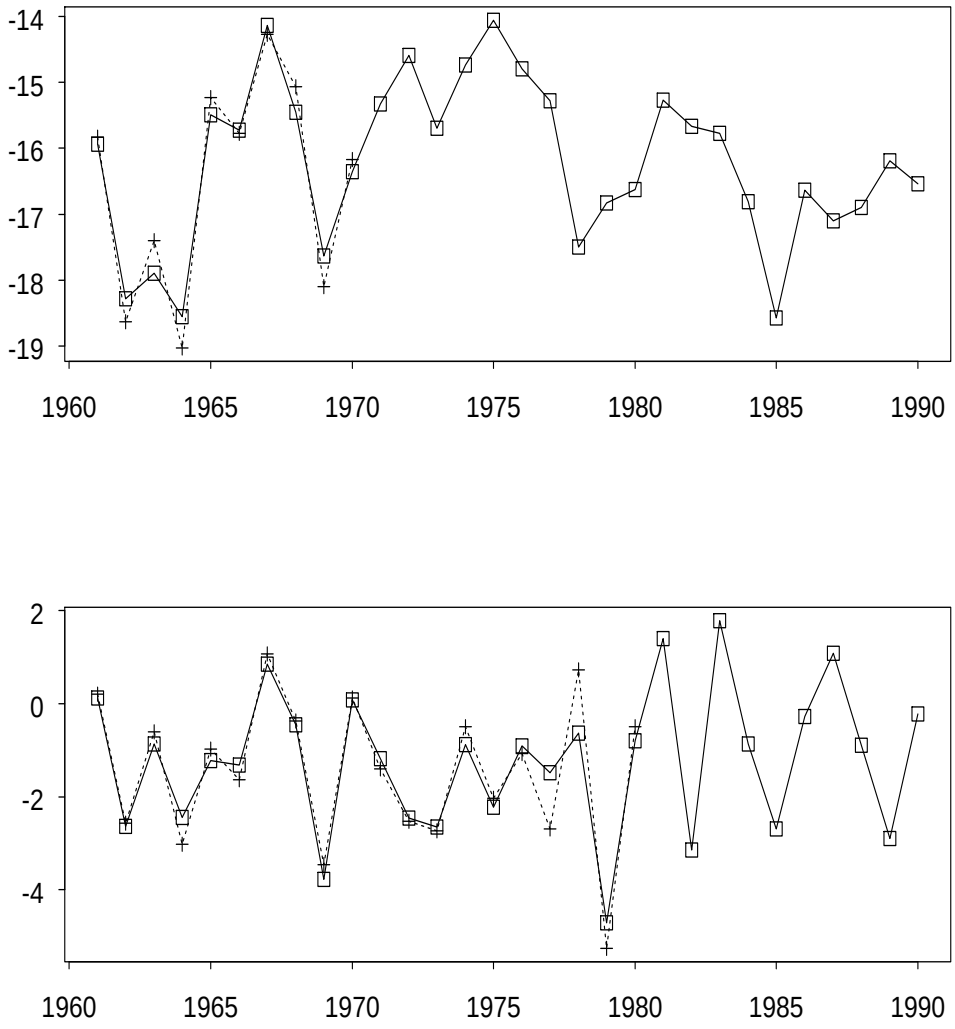

Figure 7: *Comparison of fitted values and observations at two arbitrary stations: (80 S, 119.5 W) and (45.6 N, 117.5 W). Squares are the fitted values and crosses are the observations.*

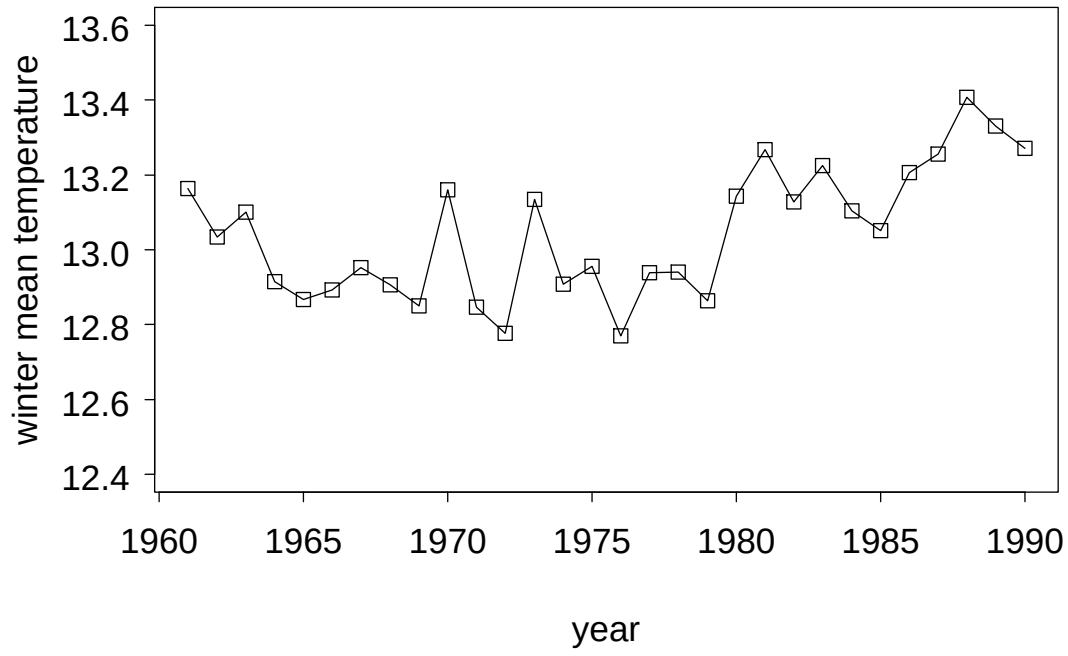

Figure 8: *Global average winter temperature (^oC) based on Model (3.3.1-3) using the 1000 stations. The grand mean temperature is* $13.0(^oC)$ *and the linear trend coefficient over the 30 year period is* $.011(^oC)/year$.

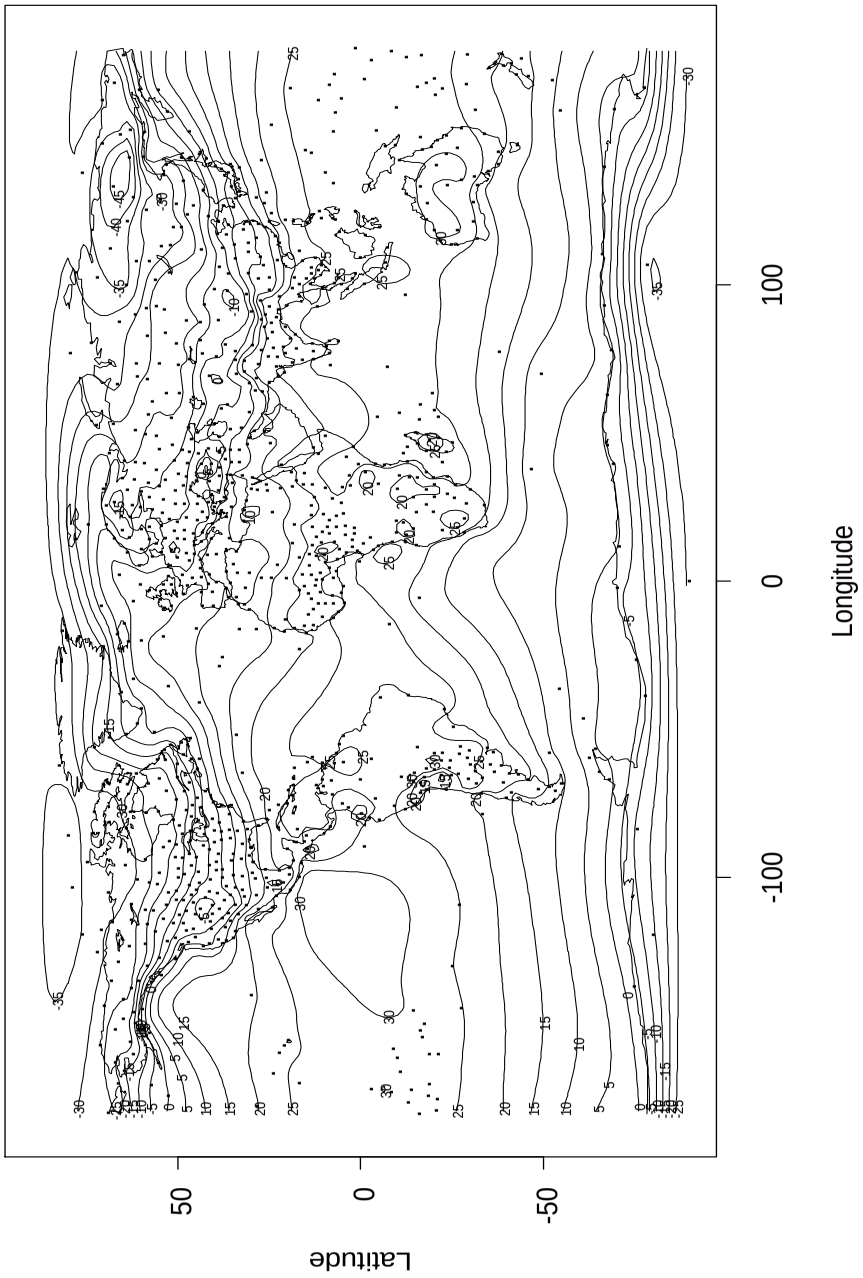

Figure 9: *Average winter temperature over the globe using the 1000 stations.*

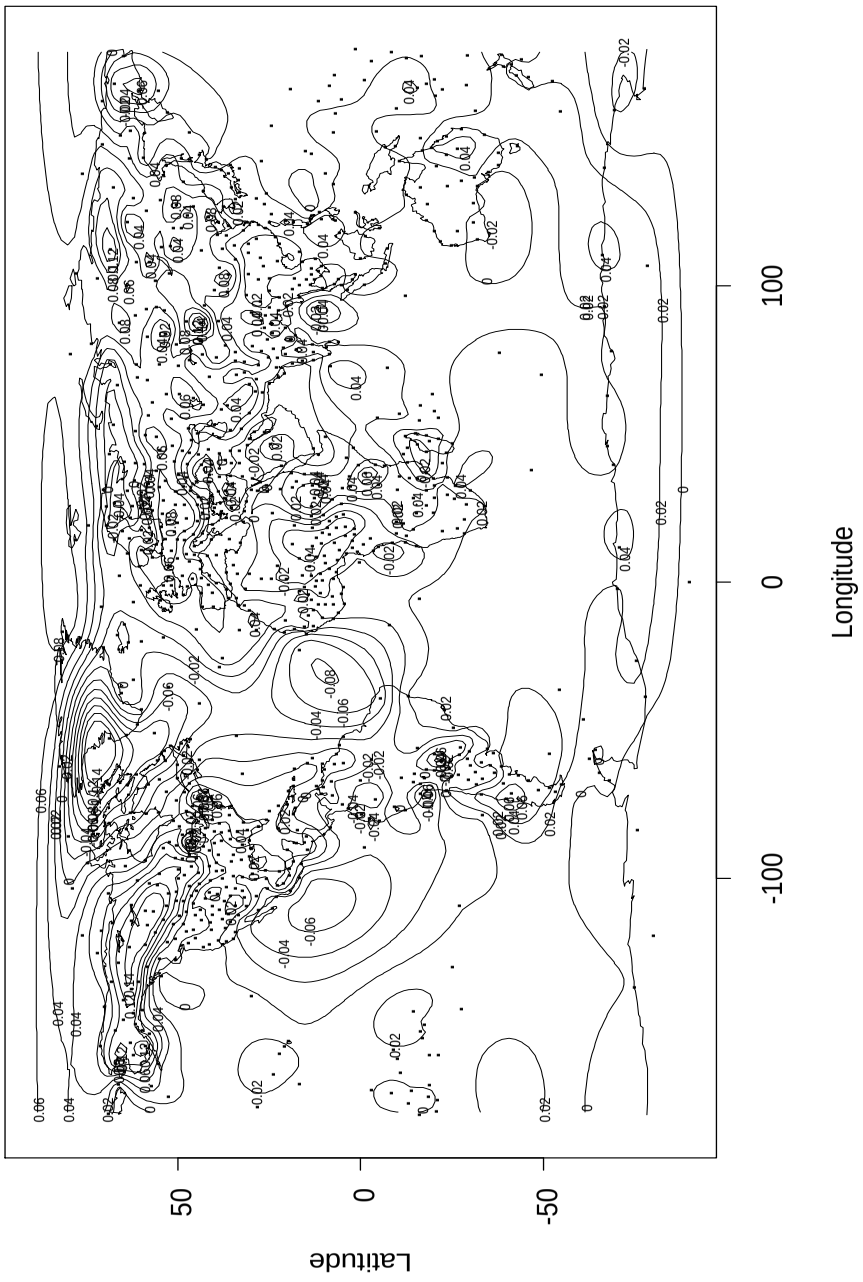

Figure 10: *Linear trend coefficient of winter temperature over the globe using the 1000 stations.*

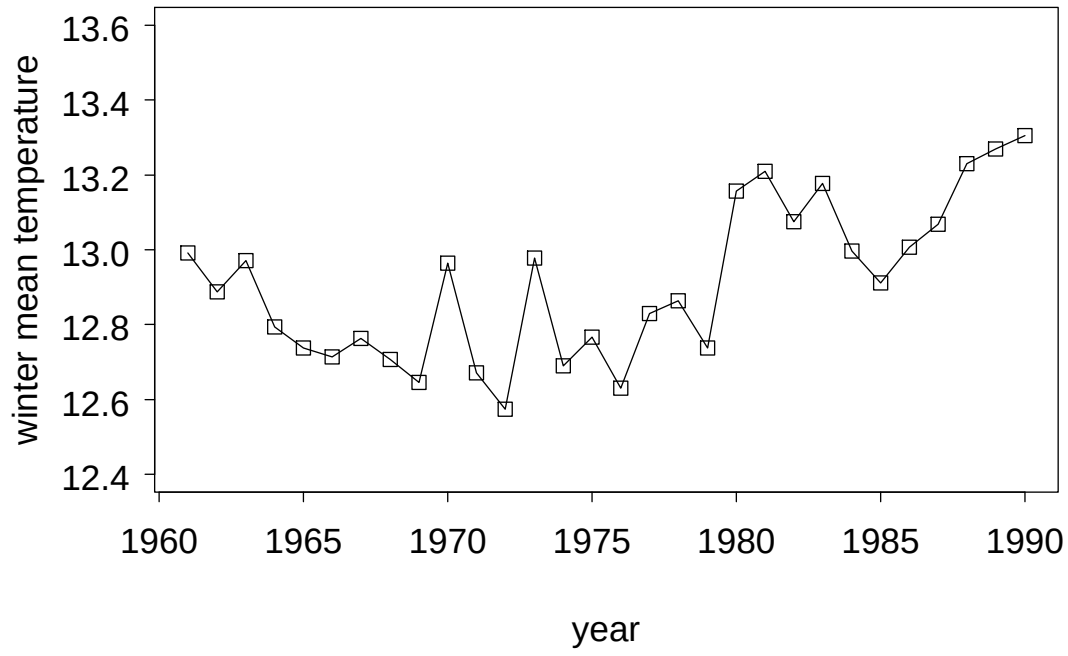

Figure 11: *Global average winter temperature (^oC) based on Model (3.3.1-3) using the 500 stations. The grand mean temperature is* $12.9(^oC)$ *and the linear trend coefficient over the 30 year period is* $.015(^oC)/year$.

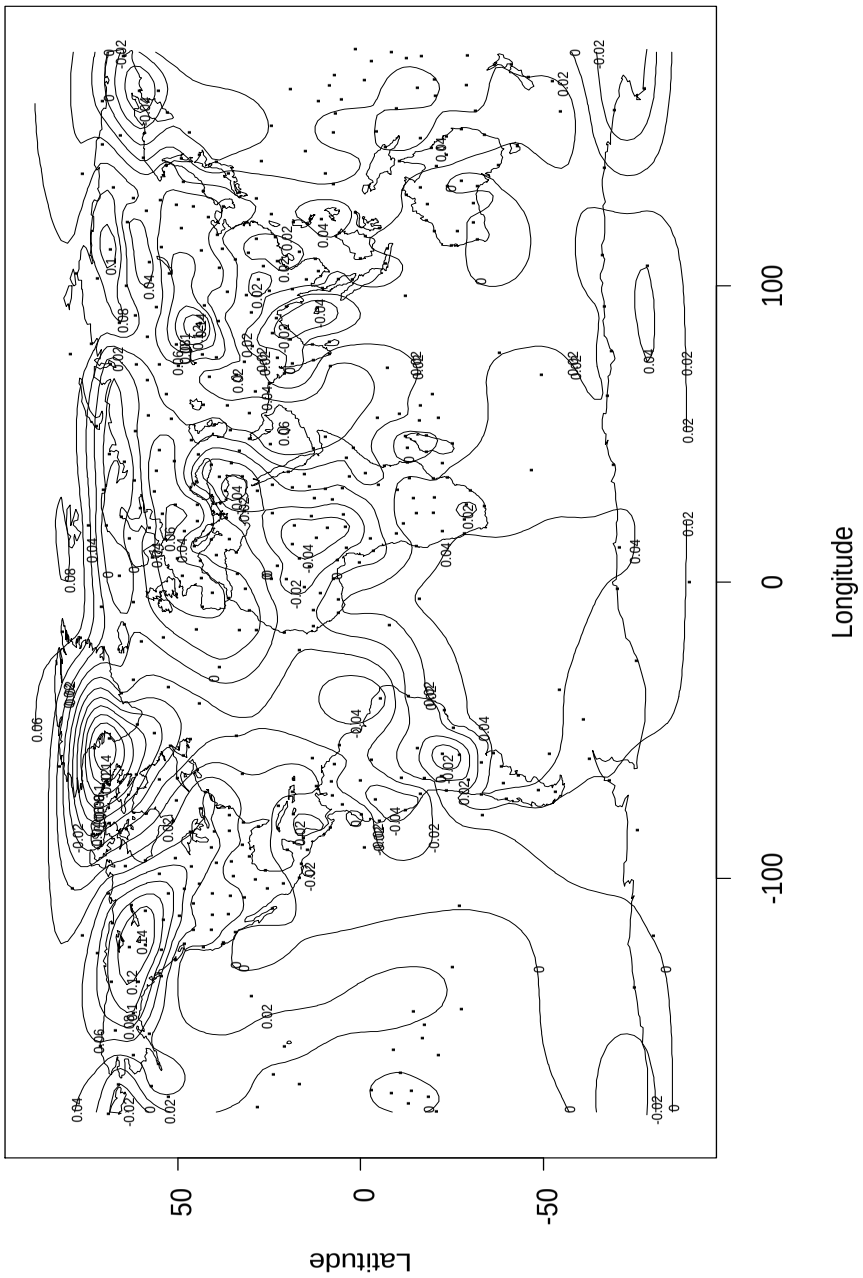

Figure 12: *Linear trend coefficient of winter temperature over the globe using the 500 stations.*

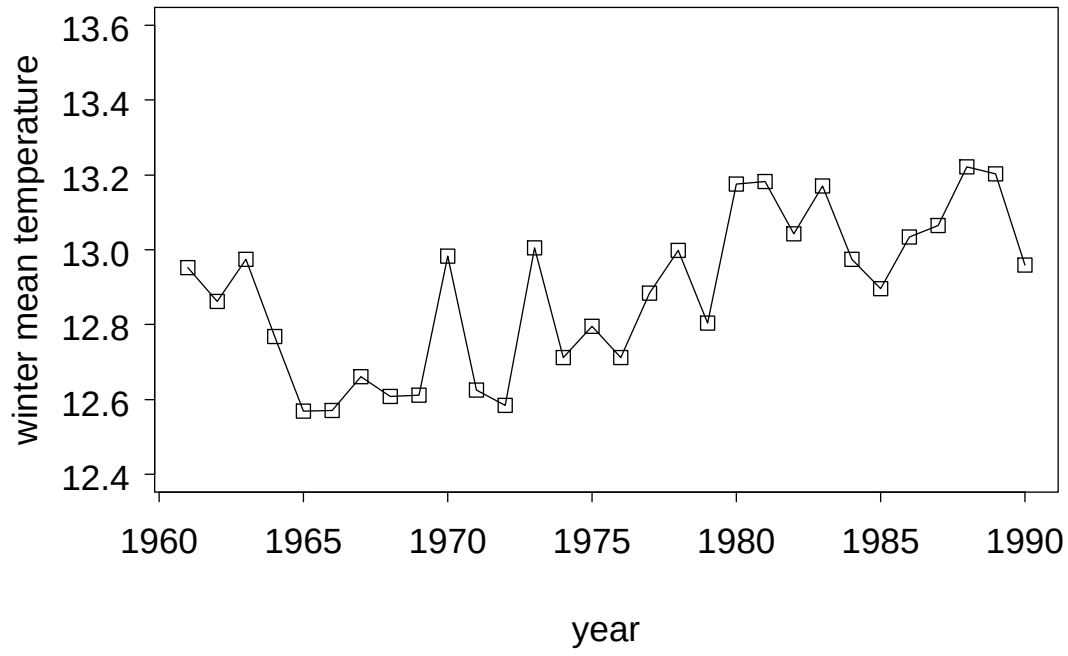

Figure 13: *Global average winter temperature (^oC) based on Model (3.3.1-3) using the 250 stations. The grand mean temperature is* $12.9(^oC)$ *and the linear trend coefficient over the 30 year period is* $.015(^oC)/year$.

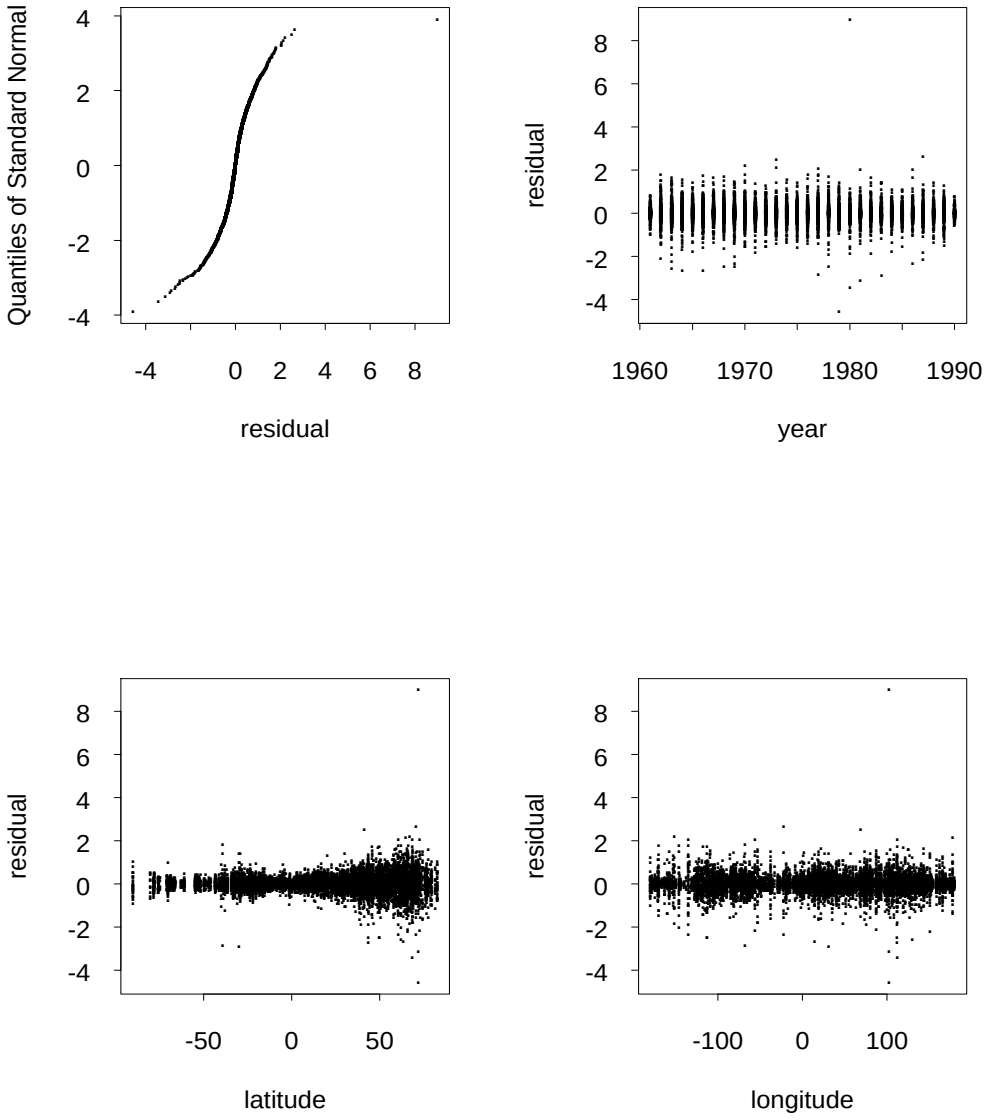

Figure 14: *Residual plots using the 500 station subset of the uncorrected version of the data.*

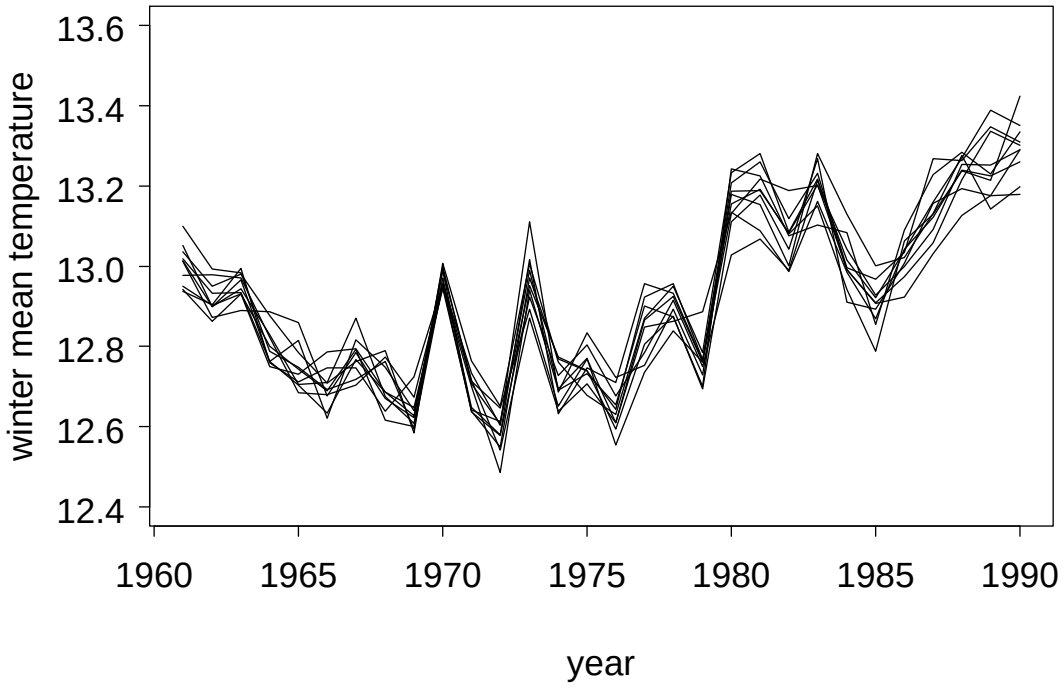

Figure 15: *SS estimates of global winter mean temperature history for 10 copies of the pseudo data. Refer to Figure 11.*

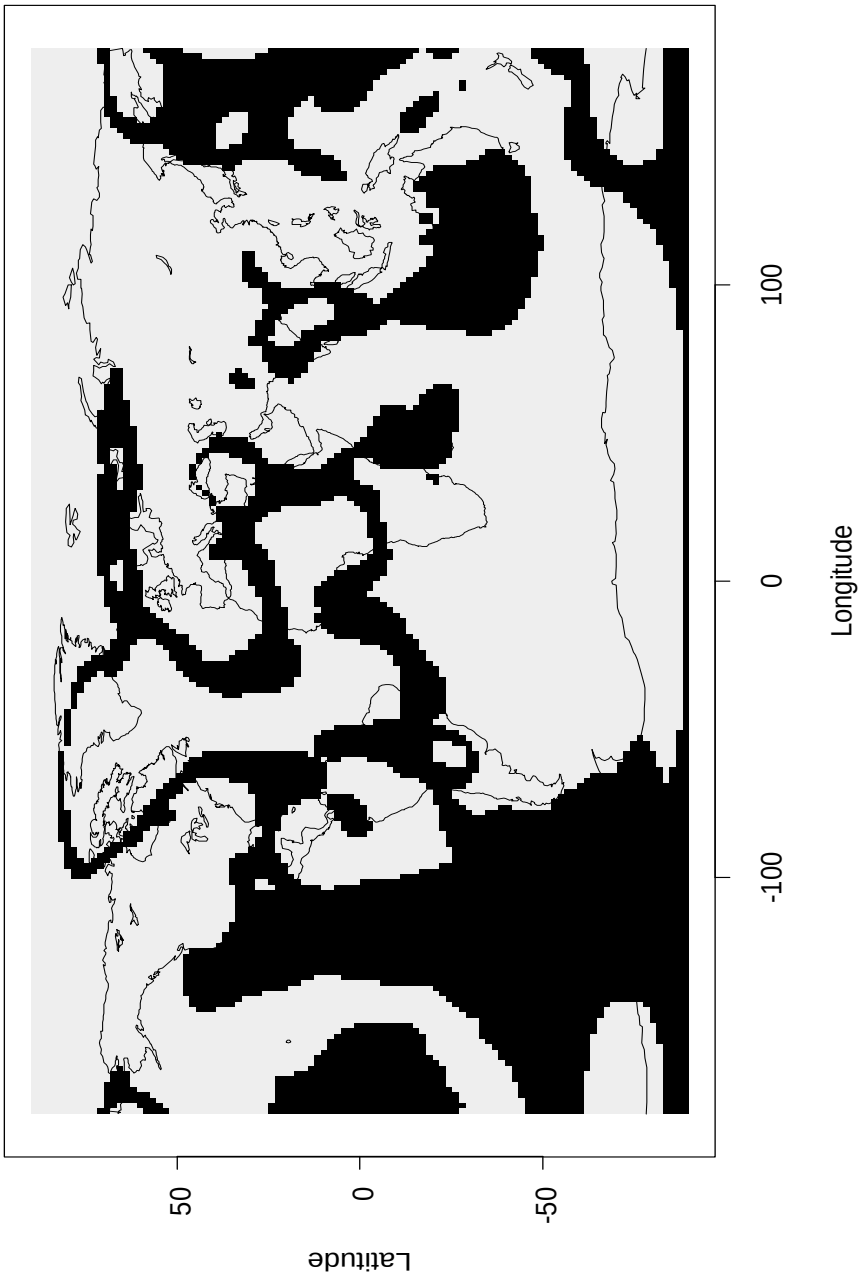

Figure 16: *Black regions are the areas where the range of 10 estimated linear trend coefficients for the pseudo data covers zero. Refer to Figure 12.*

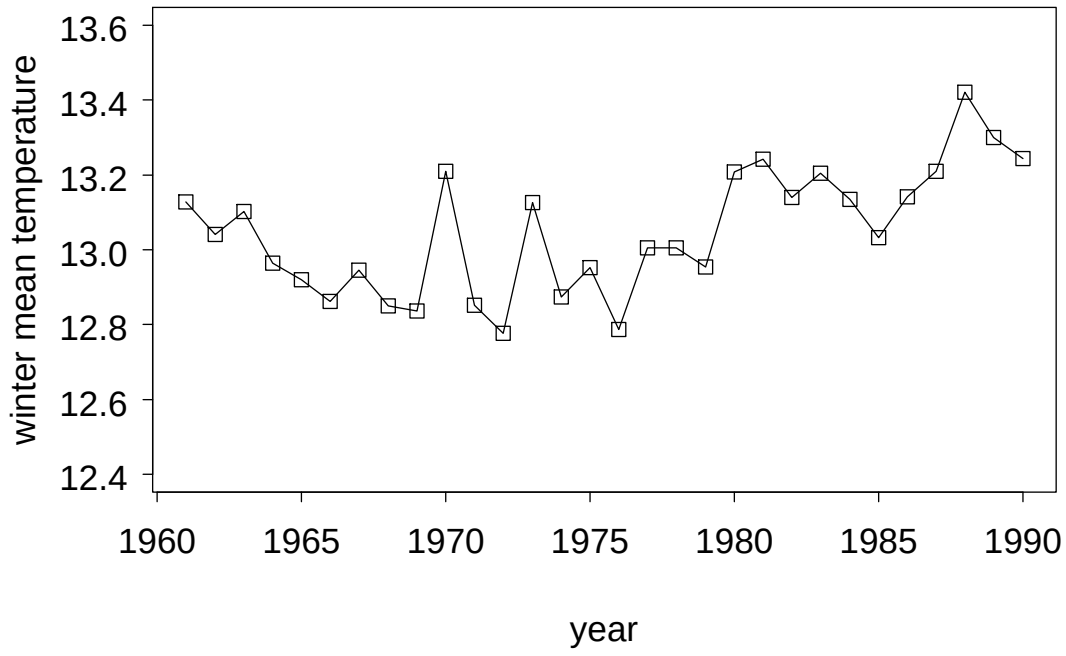

Figure 17: *Global average winter temperature* (^oC) *based on Model (3.3.1-3) using the 1000 stations' pseudo data. The grand mean temperature is* $13.0(^oC)$ *and the linear trend coefficient over the 30 year period is* $.011(^oC)/year$.

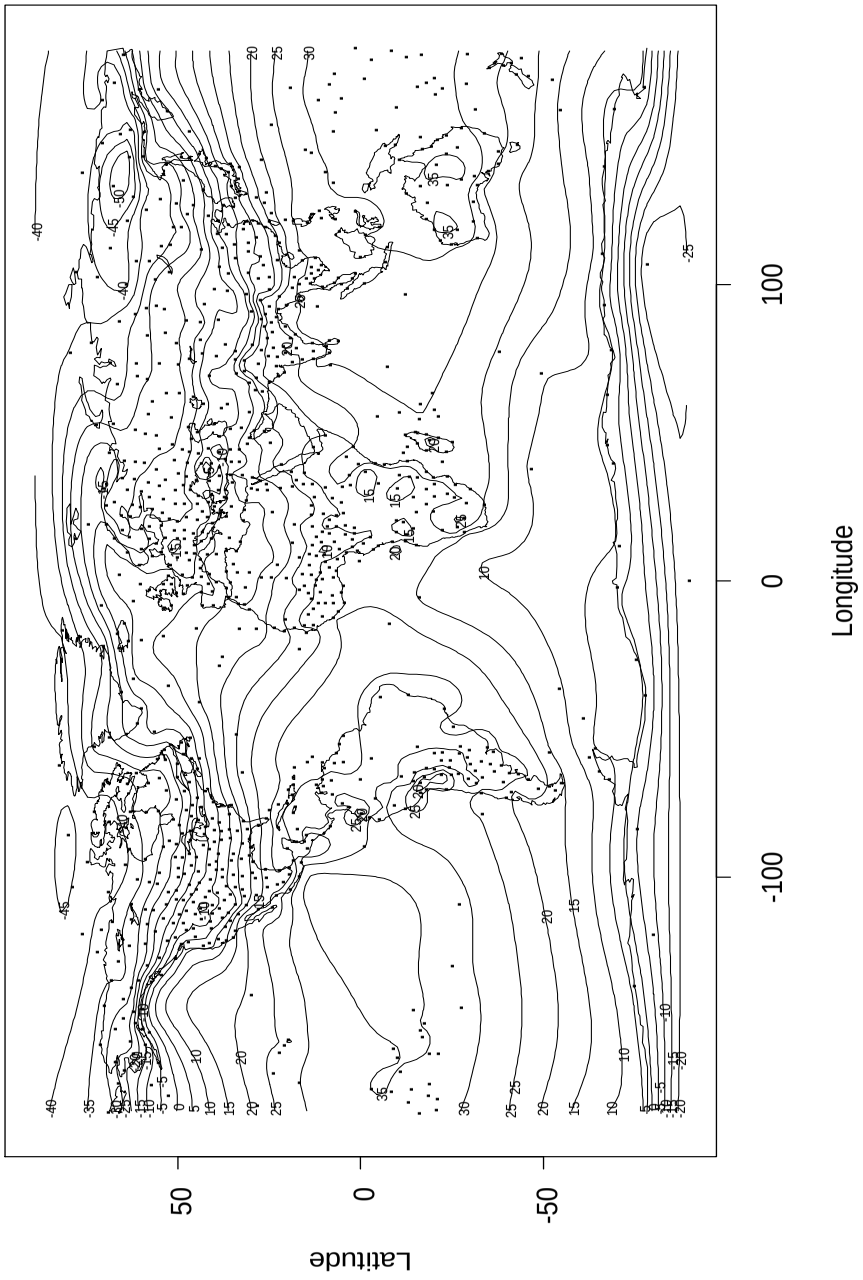

Figure 18: *Average winter temperature over the globe using the 1000 stations' pseudo data.*

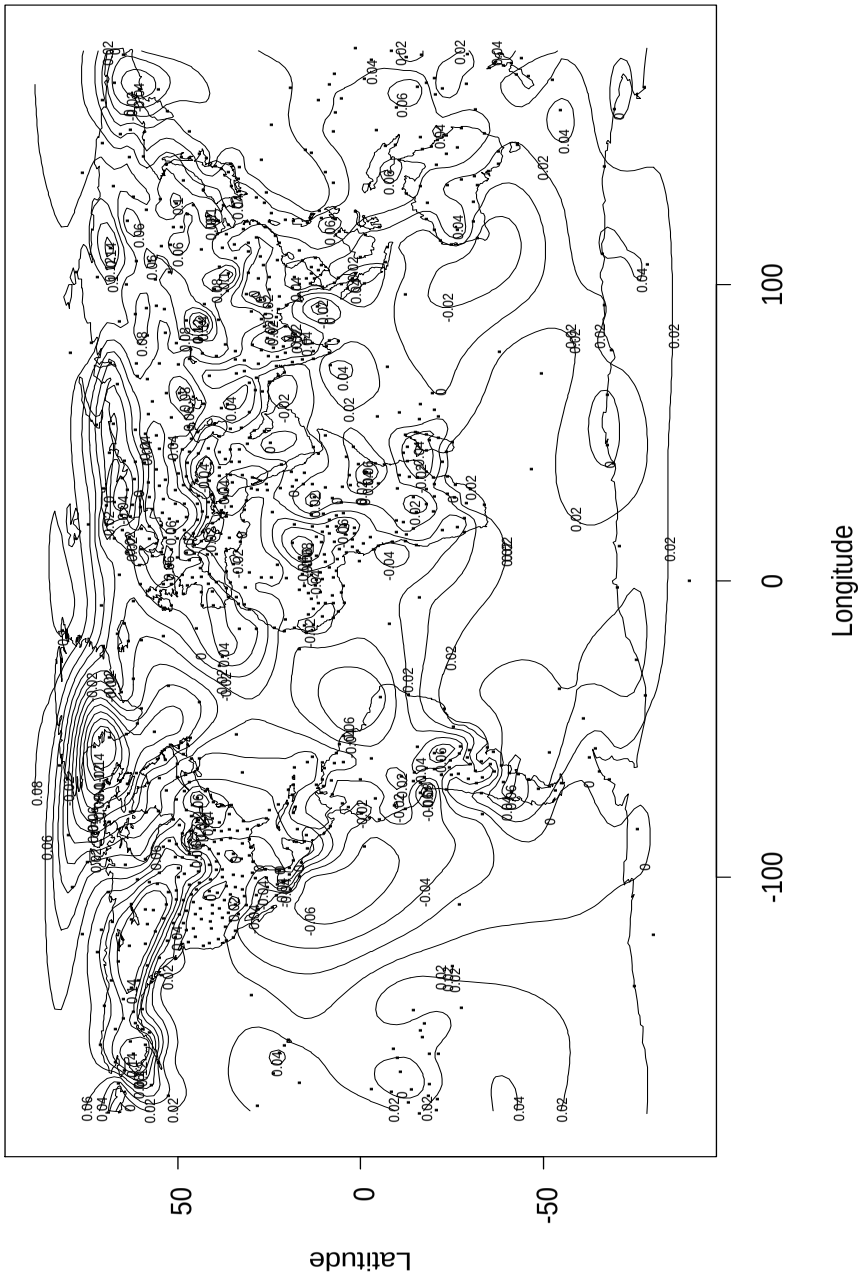

Figure 19: *Linear trend coefficient of winter temperature over the globe using the 1000 stations' pseudo data.*

Chapter 4

Backfitting vs the Gibbs sampler, and on-going research

In the previous chapters we have mentioned some on-going research problems here and there. In this chapter we will focus on constructing confidence intervals for the smoothing spline estimates using a Bayesian model and through this model studying a correspondence between optimization methods for getting smoothing spline (penalized likelihood) estimates and Monte Carlo methods for getting posterior distributions.

In Section 1, we will discuss a Bayesian model behind the SS estimates. In Section 2, the backfitting algorithm and the Gibbs sampler are considered in the same perspective. Some common issues in speeding up such as SOR, collapsing, grouping, etc., are discussed. In Section 3, more analogous situations between sampling methods and optimization methods are described.

4.1 A Bayesian model

The smoothing spline estimates have a Bayesian interpretation which has been used to construct confidence intervals. See Wahba (1978, 1983), and Gu and Wahba (1993b).

Suppose that the data we have are:

$$y_i = f(t_i) + \epsilon_i, \text{ for } i = 1, 2, \cdots, n \tag{4.1.1}$$

where ϵ_i are identically distributed independent random variables. They have

a common Gaussian distribution with a mean zero and a standard deviation σ. Assume further that

$$f(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{\alpha=1}^{p} f_\alpha(t) \tag{4.1.2}$$

where $\{\phi_\nu\}$ are M known functions (e.g., lower order polynomials), $\{d_\nu\}$ have a uniform prior, f_α as a prior is a zero-mean Gaussian process with a covariance function $\sigma^2\theta_\alpha R_\alpha(s,t)$. $\{d_\nu\}$ and $\{f_\alpha\}$ are independent. They are independent of $\{\epsilon_i\}$ too. Suppose σ and $\{\theta_\alpha\}$ are known. $\{\theta_\alpha\}$ are smoothing parameters in Section 2.1. They control the relative size of the variation in the "noise" term, ϵ, and the signal terms $\{f_\alpha\}$. $\{\theta_\alpha\}$ may be chosen using present or past data, or by pure prior (i.e. empirical Bayesian methods, or strict Bayesian methods, corresponding to the "data-driven" or "subjective" methods discussed in Section 3.3.1), or in a hierarchical Bayesian way, may be assigned a hyper-prior. In this chapter, however, we assume that they have been chosen by some method and are fixed. σ, the absolute magnitude of the noise term, is assumed to be a known parameter. It can be assigned a prior too if a strict Bayesian model is required. For example it can be assigned a Gamma prior. See Besag et. al. (1995) for some examples.

Now the posterior of $d := (d_1, \cdots, d_M)^T$ and $\{f_\alpha\}$ are proportional to

$$exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - f(t_i))^2\right\} \prod_{\alpha=1}^{p} g(f_\alpha) \tag{4.1.3}$$

where g denotes a generic density with some abuse of notations.

Rewrite (4.1.3) as

$$\begin{aligned} &exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - f(t_i))^2\right\} \prod_{\alpha=1}^{p} g(f_\alpha(t_1), \cdots, f_\alpha(t_n)) \\ &\prod_{\alpha=1}^{p} g(f_\alpha(t), t \neq t_i, i = 1, \cdots, n | f_\alpha(t_1), \cdots, f_\alpha(t_n)) \end{aligned} \tag{4.1.4}$$

Since the first part of (4.1.4) does not depend on $\{(f_\alpha(t), t \neq t_i, i = 1, \cdots, n), \alpha = 1, \cdots, p\}$, and the last part of (4.1.4) is Gaussian with $\{(f_\alpha(t_1), \cdots, f_\alpha(t_n)), \alpha = 1, \cdots, p\}$ only appearing in the means (by the assumptions about the prior of $\{f_\alpha\}$), the marginal posterior of $\{f_\alpha := (f_\alpha(t_1), \cdots, f_\alpha(t_n))^T, \alpha = 1, \cdots, p\}$ and

d is proportional to

$$\begin{aligned} & exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - f(t_i))^2\right\}\prod_{\alpha=1}^{p} g(f_\alpha(t_1),\cdots,f_\alpha(t_n)) \\ = \; & exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - f(t_i))^2\right\}\prod_{\alpha=1}^{p} exp(-\frac{1}{2\sigma^2\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha) \\ = \; & exp\left\{-\frac{1}{2\sigma^2}\left(\sum(y_i - f(t_i))^2 + \sum_{\alpha=1}^{p}\frac{1}{\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha\right)\right\} \end{aligned} \tag{4.1.5}$$

We have used $\{f_\alpha\}$ to denote both the component functions and the vector of their values at points $\{t_i\}$ when there is no confusion.

Note that since $f_\alpha \sim N(0, \sigma^2\theta_\alpha Q_\alpha)$, we know that $f_\alpha \in \mathcal{L}(Q_\alpha)$ almost surely.

Hence now to maximize the posterior (4.1.5) is equivalent to minimizing

$$\sum(y_i - f(t_i))^2 + \sum_{\alpha=1}^{p}\frac{1}{\theta_\alpha} f_\alpha^T Q_\alpha^\dagger f_\alpha \tag{4.1.6}$$

which is the same as (2.2.6). Thus we see that the SS estimate is a posterior mode when the prior is given in such a way.

We have assumed that σ is known. If we assign a prior such as $\delta := 1/\sigma^2 \sim \Gamma(a, b)$, that is δ has a density $\delta^{a-1}exp(-b\delta)$, then it can be shown that maximizing the posterior is still equivalent to minimizing (4.1.6).

A Bayesian model like this can be used to construct confidence intervals based on the posterior of the estimated functions. See Wahba (1983) and Gu and Wahba (1993b) for the formulation. See also Nychka (1988, 1990) for their frequentist properties. Analogous to the situation in which the posterior mode is computed, computing posterior variances requires huge memory too if direct matrix decomposition methods are to be used. A possible way out of this memory problem is to get the posterior distribution through Monte Carlo in a similar way as we get the posterior mode in Chapter 2, i.e., through componentwise updating. This leads to our discussion in the next section.

4.2 Backfitting vs the Gibbs sampler

Since the posterior of $(d, f_1, f_2, \cdots, f_p)$ in the Bayesian model of the last section is proportional to

$$exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - f(t_i))^2 + \sum_{\alpha=1}^{p}\frac{1}{\theta_\alpha}f_\alpha^T Q_\alpha^\dagger f_\alpha\right)\right\} \tag{4.2.1}$$

the SS estimate is the same as the posterior mode. One way to compute the posterior mode is through maximizing (4.2.1) component-wisely. That is to maximize along each component of $(d, f_1, \cdots, f_p)$ in turn until it converges. Since the conditional posterior of each component given others is proportional to (4.2.1), maximizing along each component is equivalent to computing conditional modes. Hence we see that Besag (1986)'s iterative conditional mode (ICM) algorithm is actually the component-wise descent method for optimization.

The conditional posterior of f_β, given $f_\alpha, \alpha \neq \beta$ and d, is proportional to

$$exp\left\{-\frac{1}{2\sigma^2}\left(\|y - Sd - \sum_{\alpha=1}^{p} f_\alpha\|^2 + \frac{1}{\theta_\beta}f_\beta^T Q_\beta^\dagger f_\beta\right)\right\} \tag{4.2.2}$$

hence the conditional mode is

$$f_\beta = (\frac{1}{\theta_\beta}I + Q_\beta)^{-1}Q_\beta(y - Sd - \sum_{\alpha\neq\beta} Q_\alpha) \tag{4.2.3}$$

and the conditional distribution is

$$\begin{aligned} & f_\beta | d, f_\alpha, \alpha \neq \beta, y \\ \sim\ & N\left((Q_\beta + \frac{1}{\theta_\beta}I)^{-1}Q_\beta(y - Sd - \sum_{\alpha\neq\beta} Q_\alpha), \sigma^2(Q_\beta + \frac{1}{\theta_\beta}I)^{-1}Q_\beta\right) \end{aligned} \tag{4.2.4}$$

Similarly, the conditional posterior of d, given f_α for $\alpha = 1, \cdots, p$, is proportional to

$$exp\left\{-\frac{1}{2\sigma^2}\|y - Sd - \sum_{\alpha=1}^{p} f_\alpha\|^2\right\} \tag{4.2.5}$$

hence the conditional posterior mode is

$$d = (S^T S)^{-1}S^T(y - \sum_{\alpha=1}^{p} f_\alpha) \tag{4.2.6}$$

and the conditional posterior distribution is

$$\begin{aligned} & d|f_\alpha, \alpha = 1, \cdots, p, y \\ \sim \quad & N\left((S^TS)^{-1}S^T(y - \sum_{\alpha=1}^{p} f_\alpha), \sigma^2(S^TS)^{-1}\right) \end{aligned} \tag{4.2.7}$$

(4.2.3) and (4.2.6) are the same as the updating formulae for backfitting. See (2.2.4). Viewing backfitting under this perspective, one can see that it is analogous to the Gibbs sampler (see Besag et. al. (1995), Liu (1994), Liu and et. al. (1994,1995), and Roberts and Sahu (1996)). Both the backfitting algorithm and the Gibbs sampler make use of conditional posteriors. One computes conditional modes (4.2.3) and (4.2.6) in turn to get the mode of the joint posterior (4.2.2), the other samples from conditional posteriors (4.2.4) and (4.2.7) in turn to get a sample (correlated though) from the joint posterior (4.2.2).

The advantage of using the Gibbs sampler to get a posterior sample here is similar to the advantage of using the backfitting algorithm to get the posterior mode. That is to say, we can easily get the eigen-decompositions of the matrices in the updating formulae through a tensor product structure if the data have a tensor product design. Therefore we can update those components quickly. In this way, overall computing time and space may be saved.

With an incomplete tensor product design, analogous to the EM algorithm, the Data Augmentation method of Tanner and Wong (1987) may be used together with the Gibbs sampler. However, its feasibility in our application of Chapter 3, i.e., the situation of very large data size, still needs to be investigated.

4.2.1 Issues in speeding up Gibbs sampler

There are some speeding-up methods for the Gibbs sampler which are quite similar to those for the backfitting (Gauss-Seidel) algorithm. Many authors have noted such analogous situations for non-stochastic iterative algorithms and iterative Monte Carlo algorithms. For example, see Besag et. al. (1995) and Roberts and Sahu (1996).

Barone and Frigessi (1989) proposed a stochastic relaxation method that is a direct analog of successive over relaxation (SOR) for speeding up the Gauss-Seidel algorithm. Suppose that we want to sample from a multivariate normal

distribution. The Gibbs sampler draws a new value of one component from the conditional distribution of this component given other components, say, $N(\mu_i, \sigma_i^2)$. Successive drawing through all the components will give us a sequence of (correlated) random vectors distributed as the target multivariate normal distribution (after convergence). Barone and Frigessi's stochastic relaxation method draws a new value from $N(\omega\mu_i + (1-\omega)x_i, \omega(2-\omega)\sigma_i^2)$ where ω is a constant in $(0, 2)$. Similarly for the Gauss-Seidel algorithm, a new updated value, i.e., a conditional mode μ_i (considering that GS is the same as ICM), is replaced by a linear combination of μ_i and its corresponding old value x_i, i.e., $\omega\mu_i + (1-\omega)x_i$, which is the mode of the updating distribution of Barone and Frigessi's algorithm. Not much convergence results are available about this stochastic relaxation method. See Green and Han (1992). It seems plausible that the rich material in numerical analysis literature about SOR might benefit the research on the convergence of this stochastic relaxation method.

One product of such a connection is some of the results in Roberts and Sahu (1995) in which they prove that the convergence rate in terms of Chi-square distance of the iterates and the target distribution is the same as the spectral radius of the Gauss-Seidel algorithm's updating matrix (comparing their B with (3.19) on page 72 of Young (1971)). Some results of Roberts and Sahu are actually direct analogs of the similar results for the Gauss-Seidel algorithm in Varga(1962) and Young (1971). For example, their Theorem 8 about the convergence rate of grouping components in the Gibbs sampler is a direct result of Varga (1962)'s Theorem 3.15 about the convergence rate of grouping components in the Gauss-Seidel algorithm, due to the above connection, as noted by them. See also Liu (1994), Liu, Wong and Kong (1994, 1995) for more about speeding up the Gibbs sampler through grouping and collapsing. An application in the reverse direction is the use of collapsing in the Gauss-Seidel algorithm as discussed in Section 2.3.2.

4.3 Other analogous algorithms

There are some other analogs between Monte Carlo sampling algorithms and optimization algorithms. One such example is Amit, Grenander and Piccioni (1991)'s Langevin-Hastings algorithm for sampling from $\pi(x) \propto exp\{-u(x)\}$,

$x \in R^n$. The proposed next state of x is drawn from

$$x' \sim N(x - \tau \nabla u(x), 2\tau I_n) \tag{4.3.1}$$

where τ is a small positive constant. See also Besag et. al. (1995, Section 2.3.4). Similar to the analog between Barone and Frigessi (1989)'s stochastic relaxation and SOR, here the mean of the proposal distribution is a move along the steepest descent direction, hence it is analogous the steepest descent algorithm in optimization literature.

In a problem of sampling from a high-dimensional Gaussian distribution, we may find useful an application of an analog of the conjugate gradient algorithm in optimization literature.

Consider the problem:

$$\min_{x \in R^n} f(x) \tag{4.3.2}$$

where $f(x) \propto exp\{-h(x)\}$, $h(x) = \frac{1}{2}x^T Q x - b^T x$ where Q is a positive definite matrix which may be too large to be saved. The conjugate gradient algorithm (see Luenberger (1984), p. 244) is:

1. Start with x_0, let $g_0 = \nabla h(x_0)$, $d_0 = -g_0$;
2. For $k = 1, 2, \cdots, n$, compute
   (a) $r = 2(h(d_{k-1}) + b^T d_{k-1})$;
   (b) $\alpha = -g_{k-1}^T d_{k-1}/r$;
   (c) $x_k = x_{k-1} + \alpha d_{k-1}$;
   (d) $g_k = \nabla h(x_k), \beta = \frac{(g_k - g_{k-1})^T g_k}{\alpha r}$;
   (e) $d_k = -g_k + \beta d_{k-1}$;

Note that $2(h(d) + b^T d) = d^T Q d$, $\frac{(g_k - g_{k-1})^T g_k}{\alpha} = \frac{(x_k - x_{k-1})^T Q g_k}{\alpha} = d_{k-1}^T Q g_k$, and $b = -\nabla h(0)$. The expressions used in the algorithm are chosen in order to avoid using Q directly. A property of this algorithm is that the resulted $d_k, k = 0, 1, \cdots, n-1$ are Q-conjugate directions, that is, $d_k^T Q d_l = 0$ for any $k \neq l$.

Now if we want to sample from distribution $f(x)$ instead of minimizing $f(x)$, we can make use of these directions. Since

$$\begin{aligned} & exp\{-\frac{1}{2}x^TQx + x^Tb\} \\ = \; & exp\{-\frac{1}{2}(\sum_{k=0}^{n-1} a_k d_k)^T Q(\sum_{k=0}^{n-1} a_k d_k) + (\sum_{k=0}^{n-1} a_k d_k)^T b\} \\ = \; & exp\{-\frac{1}{2}\sum_{k=0}^{n-1} a_k^2 d_k^T Q d_k + \sum_{k=0}^{n-1} a_k d_k^T b\} \\ \propto \; & exp\{-\frac{1}{2}\sum_{k=0}^{n-1} \frac{(a_k - \frac{d_k^T b}{d_k^T Q d_k})^2}{(d_k^T Q d_k)^{-1}}\} \end{aligned} \tag{4.3.3}$$

Hence we see that we can sample from $f(x)$ by $x = \sum_{k=0}^{n-1} a_k d_k$ where a_k is distributed as $N(\frac{d_k^T b}{d_k^T Q d_k}, (d_k^T Q d_k)^{-1})$. Note that $d_k^T Q d_k = 2(h(d_k) + d_k^T b)$ can be calculated without explicit use of Q. In order for this method to be feasible in practice, the calculation of h and its derivatives must be very efficient. In many situations corresponding to statistical models of some special structure, this is often the case.

Even though the conjugate gradient algorithm has been extended to non-quadratic optimization problems, the extendibility of the above sampling algorithm to non-quadratic cases is still under investigation.

Bibliography

Amit, Y., Grenander, U. and Piccioni, M. (1991), " Structural image restoration through deformable templates", *J. Amer. Statist. Assoc.*, Vol. 86, 376-387.

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.

Aronszajn, N. (1950), "Theory of reproducing kernels", *Trans. Amer. Math. Soc.*, Vol. 68, 337-404.

Barone, P. and Frigessi, A. (1989), "Improving stochastic relaxation for Gaussian random fields", *Probability in the Engineering and Informational Sciences*, Vol. 4, 369-389.

Besag, J. (1986), "On the statistical analysis of dirty pictures" (with discussion), *J. Roy. Statist. Soc. Ser. B*, Vol. 48, 259-302.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with Comments), *Statistical Science*, Vol. 10, No. 1, 3-66.

Buja, A., Hastie, T. and Tibshirani, R. (1989), "Linear smoothers and additive models", (with discussions) *Ann. Stat.*, Vol. 17, No. 2, 453-555.

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.

Girard, D.(1991), "Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization", *Ann. Statist.*, Vol. 19, 1950-1963.

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.

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*.

Hurrell, J.W. and Trenberth, K.E. (1996), "Satellite versus Surface Estimates of Air Temperature since 1979", *J. Climate*, to appear.

Hutchinson, M.(1989), "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines", *Commun. Statist.-Simula.*, Vol. 18, 1059-1076.

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 Temperature Anomaly Data Set: 1851-1988", Environmental Sciences Division Publication No. 3520, U.S. Department of Energy.

Jones, P.D., Wigley, T.M.L., Kelly, P.M.(1982), "Variations in surface air temperatures: Part 1. Northern Hemisphere, 1881-1980", *Mon. Wea. Rev.*, Vol. 110, 59-72.

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.

Knight, R.W. (1980), "A comparison of some methods for flagging erroneous observations in certain types of meteorological data", Techincal Report No. 610, Department of Statistics, University of Wisconsin -Madison.

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.

Liu, J.S., Wong, W.H. and Kong, A. (1994), "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes", *Biometrika*, Vol. 81, No. 1, 27-40.

——— (1995), "Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans" *J. R. Statist. Soc. B*, Vol. 57, No. 1, 157-169.

Luenberger, D.G. (1984), *Linear and Nonlinear Programming*, 2nd Ed., Addison-Wesley, Reading, Massachusetts.

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.

Nicholls, N., Gruza, G.V., Jouzel, J., Karl, T.R., Ogallo, L. A. and Parker, D. E., (1996), "Observed Climate Variability and Change", in *Climate Change 1995 : The Science of Climate Change*, edited by Houghton, J. T., Meira Filho, L. G., Callander, B. A., Harris, N., Kattenberg, A. and Maskell, K., 134-192, Cambridge University Press.

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.

Raper, S.C.B., Wigley, T.M.L., Mayes, P.R., Jones, P.D., and Salinger, M.J. (1984), "Variations in surface air temperatures. Part 3: The Antarctic, 1957-82", *Mon. Wea. Rev.*, Vol. 112, 1341-1353.

Roberts, G.O. and Sahu, S.K. (1996), " Updating Schemes, Correlation Strucure, Blocking and Parameterization for the Gibbs Sampler", Technical Report, University of Cambridge.

Sampson, P.D. and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure", *J. Amer. Statist. Assoc.*, Vol. 87, 108-119.

Tanner, M.A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd Ed., Springer-Verlag, New York.

Tanner, M.A. and Wong, W.H. (1987), "The Calculation of posterior distributions by data augmentation", *J. Amer. Statist. Assoc.*, Vol. 82, 528-540.

Trenberth, K.E. (Editor) (1992), *Climate System Modeling*, Cambridge University Press.

Varga, R.S (1962), *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey.

Vinnikov, K.Ya., Gruza, G.V., Zakharov, V.F., Kirillov, A.A., Kovyneva, N.P. and Ran'kova, E. Ya. (1980), "Contemporary variations of the Northern Hemisphere climate", *Meteor. Gidrol.*, Vol. 6, 5-17 (in Russian).

Vinnikov, K.Ya., Groisman, P.Ya. and Lugina, K.M. (1990), "Empirical Data on Contemporary Global Climte Changes (Temperature and Precipitation)", *J. Climate*, Vol. 3, 662-677.

Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression", *J. Roy. Stat. Soc. Ser. B*, Vol.40, No.3, 364-372.

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.

Wang, Y. (1994), "Smoothing Spline Analysis of Variance of Data from Exponential Families", Technical Report No. 928, Department of Statistics, University of Wisconsin-Madison (Thesis).

Weber, R. and Talkner, P. (1993), "Some remarks on spatial correlation function models", *Monthly Weather Review*, Vol. 121, 2611-2617.

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.