Time and Space Models on the Globe:
Thirty years (1961-90) of Dec. Jan. Feb. average temperature measurements at 1000 stations around the globe (with missing data- 23,119 observations), \( t = (t_1, t_2) = (x, P) \) where \( x \) is year, and \( P \) is (latitude, longitude). The RKHS of historical global temperature functions that was used is

\[
\mathcal{H} = [(1^{(1)}) \oplus [\phi] \oplus \mathcal{H}_{s}^{(1)})] \otimes [(1^{(2)}) \oplus \mathcal{H}_{s}^{(2)}),
\]

a collection of functions \( f(x, P) \), on

\[
\{1, 2, ..., 30\} \otimes S,
\]

where \( S \) is the sphere. \( \mathcal{H} \) and \( f \) have the corresponding (six term) decompositions given next:
\[ \mathcal{H} = [1] \oplus [\phi] \oplus [\mathcal{H}_s^{(1)}] \oplus [\mathcal{H}_s^{(2)}] \]

\[ f(x, P) = C + d\phi(x) + f_1(x) + f_2(P) \]

\[ = \text{mean} + \text{global} + \text{time} + \text{space} \]

\[ + \oplus [[\phi] \otimes \mathcal{H}_s^{(2)}] + \phi(x)f_{\phi,2}(P) + f_{12}(x, P) \]

\[ + \oplus \mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)} \]

Here \( \phi \) is a linear function which averages to 0. A sum of squares of second differences was applied to the time variable, and a spline on the sphere penalty was applied to the space variable.
\[ \begin{array}{ccc}
\beta & RKHS & RK \\
1 & \mathcal{H}_s^{(1)} & R_1(x, P; x', P') = \tilde{R}_1(x, x') \\
2 & \mathcal{H}_s^{(2)} & R_2(x, P; x', P') = \tilde{R}_2(P, P') \\
3 & [\phi] \otimes \mathcal{H}_s^{(2)} & R_3(x, P; x', P') = \phi(x)\phi(x')\tilde{R}_2(P, P') \\
4 & \mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)} & R_4(x, P; x', P') = \tilde{R}_1(x, x')\tilde{R}_2(P, P') \\
\end{array} \]

1 = time, 2 = space, 3 = time main effect \times space interaction (trend by space), 4 = smooth time \times smooth space interaction.
Find $f$ in $\mathcal{M} = \mathcal{H}^0 \oplus \sum_{\beta} \mathcal{H}^\beta$ to minimize

$$
\sum_{i=1}^{n} (y_i - f(t(i)))^2 + \sum_{\beta=1}^{4} \theta_{\beta}^{-1} \|P^\beta f\|^2,
$$

where $P^\beta$ is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^\beta$, and $\theta_{\beta}^{-1} = \lambda_{\beta}$. The minimizer $f_{\lambda}$ ($\lambda = (\lambda_1, \cdots, \lambda_4)$) is of the following form: Letting

$$
Q_\theta(s, t) = \sum_{\beta=1}^{4} \theta_{\beta} R_{\beta}(s, t),
$$

then

$$
f_\theta(t) = \sum_{\nu=1}^{2} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i Q_\theta(t(i), t).
$$

c_{n \times 1}$ and $d_{2 \times 1}$ are vectors of coefficients which satisfy

$$
(Q_\theta + I)c + Sd = y \\
S'c = 0
$$

$Q_\theta$ is the $n \times n$ matrix with $ij$th entry $Q_\theta(t(i), t(j))$, and $S$ is the $n \times 2$ matrix with $i\nu$th entry $\phi_\nu(t(i))$. 

4
This system will have a unique solution for any set of positive $\{\lambda_\beta\}$ provided $S$ is of full column rank, which we will always assume. If all 1000 stations reported for each of the 30 years, then $n = 30,000$. Results in an unpleasantly large linear system to solve.
The backfitting algorithm:

The representation (2) can certainly be written as

\[ f_\theta(t) = \sum_{\nu=1}^{2} d_\nu \phi_\nu(t) + \sum_{\alpha=1}^{4} \theta_\alpha \sum_{i=1}^{n} c_{i,\alpha} R_\alpha(t_i, t) \]  

(3)

too, where \( c_{i,\alpha} \) differs for different \( \alpha \). Since the minimizer of (2) is unique (assuming as usual that \( S \) is of full rank), we can minimize (2) within the class of functions of form (3) and get the same smoothing spline estimates as before. This leads to a problem of minimizing:

\[ \| y - Sd - \sum_{\alpha=1}^{4} \theta_\alpha Q_\alpha c_\alpha \|^2 + \sum_{\alpha=1}^{4} \theta_\alpha c_\alpha^T Q_\alpha c_\alpha \]  

(4)

over \( d \) and \( c_\alpha \), for \( \alpha = 1, 2, 3, 4 \), where

\[ Q_\alpha := (R_\alpha(t(i), t(j)))_{n \times n}. \]
The corresponding stationary equations are:

\[
\begin{align*}
(S^T S)_d &= S^T (y - \sum_{\alpha=1}^p \theta_\alpha Q_\alpha c_\alpha) \\
(\theta_\beta Q_\beta + I) Q_\beta c_\beta &= Q_\alpha (y - Sd - \sum_{\alpha \neq \beta} \theta_\alpha Q_\alpha c_\alpha),
\end{align*}
\]

for \( \beta = 1, 2, 3, 4 \).

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 \( f_\theta \) and its components. Without confusion within their context, we denote the component functions of SS estimate \( f_\theta \) evaluated at data points as \( f_0, f_1, \cdots, f_4 \) also. That is,

\[
\begin{align*}
f_0 &= Sd \\
f_\alpha &= \theta_\alpha Q_\alpha c_\alpha,
\end{align*}
\]

for \( \alpha = 1, 2, \cdots, p \). They must satisfy

\[
\begin{align*}
f_0 &= S_0 (y - \sum_{\alpha=1}^p f_\alpha) \\
f_\beta &= S_\beta (y - \sum_{\alpha \neq \beta} f_\alpha), \quad \text{for } \beta = 1, 2, 3, 4.
\end{align*}
\]

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

\[
\begin{cases}
    f^{(k)}_0 &= S_0 (y - \sum_{\alpha=1}^{\beta} f^{(k-1)}_{\alpha}) \\
    f^{(k)}_\beta &= S_\beta (y - \sum_{\alpha<\beta} f^{(k)}_{\alpha} - \sum_{\alpha>\beta} f^{(k-1)}_{\alpha}),
\end{cases}
\]

(7)

for $\beta = 1, 2, \cdots, 4$.

This is exactly the backfitting algorithm studied in Buja, Hastie and Tibshirani (1989), "Linear Smoothers and Additive Models", Ann. Statist. 17, No2 453-510, in JSTOR.
Rewrite the equations (6) as

\[
\begin{pmatrix}
I & S_0 & \cdots & S_0 \\
S_1 & I & \cdots & S_1 \\
\vdots & & & \vdots \\
S_4 & S_4 & \cdots & I
\end{pmatrix}
\begin{pmatrix}
f_0 \\
f_1 \\
\vdots \\
f_4
\end{pmatrix}
= 
\begin{pmatrix}
S_0y \\
S_1y \\
\vdots \\
S_4y
\end{pmatrix}
\] (8)
It is clear that the backfitting algorithm we have just described, (7), is a (block) Gauss-Seidel algorithm.

Having known \( f_0 (= Sd) \), we know \( d \) immediately. By (3), \((Q_\theta + I)c = y - Sd\), hence

\[
c = y - Sd - Q_\theta c = y - \sum_{\alpha=0}^{4} f_\alpha \quad (9)
\]

Therefore \( c \) is available after we get the \( f_\alpha \)'s.

One advantage of the backfitting algorithm is that it enables us to take advantage of some special structures of \( Q_\alpha \) 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 \(i = 1, 2, \cdots, n_1 = 30\) and \(j = 1, 2, \cdots, n_2 = 1000\). That is, the data have a tensor product design. Hence the sample size \(n = n_1 n_2 = 30,000\). Then the \(S\) and \(Q_\alpha\)'s have the following forms:

\[
S = 1 \otimes \tilde{S} \\
Q_1 = 11^T \otimes Q_t \\
Q_2 = Q_s \otimes 11^T \\
Q_3 = Q_s \otimes \phi \phi^T \\
Q_4 = Q_s \otimes Q_t
\]

where 1 is a vector of ones of appropriate length, \(\phi = (\phi(1), \cdots, \phi(n_1))^T\), \(\tilde{S} = (1 \phi)_{n_1 \times 2}\), \(Q_s\) 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), because $Q_\theta = \sum_{\alpha=1}^{4} \theta_\alpha Q_\alpha$ does not have a tensor-product structure even though every single $Q_\alpha$ 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 (7) involves just a few matrix multiplications.

Unfortunately there were about 3000 missing data points which destroyed the tensor product structure, but that was gotten around by a generalization of the leaving-out-one lemma.
The Leaving-Out-K Lemma

Let $\mathcal{H}$ be an RKHS with subspace $\mathcal{H}^0$ of dimension $M$ and for $f \in \mathcal{H}$ let $\|Pf\|^2 = \sum_{\beta=1}^{p} \theta_{\beta}^{-1} \|P_{\beta}f\|^2$. Let $f[K]$ be the solution to the variational problem: Find $f \in \mathcal{H}$ to minimize

$$\sum_{i=1}^{n} (y_i - f(t(i)))^2 + \|Pf\|^2,$$

where $S_K = \{i_1, \cdots, i_K\}$ is a subset of $1, \cdots, n$ with the property that the above 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

$$\sum_{i=1}^{n} (y_i - f(t(i)))^2 + \sum_{i \in S_K} (y_i^* - f(t(i)))^2 + \|Pf\|^2$$

is $f[K]$. 

12
Let $y$ be partitioned as

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

where $y^{(1)}$ are observed and $y^{(2)}$ have been imputed. and let $A(\lambda)$ be defined as before by $\tilde{f} = A(\lambda)y$. Let $A(\lambda)$ be partitioned corresponding to (10) as

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

Then, by the Leaving-Out-K Lemma,

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

(12)
and, if furthermore $(I - A_{22}) \succ 0$, then

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

(13)
There is an easy necessary and sufficient condition for \((I - A_{22}) \succ 0\)

Pre-Imputation Lemma:

Let \(\Gamma_1\) 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 (10) as

\[
\begin{pmatrix}
\Gamma_{11} \\
\vdots \\
\Gamma_{21}
\end{pmatrix}.
\]  

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

Proof by contradiction, if 1 is an eigenvalue, then the problem does not have a unique solution.
The Imputation Lemma:

Let $g^{(2)}_o$ be a $K$-vector of initial values for an imputation of $(f^K(t(i_1)), \ldots, f^K(t(i_K)))'$, and suppose $0 \prec (I - A_{22})$. Let successive imputations $g^{(2)}(\ell)$ for $\ell = 1, 2, \cdots$ be obtained via

$$
\begin{pmatrix}
g_1^{(\ell)} \\
\vdots \\
g_2^{(\ell)}
\end{pmatrix}
= A(\lambda)
\begin{pmatrix}
y_1 \\
\vdots \\
y_{2(\ell-1)}
\end{pmatrix}.
$$

Then

$$
\lim_{\ell \to \infty}
\begin{pmatrix}
g_1^{(1)} \\
\vdots \\
g_2^{(2)}
\end{pmatrix}
= \begin{pmatrix}
f^K(t(1)) \\
\vdots \\
f^K(t(n))
\end{pmatrix}.
$$
Proof: By the Leaving-Out-$K$ Lemma,

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

so we only need to show that

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

But

\[
g^{(2)}_{(\ell)} = A_{21}y^{(1)} + A_{22}[A_{21}y^{(1)} + A_{22}g^{(2)}_{(\ell-1)}] = \ldots = (I + A_{22} + \cdots + A_{22}^{\ell-1})A_{21}y^{(1)} + A_{22}^{\ell}g^{(2)}_{(o)}.
\]

so that assuming $0 \prec (I - A_{22})$ then $A_{22}^{\ell}$ tends to 0, giving

\[
g^{(2)}_{(\ell)} \to (I - A_{22})^{-1} A_{21}y^{(1)},
\]

and the result follows.
We remark that the randomized trace technique works perfectly well in conjunction with the imputation technique. The components of the noise vector $\xi$ in the randomization technique are generated only where there are observations.