

### General Formulation

The general form of PLS functional in an RKHS  $\mathcal{H} = \bigoplus_{\beta=0}^p \mathcal{H}_\beta$  is

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda J(\eta),$$

where  $J(f) = J(f, f) = \sum_{\beta=1}^p \theta_\beta^{-1} (f, f)_\beta$  and  $(f, g)_\beta$  are IPs in  $\mathcal{H}_\beta$  with RKs  $R_\beta(x, y)$ . The penalty  $\lambda \sum_{\beta=1}^p \theta_\beta^{-1} (f, f)_\beta$  is overparameterized by  $(\lambda, \theta_\beta)$ , but only the ratios  $\lambda/\theta_\beta$  matter.

$J(f, g) = \sum_{\beta=1}^p \theta_\beta^{-1} (f, g)_\beta$  is an IP in  $\mathcal{H}_J = \bigoplus_{\beta=1}^p \mathcal{H}_\beta$ , with an RK  $R_J(x, y) = \sum_{\beta=1}^p \theta_\beta R_\beta(x, y)$  and a null space  $\mathcal{N}_J = \mathcal{H}_0$ . The minimizer of PLS is of the form

$$\eta(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{i=1}^n c_i R_J(x_i, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c},$$

where  $\{\phi_\nu\}_{\nu=1}^m$  is a basis of  $\mathcal{N}_J = \mathcal{H}_0$ .

Slide 1

### Numerical Problem

Plugging in the solution expression, PLS becomes

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T (\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda \mathbf{c}^T Q\mathbf{c},$$

where  $S_{i,\nu} = \phi_\nu(x_i)$  and  $Q_{i,j} = R_J(x_i, x_j)$ .

If  $S$  is of full column rank, the minimizer of PLS is unique, although  $(\mathbf{c}, \mathbf{d})$  may not be. The linear system

$$(Q + n\lambda I)\mathbf{c} + S\mathbf{d} = \mathbf{Y},$$

$$S^T \mathbf{c} = 0$$

yields a solution, which is all one needs.

Let  $S = (F_1, F_2) \begin{pmatrix} R \\ 0 \end{pmatrix} = F_1 R$  be the QR-D of  $S$ . One has

$$\mathbf{c} = F_2 (F_2^T Q F_2 + n\lambda I)^{-1} F_2^T \mathbf{Y},$$

$$\mathbf{d} = R^{-1} (F_1^T \mathbf{Y} - F_1^T Q \mathbf{c}).$$

Slide 2

### Smoothing Matrix

Denote  $\hat{\mathbf{Y}} = (\eta_\lambda(x_1), \dots, \eta_\lambda(x_n))^T = Q\mathbf{c} + S\mathbf{d}$  and  $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = n\lambda\mathbf{c}$ .  $\hat{\mathbf{Y}} = A(\lambda)\mathbf{Y}$  and  $\mathbf{e} = (I - A(\lambda))\mathbf{Y}$ , where

$$A(\lambda) = I - n\lambda F_2 (F_2^T Q F_2 + n\lambda I)^{-1} F_2^T$$

Slide 3

is known as the *smoothing matrix*. It will be seen that the diagonals  $a_{i,i}$  of  $A(\lambda)$  play important roles in various places. The eigenvalues of  $A(\lambda)$  are in the range  $[0, 1]$ . An alternative expression is

$$A(\lambda) = I - n\lambda (M^{-1} - M^{-1} S (S^T M^{-1} S)^{-1} S^T M^{-1}),$$

where  $M = Q + n\lambda I$ .

- The hat matrix  $H = X(X^T X)^{-1} X^T$  of an ordinary LS regression  $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$  has eigenvalues in  $\{0, 1\}$ .

### Weighted Least Squares

For  $\epsilon_i$  having unequal variance with known ratios, minimize

$$\frac{1}{n} \sum_{i=1}^n w_i (Y_i - \eta(x_i))^2 + \lambda J(\eta).$$

A solution can be obtained by solving

$$(Q_w + n\lambda I)\mathbf{c}_w + S_w \mathbf{d} = \mathbf{Y}_w,$$

$$S_w^T \mathbf{c}_w = 0,$$

Slide 4

where  $Q_w = W^{1/2} Q W^{1/2}$ ,  $\mathbf{c}_w = W^{-1/2} \mathbf{c}$ ,  $S_w = W^{1/2} S$ , and  $\mathbf{Y}_w = W^{1/2} \mathbf{Y}$ , for  $W = \text{diag}(w_i)$ .

Write  $\hat{\mathbf{Y}}_w = W^{1/2} \hat{\mathbf{Y}} = A_w(\lambda) \mathbf{Y}_w$  and  $\mathbf{e}_w = \mathbf{Y}_w - \hat{\mathbf{Y}}_w$ ;  $\mathbf{e}_w = n\lambda \mathbf{c}_w$ .

$$\begin{aligned} A_w(\lambda) &= I - n\lambda F_2 (F_2^T Q_w F_2 + n\lambda I)^{-1} F_2^T \\ &= I - n\lambda (M_w^{-1} - M_w^{-1} S_w (S_w^T M_w^{-1} S_w)^{-1} S_w^T M_w^{-1}), \end{aligned}$$

where  $F_2^T F_2 = I$ ,  $F_2^T S_w = 0$ , and  $M_w = Q_w + n\lambda I$ .

### Smoothing Parameter Selection

As an estimate of  $\eta$  based on data collected from  $x_i$ ,  $i = 1, \dots, n$ , the performance of  $\eta_\lambda$  is to be assessed via the loss

$$L(\lambda) = n^{-1} \sum_{i=1}^n (\eta_\lambda(x_i) - \eta(x_i))^2.$$

Slide 5

To select a  $\lambda$  that nearly minimizes  $L(\lambda)$ , one may use the minimizer of Mallows'  $C_L$ ,

$$U(\lambda) = \frac{1}{n} \mathbf{Y}^T (I - A(\lambda))^2 \mathbf{Y} + 2 \frac{\sigma^2}{n} \text{tr} A(\lambda).$$

◇ **Theorem:** If  $nR(\lambda) \rightarrow \infty$  as  $n \rightarrow \infty$ , where  $R(\lambda) = E[L(\lambda)]$ , then  $U(\lambda) - L(\lambda) - n^{-1} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = o_p(L(\lambda))$ .

- $U(\lambda)$  assumes a known  $\sigma^2$ .

### Generalized Cross-Validation

If validation data were available,  $Y_i^* = \eta(x_i) + \epsilon_i^*$ , one may minimize  $n^{-1} \sum_{i=1}^n (\eta_\lambda(x_i) - Y_i^*)^2$ . Lacking  $Y_i^*$ , one may cross-validate via

$$V_0(\lambda) = \frac{1}{n} \sum_{i=1}^n (\eta_\lambda^{[i]}(x_i) - Y_i)^2,$$

where  $\eta_\lambda^{[k]}$  minimizes the “delete-one” PLS functional

$$\frac{1}{n} \sum_{i \neq k} (Y_i - \eta(x_i))^2 + \lambda J(\eta).$$

It can be shown that  $\eta_\lambda^{[i]}(x_i) - Y_i = (\eta_\lambda(x_i) - Y_i)/(1 - a_{i,i})$ , so

$$V_0(\lambda) = \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \eta_\lambda(x_i))^2}{(1 - a_{i,i})^2}.$$

Replacing  $a_{i,i}$  by their average, one has generalized cross-validation,

$$V(\lambda) = \frac{n^{-1} \mathbf{Y}^T (I - A(\lambda))^2 \mathbf{Y}}{\{n^{-1} \text{tr}(I - A(\lambda))\}^2}.$$

Slide 6

### Optimality of GCV, Variance Estimate

◇ **Theorem:** If  $nR(\lambda) \rightarrow \infty$  and  $\{n^{-1}\text{tr}A(\lambda)\}^2/n^{-1}\text{tr}A^2(\lambda) \rightarrow 0$  as  $n \rightarrow \infty$ , then  $V(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\lambda))$ .

• The theorem does not need normality, only uniformly bounded fourth moments of  $\epsilon_i$ .

A good estimate of  $\sigma^2 = \text{var}[\epsilon_i]$  is given by

$$\hat{\sigma}_v^2 = \frac{\mathbf{Y}^T(I - A(\lambda_v))^2\mathbf{Y}}{\text{tr}(I - A(\lambda_v))},$$

where  $\lambda_v$  minimizes  $V(\lambda)$ . The estimate is asymptotically consistent, along with many others. Its excellent finite sample performance, however, is not widely shared.

Slide 7

### Restricted Maximum Likelihood

Under the Bayes model,  $Y_i = \sum_{\nu=1}^m d_\nu \phi_\nu(x_i) + \eta_1(x_i) + \epsilon_i$ , with  $\epsilon_i \sim N(0, \sigma^2)$  and  $E[\eta_1(x)\eta_1(y)] = bR_J(x, y)$ , consider the likelihood of  $\mathbf{Z} = F_2^T \mathbf{Y}$ , where  $F_2^T S = 0$  kills the fixed effects.

The minus log likelihood of  $(\sigma^2, b)$  based on  $\mathbf{Z}$  is given by

$$\frac{1}{2b} \mathbf{Z}^T(Q^* + n\lambda I)^{-1}\mathbf{Z} + \frac{1}{2} \log |Q^* + n\lambda I| + \frac{n-m}{2} \log b,$$

where  $Q^* = F_2^T Q F_2$  and  $n\lambda = \sigma^2/b$ . Plugging in

$\hat{b} = \mathbf{Z}^T(Q^* + n\lambda I)^{-1}\mathbf{Z}/(n - m)$ , some algebra shows that the profile likelihood of  $\lambda$  is monotone in

$$M(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda))\mathbf{Y}}{|I - A(\lambda)|_+^{1/(n-m)}},$$

where  $|B|_+$  is the product of positive eigenvalues of  $B$ . The variance estimate is given by  $\hat{\sigma}_m^2 = \mathbf{Y}^T(I - A(\lambda_m))\mathbf{Y}/(n - m)$ .

Slide 8

**Performances of  $U(\lambda)$ ,  $V(\lambda)$ , and  $M(\lambda)$**

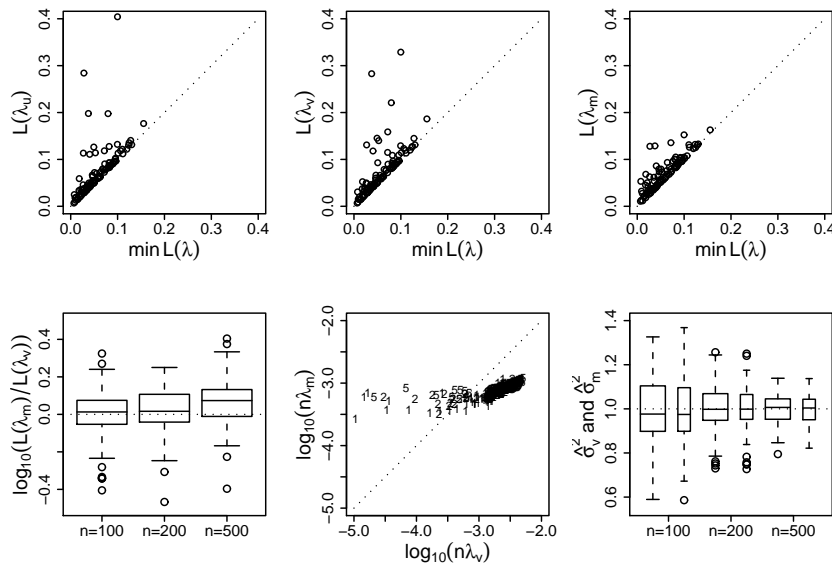
One hundred replicates of samples of size  $n = 100$  were generated from  $Y_i = \eta(x_i) + \epsilon_i$ ,  $x_i = (i - 0.5)/n$ ,  $i = 1, \dots, n$ , where  $\eta(x) = 1 + 3 \sin(2\pi x - \pi)$  and  $\epsilon_i \sim N(0, 1)$ .

Slide 9

Cubic smoothing splines were calculated with  $\lambda$  minimizing  $U(\lambda)$ ,  $V(\lambda)$ , and  $M(\lambda)$ , and with  $\lambda$  on the grid  $\log_{10} n\lambda = (-6)(0.1)(0)$ . The mean square error  $L(\lambda) = n^{-1} \sum_{i=1}^n (\eta_\lambda(x_i) - \eta(x_i))^2$  was calculated for all estimates.

Part of the simulation was repeated for sample sizes  $n = 200$  and  $n = 500$ , each with 100 replicates. The variance estimates  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_m^2$  were also calculated.

**Performances of  $U(\lambda)$ ,  $V(\lambda)$ , and  $M(\lambda)$**



Slide 10

### **$U(\lambda)$ , $V(\lambda)$ , and $M(\lambda)$ for Weighted Data**

For weighted data with  $E[\epsilon_i^2] = \sigma^2/w_i$ , one may use the loss  $L_w(\lambda) = n^{-1} \sum_{i=1}^n w_i (\eta_\lambda(x_i) - \eta(x_i))^2$ .  $C_L$  and GCV are given by

$$U_w(\lambda) = \frac{1}{n} \mathbf{Y}_w^T (I - A_w(\lambda))^2 \mathbf{Y}_w + 2 \frac{\sigma^2}{n} \text{tr} A_w(\lambda),$$

$$V_w(\lambda) = \frac{n^{-1} \mathbf{Y}_w^T (I - A_w(\lambda))^2 \mathbf{Y}_w}{\{n^{-1} \text{tr}(I - A_w(\lambda))\}^2}.$$

Slide 11

Under conditions, one has  $U_w(\lambda) - L_w(\lambda) - n^{-1} \boldsymbol{\epsilon}^T W \boldsymbol{\epsilon} = o_p(L_w(\lambda))$  and  $V_w(\lambda) - L_w(\lambda) - n^{-1} \boldsymbol{\epsilon}^T W \boldsymbol{\epsilon} = o_p(L_w(\lambda))$ .

Under the Bayes model, one may work on the likelihood of the contrasts of  $\mathbf{Y}_w$ , leading to the REML score

$$M_w(\lambda) = \frac{n^{-1} \mathbf{Y}_w^T (I - A_w(\lambda)) \mathbf{Y}_w}{|I - A_w(\lambda)|_+^{1/(n-m)}}.$$

### **$U(\lambda)$ , $V(\lambda)$ , and $M(\lambda)$ for Replicated Data**

Consider  $Y_{i,j} = \eta(x_i) + \epsilon_{i,j}$ , where  $i = 1, \dots, n$ ,  $j = 1, \dots, w_i$ , and  $\epsilon_{i,j} \sim N(0, \sigma^2)$ . One has

$$\sum_{i=1}^n \sum_{j=1}^{w_i} (Y_{i,j} - \eta(x_i))^2 = \sum_{i=1}^n w_i (\bar{Y}_i - \eta(x_i))^2 + \sum_{i=1}^n \sum_{j=1}^{w_i} (Y_{i,j} - \bar{Y}_i)^2,$$

Slide 12

where  $\bar{Y}_i = \sum_{j=1}^{w_i} Y_{i,j}/w_i$ . The size  $N = \sum_{i=1}^n w_i$  PLS is equivalent to a size  $n$  weighted PLS. Write  $\tilde{\sigma}^2 = \sum_i \sum_j (Y_{i,j} - \bar{Y}_i)^2 / (N - n)$ .

In terms of  $\mathbf{Y}_w$  and  $A_w(\lambda)$  in the weighted PLS, one has

$$U(\lambda) = \frac{1}{N} \mathbf{Y}_w^T (I_n - A_w(\lambda))^2 \mathbf{Y}_w + 2 \frac{\tilde{\sigma}^2}{N} \text{tr} A_w(\lambda) + \frac{N - n}{N} \tilde{\sigma}^2,$$

$$V(\lambda) = \frac{N^{-1} \{ \mathbf{Y}_w^T (I_n - A_w(\lambda))^2 \mathbf{Y}_w + (N - n) \tilde{\sigma}^2 \}}{\{1 - N^{-1} \text{tr} A_w(\lambda)\}^2},$$

$$M(\lambda) = \frac{N^{-1} \{ \mathbf{Y}_w^T (I_n - A_w(\lambda)) \mathbf{Y}_w + (N - n) \tilde{\sigma}^2 \}}{|I_n - A_w(\lambda)|_+^{1/(N-m)}}.$$

### Posterior Moments under Bayes Model

Consider the Bayes model,  $Y_i = \sum_{\nu=1}^m \psi_{\nu}(x_i) + \sum_{\beta=1}^p \eta_{\beta}(x_i) + \epsilon_i$ , where  $\psi_{\nu}$  diffuse in  $\text{span}\{\phi_{\nu}\}$  and  $E[\eta_{\beta}(x)\eta_{\beta}(y)] = b\theta_{\beta}R_{\beta}(x, y)$ , independent of each other. It can be shown that

$$E[\psi_{\nu}(x)|\mathbf{Y}] = \phi_{\nu}(x)\mathbf{e}_{\nu}^T \mathbf{d}, \quad E[\eta_{\beta}(x)|\mathbf{Y}] = \boldsymbol{\xi}_{\beta}^T \mathbf{c},$$

$$b^{-1}\text{Cov}[\psi_{\nu}(x), \psi_{\mu}(x)|\mathbf{Y}] = \phi_{\nu}(x)\phi_{\mu}(x)\mathbf{e}_{\nu}^T (S^T M^{-1} S)^{-1} \mathbf{e}_{\mu},$$

$$b^{-1}\text{Cov}[\psi_{\nu}(x), \eta_{\beta}(x)|\mathbf{Y}] = -\phi_{\nu}\mathbf{e}_{\nu}^T \mathbf{d}_{\beta},$$

$$b^{-1}\text{Cov}[\eta_{\beta}(x), \eta_{\gamma}(x)|\mathbf{Y}] = \theta_{\beta}R_{\beta}(x, x)\delta_{\beta,\gamma} - \mathbf{c}_{\beta}^T \boldsymbol{\xi}_{\gamma},$$

where  $(\mathbf{c}, \mathbf{d})$  minimize  $(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T (\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda\mathbf{c}^T Q\mathbf{c}$  and  $(\mathbf{c}_{\beta}, \mathbf{d}_{\beta})$  minimize the same expression but with  $\boldsymbol{\xi}_{\beta} = \theta_{\beta}R_{\beta}(\mathbf{x}, x)$  replacing  $\mathbf{Y}$ . Remember that  $S_{i,\nu} = \phi_{\nu}(x_i)$ ,  $Q_{i,j} = \sum_{\beta} \theta_{\beta}R_{\beta}(x_i, x_j)$ , and  $M = Q + n\lambda Q$ .

Slide 13

### Posterior Moments under Bayes Model

For the overall function  $\eta = \sum_{\nu=1}^m \psi_{\nu} + \sum_{\beta=1}^p \eta_{\beta}$ , one has

$$b^{-1}\text{Var}[\eta(x)|\mathbf{Y}] = R_J(x, x) + \boldsymbol{\phi}^T (S^T M^{-1} S)^{-1} \boldsymbol{\phi} - 2\boldsymbol{\phi}^T \mathbf{d}_{\xi} - \boldsymbol{\xi}^T \mathbf{c}_{\xi},$$

where  $(\mathbf{c}_{\xi}, \mathbf{d}_{\xi})$  minimize the PLS score with  $\boldsymbol{\xi} = \sum_{\beta} \theta_{\beta}R_{\beta}(\mathbf{x}, x)$  replacing  $\mathbf{Y}$ . This can be obtained by summing terms.

On the sampling points  $x_i$ , it can be shown that

$$\text{Var}[\eta(x_i)|\mathbf{Y}] = \sigma^2 a_{i,i}; \text{ remember that } \sigma^2 = (n\lambda)b.$$

For weighted data with  $E[\epsilon_i^2] = \sigma^2/w_i$ , replace in the formulas  $S$ ,  $Q$ ,  $\mathbf{Y}$ ,  $\boldsymbol{\xi}$ ,  $\mathbf{c}$ , and  $A(\lambda)$  by  $S_w$ ,  $Q_w$ ,  $\mathbf{Y}_w$ ,  $W^{1/2}\boldsymbol{\xi}$ ,  $W^{-1/2}\mathbf{c}$ , and  $W^{-1/2}A_w(\lambda)W^{-1/2}$ , respectively.

Slide 14

### Bayesian Confidence Intervals

Given  $\sigma^2$  and  $n\lambda$ , Bayesian confidence intervals of  $\eta(x)$  and its components at any  $x \in \mathcal{X}$  can be constructed based on the posterior moments.

Using  $\lambda_v$  and  $\sigma_v^2$ , the point-wise construction demonstrates an across-the-function coverage property. Define

$$\text{ACP}(\alpha) = \frac{1}{n} \#\{i : |\eta_{\lambda_v}(x_i) - \eta(x_i)| < z_{\alpha/2} \hat{\sigma}_v \sqrt{a_{i,i}}\}.$$

One has  $E[\text{ACP}(\alpha)] \approx 1 - \alpha$  for  $n$  large. The coverage property does not hold for component-wise intervals.

The coverage property is largely based on heuristic analysis and numerical simulations, though more rigorous treatment is available for polynomial smoothing splines.

Slide 15

### Computation: Generic Algorithms

Generic algorithms have been developed to solve the linear system

$$\begin{aligned} (Q + n\lambda I)\mathbf{c} + S\mathbf{d} &= \mathbf{Y}, \\ S^T \mathbf{c} &= 0, \end{aligned}$$

The key is to efficiently evaluate  $U(\lambda)$ ,  $V(\lambda)$ , or  $M(\lambda)$ .

For a single  $\lambda$ , after a one-time tridiagonalization of  $F_2^T Q F_2$  costing  $(4/3)n^3 + O(n^2)$  flops, the scores are available in  $O(n)$  flops.

With  $\theta_\beta$  hidden in  $Q$ , gradient and Hessian are available in  $(p-1)(4/3)n^3 + O(n^2)$  extra flops to facilitate a Newton-type iteration.

- Fixing  $\lambda$  and  $\theta_\beta$ , it takes  $n^3/3 + O(n^2)$  flops to solve the system.

Slide 16

### Software: R Package gss

The generic algorithms have been implemented in a collection of RATFOR routines known as RKPACk. A user-friendly front end is available through the `ssanova` suite in the R package `gss`.

Slide 17

The usage of `ssanova` and the affiliated methods `predict`, `summary`, `residuals`, and `fitted` are similar to that of the `lm` suite for standard linear regression. Implemented are cubic and linear splines, with the terms in ANOVA decomposition satisfying the integration side conditions.

- For polynomial splines,  $O(n)$  algorithms exist for the calculation of  $\eta_\lambda$ , but not for posterior variance.
- Faster computation is possible for lower-dimensional approximations of  $\eta_\lambda$ .

### Cosine Diagnostics

Consider  $\eta = \sum_{\beta=0}^p f_\beta$ , where  $f_0 \propto 1$ . Evaluating a fit at  $x_i$ , one has  $\mathbf{Y} = \mathbf{f}_0 + \mathbf{f}_1 + \cdots + \mathbf{f}_p + \mathbf{e}$ , where  $\mathbf{f}_\beta = (f_\beta(x_1), \dots, f_\beta(x_n))^T$ . Projecting onto  $\{\mathbf{1}\}^\perp$ , one gets  $\mathbf{Y}^* = \mathbf{f}_1^* + \cdots + \mathbf{f}_p^* + \mathbf{e}^*$ . A set of diagnostics are largely based on cosines among the vectors.

Slide 18

- The collinearity indices  $\kappa_\beta$  of  $(\mathbf{f}_1^*, \dots, \mathbf{f}_p^*)$  (i.e.,  $\sqrt{\text{VIF}_\beta}$ ) indicates identifiability problems.
- The “SS decomposition”  $\pi_\beta = (\mathbf{f}_\beta^*)^T \hat{\mathbf{Y}}^* / \|\hat{\mathbf{Y}}^*\|^2$ , where  $\hat{\mathbf{Y}}^* = \mathbf{f}_1^* + \cdots + \mathbf{f}_p^*$ , gives the relative magnitudes of terms.
- A small  $\cos(\mathbf{f}_\beta^*, \mathbf{Y}^*)$  or a large  $\cos(\mathbf{f}_\beta^*, \mathbf{e}^*)$  make  $f_\beta$  suspect, so does a very small  $\|\mathbf{f}_\beta^*\|$  compared to  $\|\mathbf{Y}^*\|$ .
- The signal-to-noise ratio may be measured by  $\cos(\mathbf{Y}^*, \mathbf{e}^*)$  or  $R^2 = \|\mathbf{Y}^* - \mathbf{e}^*\|^2 / \|\mathbf{Y}^*\|^2$ .

### Fast Algorithm for Polynomial Splines

Polynomial splines with knots  $\xi_1 < \dots < \xi_q$  form a linear space of dimension  $q$ . There exists a *local-support* basis  $\{B_j(x)\}_{j=1}^q$ , with each  $B_j$  supported on at most  $2m$  adjacent intervals  $[0, \xi_1], [\xi_1, \xi_2], \dots, [\xi_q, 1]$ , and at most  $2m$   $B_j$ 's are nonzero at any  $x \in [0, 1]$ .

Slide 19

Plugging  $\eta(x) = \sum_{j=1}^q c_j B_j(x)$  into PLS, one has

$$(\mathbf{Y} - X\mathbf{c})^T(\mathbf{Y} - X\mathbf{c}) + n\lambda\mathbf{c}^T J\mathbf{c},$$

where  $X$  is  $n \times q$  with the  $(i, j)$ th entry  $B_j(x_i)$  and  $J$  is  $q \times q$  with the  $(i, j)$ th entry  $\int_0^1 B_i^{(m)} B_j^{(m)} dx$ . Sorting  $B_j$ 's by their supports,  $X^T X + n\lambda I$  is *banded*.

Banded Cholesky decomposition can be used to obtain

$\mathbf{c} = (X^T X + n\lambda J)^{-1} X^T \mathbf{Y}$ , and a recursive scheme can be used to evaluate  $\text{tr}A(\lambda) = \text{tr}\{(X^T X + n\lambda J)^{-1}(X^T X)\}$ , with  $O(q)$  flops.

### Monte Carlo Cross-Validation

Many smoothing problems can be solved through systems of the form  $(X^T X + n\lambda J)\mathbf{c} = X^T \mathbf{Y}$ , although a banded matrix isn't always available. With a sparse  $(X^T X + n\lambda J)$ , one may cash on a fast multiplication  $(X^T X + n\lambda J)\mathbf{c}$  to solve for  $\mathbf{c}$  iteratively.  $A(\lambda)$  is lost in the iterations, however.

Slide 20

For  $\boldsymbol{\epsilon} \sim N(0, I)$ ,  $E[\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon}] = \text{tr}A(\lambda)$ , and one may generate  $\boldsymbol{\epsilon} \sim N(0, I)$  and run the iterations with  $\boldsymbol{\epsilon}$  replacing  $\mathbf{Y}$ , and estimate  $\text{tr}A(\lambda)$  by  $\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon}$ . The optimality of  $U(\lambda)$  and  $V(\lambda)$  is intact with estimated  $\text{tr}A(\lambda)$ .

- Use the same  $\boldsymbol{\epsilon}$  for all  $\lambda$  to get smooth scores  $U(\lambda)$  or  $V(\lambda)$ .
- $\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon}/\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$  is a better estimator of  $\text{tr}A(\lambda)$  than  $\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon}$ .