General Formulation

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

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

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

$J(f, g) = \sum_{\beta=1}^{p} \theta_{\beta}^{-1} \langle f, g \rangle_{\beta}$ is an IP in $H_J = \oplus_{\beta=1}^{p} H_{\beta}$, with an RK $R_J(x, y) = \sum_{\beta=1}^{p} \theta_{\beta} R_{\beta}(x, y)$ and a null space $N_J = H_0$. The minimizer of PLS is of the form

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

where $\{\phi_{\nu}\}_{\nu=1}^{m}$ is a basis of $N_J = H_0$.

Numerical Problem

Plugging in the solution expression, PLS becomes

$$(Y - Sd - Qc)^T (Y - Sd - Qc) + n\lambda c^T Qc,$$

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 $(c, d)$ may not be. The linear system

$$(Q + n\lambda I)c + Sd = Y,$$

$$S^T c = 0$$

yields a solution, which is all one needs.

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

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

$$d = R^{-1} (F_1^T Y - F_1^T Qc).$$
Smoothing Matrix

Denote \( \hat{Y} = (\eta_\lambda(x_1), \ldots, \eta_\lambda(x_n))^T = Qc + Sd \) and 
\( e = Y - \hat{Y} = n\lambda c. \hat{Y} = A(\lambda)Y \) and 
\( e = (I - A(\lambda))Y \), where 
\[ A(\lambda) = I - n\lambda F_2(F_2^T Q F_2 + n\lambda I)^{-1} F_2^T \]
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 
\( Y = X\beta + \epsilon \) has eigenvalues in \([0, 1]\).

Weighted Least Squares

For \( \epsilon_i \) having unequal variances but 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 
\[ (Q_w + n\lambda I)c_w + S_w d = Y_w, \]
\[ S_w^T c_w = 0, \]
where \( Q_w = W^{1/2} Q W^{1/2} \), \( c_w = W^{-1/2} c \), \( S_w = W^{1/2} S \), and 
\( Y_w = W^{1/2} Y \), for \( W = \text{diag}(w_i) \).

Write \( \hat{Y}_w = W^{1/2} \hat{Y} = A_w(\lambda)Y_w \) and \( e_w = Y_w - \hat{Y}_w \); 
\( e_w = n\lambda c_w \).
\[ A_w(\lambda) = I - n\lambda F_2(F_2^T Q_w F_2 + n\lambda I)^{-1} F_2^T \]
\[ = I - n\lambda (M_w^{-1} - M_w^{-1} S_w (S_w^T M_w^{-1} S_w)^{-1} S_w^T M_w^{-1}), \]
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 $x_i, i = 1, \ldots, n$, the performance of $\eta_\lambda$ is to be assessed via the loss

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

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

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

\begin{itemize}
  \item $U(\lambda)$ assumes a known $\sigma^2$.
\end{itemize}

\begin{itemize}
  \item Theorem: If $nR(\lambda) \to \infty$ as $n \to \infty$, where $R(\lambda) = E[L(\lambda)]$, then

$$U(\lambda) - L(\lambda) - n^{-1} \epsilon^T \epsilon = o_p(L(\lambda)).$$

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 $\eta_\lambda^{[i]}$ 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,

$$V(\lambda) = \frac{n^{-1} Y^T (I - A(\lambda)) Y}{\{n^{-1} \text{tr}(I - \alpha A(\lambda))\}^2}, \text{ for } \alpha = 1.$$
Optimality of GCV, Variance Estimate

\begin{itemize}
  \item \textbf{Theorem}: If \( nR(\lambda) \to \infty \) and \( \left\{ n^{-1}\text{tr}A(\lambda) \right\}^2/n^{-1}\text{tr}A^2(\lambda) \to 0 \) as \( n \to \infty \), then \( V(\lambda) - L(\lambda) - n^{-1}\epsilon^T\epsilon = o_p(L(\lambda)) \).
  
  \item The theorem does not need normality, only uniformly bounded fourth moments of \( \epsilon_i \).
\end{itemize}

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

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

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.

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 \( Z = F_2^T Y \), where \( F_2^T S = 0 \) kills the fixed effects.

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

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

where \( Q^* = F_2^T Q F_2 \) and \( n\lambda = \sigma^2/b \). Plugging in \( \hat{b} = Z^T(Q^* + n\lambda I)^{-1}Z/(n - m) \), some algebra shows that the profile likelihood of \( \lambda \) is monotone in

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

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

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

Cubic smoothing splines were fitted with $\lambda$ minimizing $U(\lambda)$, $V(\lambda)$ with $\alpha = 1, 1.4$, and $M(\lambda)$, and with $\lambda$ on $\log_{10} n \lambda = (-6)(0.1)(0)$. 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.
\[ U(\lambda), V(\lambda), \text{ and } M(\lambda) \text{ for Weighted Data} \]

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

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

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

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

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

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

\[ U(\lambda), V(\lambda), \text{ and } M(\lambda) \text{ for Replicated Data} \]

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

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

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 \( \bar{\sigma}^2 = \sum_i \sum_j (Y_{i,j} - \bar{Y}_i)^2/(N - n) \).

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

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

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

\[
M(\lambda) = \frac{N^{-1} \{ \mathbf{Y}_w^T (I_n - A_w(\lambda)) \mathbf{Y}_w + (N - n)\bar{\sigma}^2 \}}{|I_n - A_w(\lambda)|^{1/(N-m)}}.
\]
**Bayesian Confidence Intervals**

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 \( \psi_{\nu} \) is diffuse in span\{\( \phi_{\nu} \)\} and \( E[\eta_{\beta}(x)\eta_{\beta}(y)] = b\theta_{\beta}R_{\beta}(x,y) \), independent of each other, where \( b = n\lambda/\sigma^2 \). The PLS estimate

\[
\hat{\eta}(x) = \sum_{\nu} d_{\nu} \phi_{\nu}(x) + \sum_i c_i (\sum_{\beta} \theta_{\beta} R_{\beta}(x_i, x)) = \phi^T d + \sum_{\beta} \xi_{\beta}^T c
\]

gives the posterior mean of \( \sum_{\nu} \psi_{\nu}(x) + \sum_{\beta} \eta_{\beta}(x) \). In fact, \( E[\psi_{\nu}(x)|Y] = d_{\nu} \phi_{\nu}(x) \) and \( E[\eta_{\beta}(x)|Y] = \xi_{\beta}^T c \). One may also calculate the posterior (co)variances of \( \psi_{\nu}(x) \) and \( \eta_{\beta}(x) \). Bayesian confidence intervals can be constructed for sums of \( \psi_{\nu}(x), \eta_{\beta}(x) \).

For the overall function \( \eta = \sum_{\nu} \psi_{\nu} + \sum_{\beta} \eta_{\beta} \) on the sampling points, \( \text{Var} [\eta(x_i)|Y] = \sigma^2 a_{i,i} \) for \( a_{i,i} \) the \( i \)th diagonal of \( A(\lambda) \).

For \( E[\epsilon_i^2] = \sigma^2/w_i \), posterior means are given by PWLS estimate, and \( \text{Var} [\eta(x_i)|Y] = \sigma^2 a_{i,i}/w_i \) for \( a_{i,i} \) the \( i \)th diagonal of \( A_w(\lambda) \).

**Efficient Approximation**

Under mild conditions, the minimizer \( \hat{\eta} \) of \( \sum_i (Y_i - \eta(x_i)) + \lambda J(\eta) \) converges to the truth \( \eta_0 \) at a rate \( V(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p) \) as \( n \to \infty \) and \( \lambda \to 0 \), where \( V(g) = \int_\mathcal{X} g^2(x)f(x)dx \) for \( f(x) \) the limiting density of \( x_i, r > 1 \) codes the smoothness of functions in \( \mathcal{H} = \{ f : J(f) < \infty \} \), and \( p \in [1,2] \) depends on how smooth \( \eta_0 \) is.

For \( J(f) = \int_0^1 \int_0^1 f^2 \) on \( \mathcal{X} = [0,1], r = 4, \) and \( p = 2 \) if \( \int_0^1 (\eta_0^{(4)})^2 < \infty \).

Consider the minimizer \( \hat{\eta}^* \) of PLS in a space

\( \mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, x), j = 1, \ldots, q\}, \)

where \( \{z_j\} \subseteq \{x_i\} \) is a random subset; \( \hat{\eta}^* \) has the same convergence rate as \( \hat{\eta} \) when \( q\lambda^{2/r} \to \infty \). For \( r = 4 \) and \( p = 2 \), the optimal rate \( O_p(n^{-8/9}) \) is attained at \( \lambda \asymp n^{-4/9} \), thus one needs \( q \asymp n^{2/9+\epsilon} \), \( \forall \epsilon > 0 \); \( q = n \) gives the exact solution \( \hat{\eta} \).
Efficient Approximation

To compute \( \hat{\eta}^*(x) = \sum_\nu d_\nu \phi_\nu(x) + \sum_j c_j R_J(z_j, x) \), one minimizes
\[
(Y - Sd - Rc)^T(Y - Sd - Rc) + n\lambda c^TQc,
\]
w.r.t. \( c \) and \( d \), where \( S_i,\nu = \phi_\nu(x_i) \), \( R_{i,j} = R_J(x_i, z_j) \), and \( Q_{j,k} = R_J(z_j, z_k) \); \( Q \) is part of \( R \). When \( q = n \), \( R = Q \).

Given \( \{z_j\} \), \( \hat{\eta}^*(x) \) is also a posterior mean in a Bayes model,
\[
Y_i = \sum_{\nu=1}^m \psi_\nu(x_i) + \sum_{\beta=1}^p \eta_\beta(x_i) + \epsilon_i,
\]
where \( \psi_\nu \) is diffuse in span\( \{\phi_\nu\} \) and \( E[\eta_\beta(x)\eta_\gamma(y)] = b\theta_\beta\theta_\gamma R_\beta(x, z^T)Q^+R_\gamma(z, y) \), where \( b = n\lambda/\sigma^2 \) and \( Q^+ \) is the Moore-Penrose inverse of \( Q = R_J(z, z^T) \).

Bayesian confidence intervals can also be constructed for sums of \( \psi_\nu(x) \) and \( \eta_\beta(x) \) under such a Bayes model, and when \( q = n \), the results reduce to the ones discussed earlier.

Computation: Generic Algorithms

Fixing smoothing parameters, one simply solves
\[
\begin{pmatrix} S^T S & S^T R \\ R^T S & R^T R + n\lambda Q \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} S^T Y \\ R^T Y \end{pmatrix},
\]
and to minimize \( U(\lambda) \), \( V(\lambda) \), or \( M(\lambda) \), one needs an quasi-Newton outer loop with numerical derivatives. The computation is \( O(nq^2) \).

For \( q = n \), \( R = Q \), and one may solve the linear system
\[
(Q + n\lambda I)c + Sd = Y,
\]
\[
S^T c = 0,
\]
which can be “diagonalized;” one can use analytical derivatives in the “diagonalized” system to minimize \( U(\lambda) \), \( V(\lambda) \), or \( M(\lambda) \). The computation is \( O(n^3) \), but runs faster at least up to \( n = 1000 \).
Algorithms assuming $R = Q$ are implemented in RKPACK, a collection of RATFOR routines, with a front end in the ssanova0 suite of the R package gss. Algorithms for a general $R$ are implemented in the ssanova suite of gss.

To fit a cubic spline $\hat{\eta}^*(x)$ to $(x_i, Y_i)$ on $\mathcal{X} = [a, b]$, use

```r
fit <- ssanova(y~x,method="v",alpha=1.4)
```

and to evaluate $\hat{\eta}^*(x)$ and standard error on a grid,

```r
est <- predict(fit,data.frame(x=xx),se=TRUE)
```

One may fit $\hat{\eta}(x)$ using ssanova0, but $\alpha$ is not an option.

Due to the random selection of $\{z_j\}$, repeated calls to ssanova return slightly different results.

---

Cosine Diagnostics

Consider $\eta = \sum_{\beta=0}^{p} f_\beta$, where $f_0 \propto 1$. Evaluating a fit at $x_i$, one has $Y = f_0 + f_1 + \cdots + f_p + e$, where $f_\beta = (f_\beta(x_1), \ldots, f_\beta(x_n))^T$.

Projecting onto $\{1\}^\perp$, one gets $Y^* = f_1^* + \cdots + f_p^* + e^*$. A set of diagnostics are largely based on cosines among the vectors.

- The collinearity indices $\kappa_\beta$ of $(f_1^*, \ldots, f_p^*)$ (i.e., $\sqrt{\text{VIF}_\beta}$) indicates identifiability problems.
- The “SS decomposition” $\pi_\beta = (f_\beta^*)^T \hat{Y}^* / \| \hat{Y}^* \|^2$, where $\hat{Y}^* = f_1^* + \cdots + f_p^*$, gives the relative magnitudes of terms.
- A small $\cos(f_\beta^*, Y^*)$ or a large $\cos(f_\beta^*, e^*)$ make $f_\beta$ suspect, so does a very small $\| f_\beta^* \|$ compared to $\| Y^* \|$.
- The signal-to-noise ratio may be measured by $\cos(Y^*, e^*)$ or $R^2 = \| Y^* - e^* \|^2 / \| Y^* \|^2$. 

---

C. Gu Spring 2015
Square Error Projection

To “test” the hypotheses $H_0 : \eta \in \mathcal{H}_0$ vs $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$ with an infinite-dimensional null, say for $\mathcal{H}_0 = \{ \eta : \eta = \eta_0 + \eta_1 + \eta_2 \}$ and $\mathcal{H}_1 = \{ \eta : \eta = \eta_{1,2} \}$, one may use the square error projection.

Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, its square error projection $\tilde{\eta} \in \mathcal{H}_0$ minimizes

$$SE(\hat{\eta}, \eta) = \sum_{i=1}^{n} (\hat{\eta}(x_i) - \eta(x_i))^2$$

over $\eta \in \mathcal{H}_0$. One has $SE(\hat{\eta}, \eta_c) = SE(\hat{\eta}, \tilde{\eta}) + SE(\tilde{\eta}, \eta_c)$, where $\eta_c \in \mathcal{H}_0$ is a degenerate fit such as a constant, and one loses little cutting out $\mathcal{H}_1$ if the ratio $\rho = SE(\hat{\eta}, \tilde{\eta})/SE(\hat{\eta}, \eta_c)$ is small.

The projection is numerically ill-posed in an infinite dimensional $\mathcal{H}_0$, but practically doable “within” $\mathcal{H}^*$.

```r
fit <- ssanova(y~x1*x2)
project(fit, inc=c("x1","x2"))
```

Fast Algorithm for Polynomial Splines

Polynomial splines with knots $\xi_1 < \cdots < \xi_q$ 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 $[0, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_q, 1]$, and at most $2m$ $B_j$’s are nonzero at any $x \in [0, 1]$.

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

$$(Y - Xc)^T (Y - Xc) + n\lambda c^T Jc,$$

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 $c = (X^T X + n\lambda J)^{-1} X^T Y$, 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.