Statistics 860 Lecture 18©G. Wahba 2016

Numerical Methods for Very Large Data Sets

- What do iterative methods for the solution of large linear systems do? Early stopping of iterative methods as a smoothing-regularization method.

- lecture18b: -

- Reprise of SS-ANOVA models in time and space

- Backfitting in Smoothing Spline ANOVA (probably not discussed)

- Iterative imputation as a trick for missing data in regular patterns

Early stopping of iteration as a regularization/tuning method.

- G. Wahba. Three topics in ill posed problems. In H. Engl and C. Groetsch, editors, *Proceedings of the Alpine-U.S. Seminar on Inverse and Ill Posed Problems*, pages 37–51. Academic Press, 1987. `illpose.pdf`

- G. Wahba, D. Johnson, F. Gao, and J. Gong. Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation. *Mon. Wea. Rev.*, 123:3358-3369, 1995. `wahba.johnson.gao.gong.1995.pdf`

References for Backfitting, imputation - (lect18b).

- Z. Luo. Backfitting in smoothing spline ANOVA. *The Annals of Statistics*, 26:1733–1759, 1998. `luo:annstat1998.pdf`

- Wahba, G. and Luo, Z. " Smoothing Spline ANOVA Fits for Very Large, Nearly Regular Data Sets, with Application to Historical Global Climate Data" TR 952, October 1995. Slightly revised version in Annals of Numerical Mathematics 4 (1997) 579-598. (Festschrift in Honor of Ted Rivlin, C.Micchelli, Ed.) `lreg.rev.pdf`

- Luo, Z. , Wahba, G, and Johnson, D. R. " Spatial-Temporal Analysis of Temperature Using Smoothing Spline ANOVA " J. Climate 11, 18-28 (1998). `luo:wahba:johnson:1998.pdf`

Basic References in Matrix Computations and Optimization

- G. Golub and C. VanLoan. *Matrix Computations, Third Edition*. Johns Hopkins University Press, pp694, 1996.

- J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 1999.

Early stopping in the Richardson/Landweber/Fridman/Cimino/Picard iteration.

Solve $Mx = y$ where M is a large, non-negative definite matrix, with eigenvalues λ_ν and eigenvectors u_ν by this iterative method. The kth iterate is

$$x^k = x^{k-1} + \beta M(y - Mx^{k-1}).$$

The desired “exact’ solution is

$$x = M^\dagger y = \sum_{\lambda_\nu \neq 0} \frac{(y, u_\nu)}{\lambda_\nu} u_\nu,$$

where $M^\dagger$ is the Moore-Penrose generalized inverse. If $\beta\lambda_1^2 < 1$, then (in theory) the kth iterate approaches the desired solution as $k \to \infty$.

$$
\begin{aligned}
x^k &= x^{k-1} + \beta M(y - Mx^{k-1}) \\
&= (I - \beta M^2)x^{k-1} + \beta My \\
&= (I - \beta M^2)[(I - \beta M^2)x^{k-2} + \beta My] + \beta My \\
&\vdots
\end{aligned}
$$

giving

$$x^k = (I - \beta M^2)^k \, x^0$$

$$+[(I - \beta M^2)^{k-1} + (I - \beta M^2)^{k-2} + \cdots + I]\beta My.$$

Lemma: (proof later)

$$[(I - \beta M^2)^{k-1} + (I - \beta M^2)^{k-2} + \cdots + I]\beta M^2 =$$

$$I - (I - \beta M^2)^k.$$

Right multiply by $M^\dagger$, use $M^2 M^\dagger = M$ to get:

$$[(I - \beta M^2)^{k-1} + (I - \beta M^2)^{k-2} + \cdots + I]\beta M =$$

$$[I - ((I - \beta M^2)^k]M^\dagger.$$

Proof of Lemma:

For $|\theta| < 1$, we have the familiar formula

$$\frac{1}{1-\theta} = 1 + \theta + \theta^2 + \ldots + \theta^k[1 + \theta + \ldots$$

and setting $\theta = (1-\rho)$ we get

$$\frac{1-(1-\rho)^k}{1-(1-\rho)} = 1 + (1-\rho) + \ldots + (1-\rho)^{k-1}$$

and

$$1-(1-\rho)^k = \left[1 + (1-\rho) + \ldots + (1-\rho)^{k-1}\right]\rho.$$

Let $B = \Gamma D \Gamma'$ with $O \prec B \prec I$. This lets us write

$$I-(I-B)^k = \left[I + (I-B) + \ldots + (I-B)^{k-1}\right]B.$$

Setting $B = \beta M^2$ gives the lemma.

Setting $x^0 = 0$ the result is

$$\begin{aligned} x^k &= (I - (I - \beta M^2)^k) M^\dagger y \\ &= \sum_{\lambda_\nu \neq 0} \left(1 - (1 - \beta \lambda_\nu^2)^k\right) \frac{(y, u_\nu)}{\lambda_\nu} u_\nu \end{aligned}$$

Compare to

$$M^\dagger y = \sum_{\lambda_\nu \neq 0} \frac{(y, u_\nu)}{\lambda_\nu} u_\nu$$

or to a regularized estimate:

$$y = Mx + \epsilon$$

$$(y - Mx)^2 + \lambda x' x$$

gives

$$x = \sum_{\lambda_\nu \neq 0} \left(\frac{\lambda_\nu^2}{\lambda_\nu^2 + \lambda} \right) \frac{(y, u_\nu)}{\lambda_\nu} u_\nu$$

Early Stopping

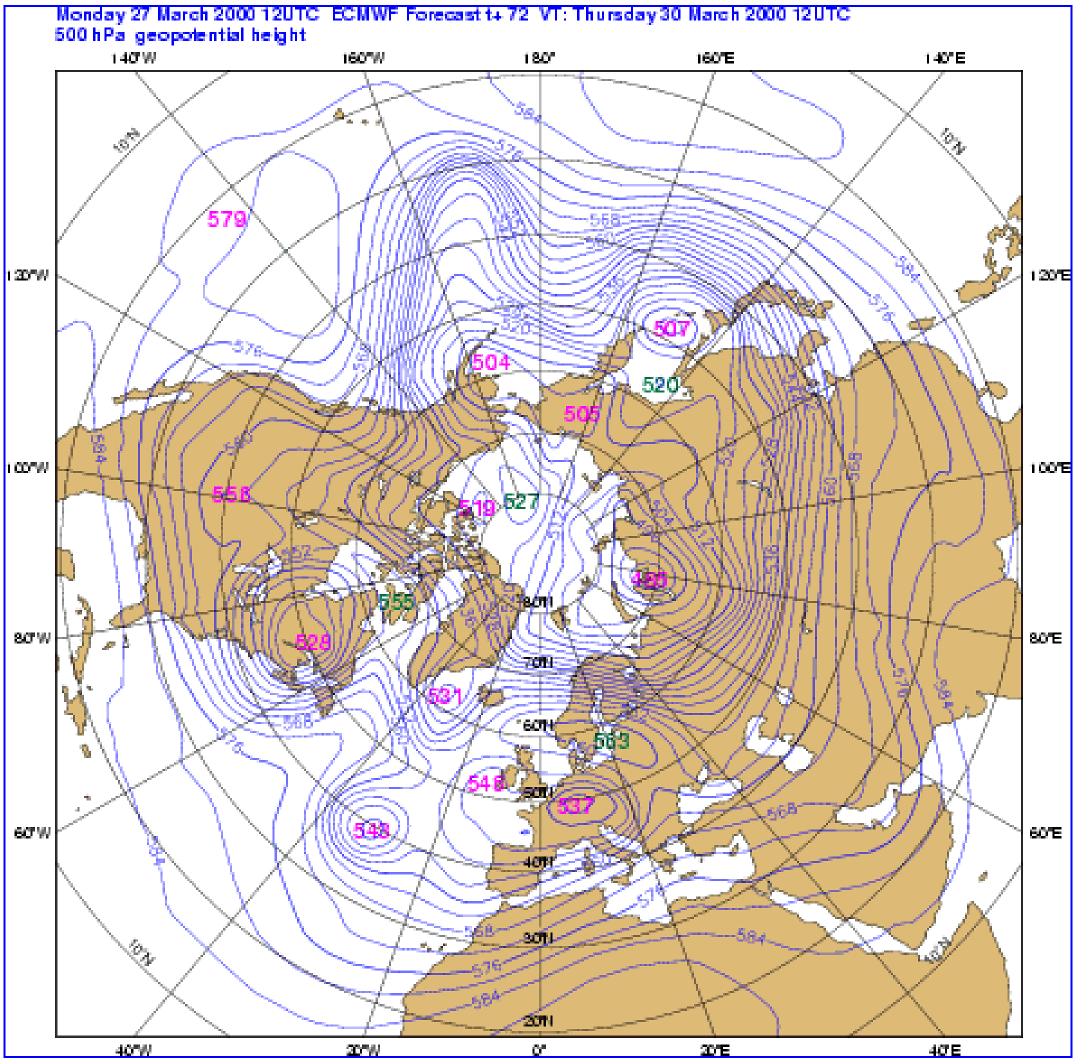
A semi-realistic 'toy' problem to test and demonstrate the feasibility and efficiency of choosing both k and λ via GCV or UBR, in conjunction with the randomized trace estimation. Used pre-conditioned conjugate gradient algorithm, with early stopping. See `cj.pdf` for the conjugate gradient algorithm.

ECMWF Gridded Level IIIB FGGE data for the 500mb height for January 2, 1979, was used to obtain a spherical harmonic representation for the 500mb height field of the form

$$f(P) = \sum_{\ell=0}^{30} \sum_{s=-\ell}^{\ell} x_{\ell s} Y_{\ell s}(P),$$

where P is a point on the sphere, and the $Y_{\ell s}$ are spherical harmonics. This representation was obtained by solving a variational problem given the gridded data. The amount of smoothing was chosen to make the resulting contour plots match the ECMWF plots visually.

500mb heght forecast from ECMWF:



March 27 72 hour forecast 500 hPa geopotential height (in 10's of meters), from ECMWF.

Simulated observational data at $n = 600$ North American radiosonde stations generated by

$$y_i = f(P_i) + \epsilon_i$$

where $\epsilon = (\epsilon_1, ..., \epsilon_{600})' \sim \mathcal{N}(0, \sigma^2 I)$, and the P_i are station locations. $\sigma = 9m.$ is realistic observational error. An approximate spline on the sphere can be obtained by letting $\widehat{x}_\lambda = (\widehat{x}_{00,\lambda}, \widehat{x}_{10,\lambda}, ...)$ be the minimizer of

$$\sum_{i=1}^{n} (y_i - \sum_{\ell=0}^{30} \sum_{s=-\ell}^{\ell} x_{\ell s} Y_{\ell s}(P_i))^2$$

$$+\lambda \sum_{\ell=0}^{30} \sum_{s=-\ell}^{\ell} [(\ell)(\ell+1)]^2 x_{\ell s}^2.$$

The penalty functional $J(f) = \sum_{\ell s} [(\ell)(\ell+1)]^2 x_{\ell s}^2$ is a multiple of $J(f) = \int_{\mathcal{S}} (\triangle f)^2$ where $\triangle$ is the Laplacian on the sphere (see Wahba(1981,1982a))

`sphspl.pdf`.

Letting K be the 600×960 matrix with entries $Y_{\ell s}(P_i)$ and D be the diagonal matrix with $\ell s, \ell s$ entries $[\ell(\ell+1)]^2$, then the minimizer $\widehat{x}_\lambda$ satisfies

$$(K'K + \lambda D)\widehat{x}_\lambda = K'y.$$

A preconditioned conjugate gradient algorithm with (symmetric, invertible) preconditioner C replaces $\widehat{x}_\lambda$ by $C^{-1}w$ and solves for w in

$$C^{-1}(K'K + \lambda D)C^{-1}w = C^{-1}K'y.$$

See Golub and van Loan (1989), Section 10.3, or `cj.pdf`. In the experiment below C was taken as $[diag(K'K + \lambda D)]^{1/2}$.

The predictive mean square error is

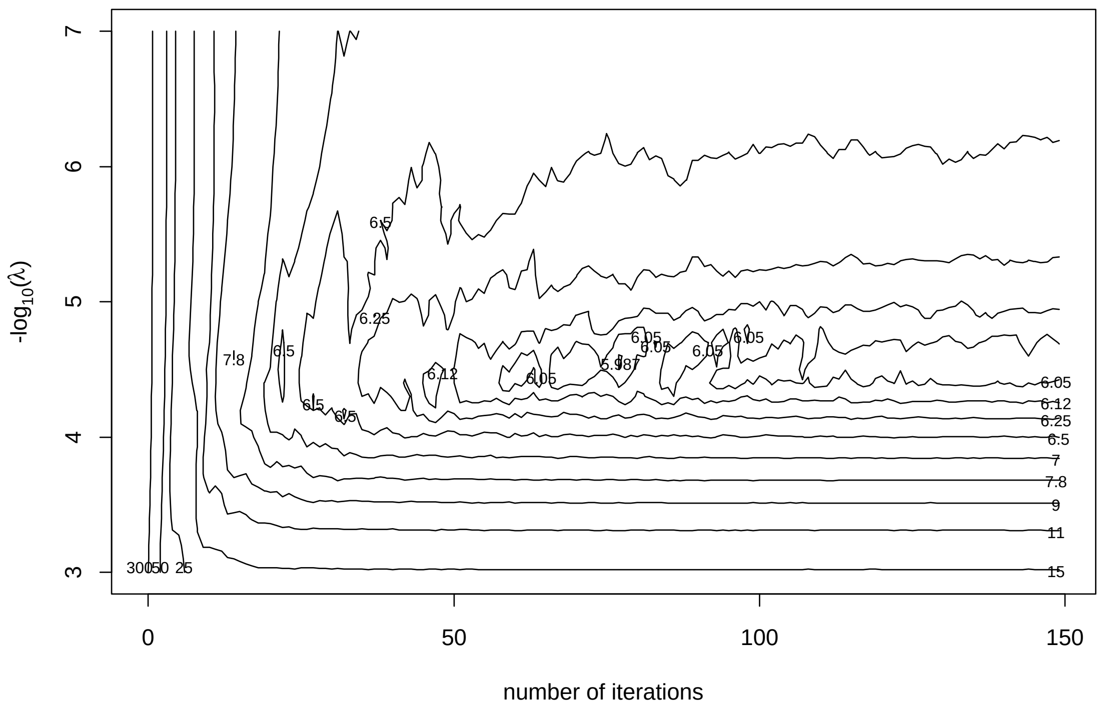
$$R(\lambda, k) = \frac{1}{n}\sum_{i=1}^{n}(f_\lambda^k(P_i) - f(P_i))^2$$

where

$$f_\lambda^k(P) = \sum_{\ell s} \widehat{x}_{\ell s,\lambda}^k Y_{\ell s}(P), \qquad (1)$$

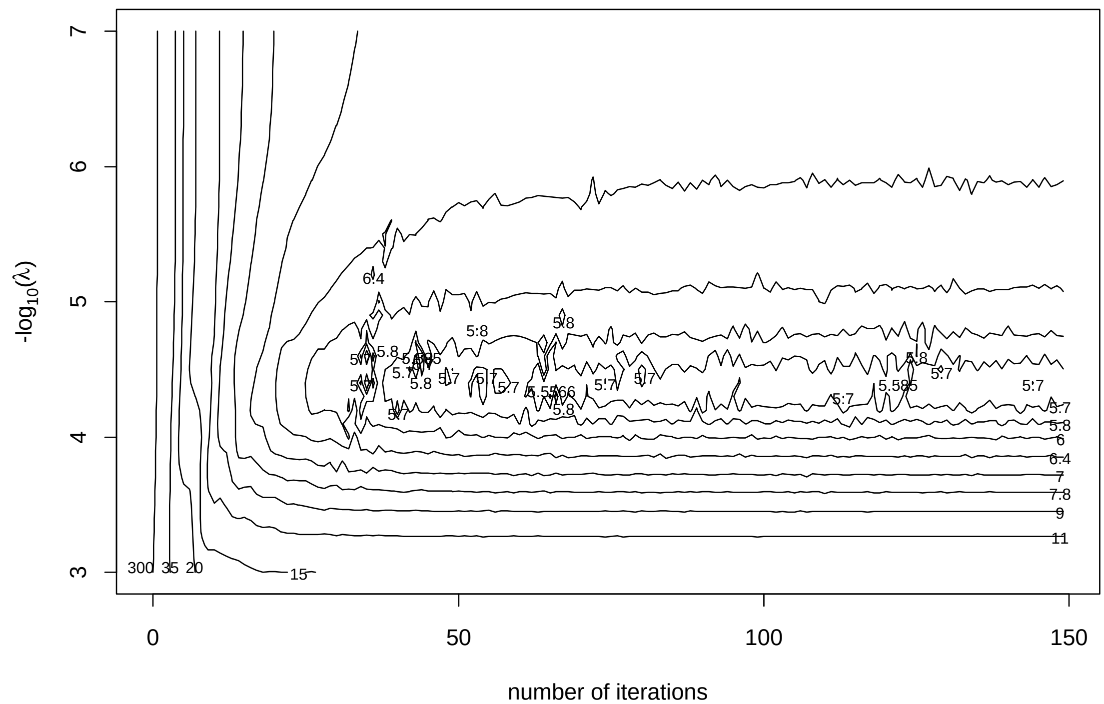
$$\widehat{x}_\lambda^k = \{\widehat{x}_{\ell s,\lambda}^k\}, \qquad (2)$$

and $\widehat{x}_\lambda^k$ is the approximate solution after k iterations.



The root predictive mean square error $R^{1/2}$ as a function of $\log_{10}(\lambda)$ and k, where k is the number of iterations in the cj iterative solution.

$R(\lambda, k)$ is minimized at around $-log_{10}(\lambda) = 4.5$, and $k = 75$. The value of $R^{1/2}(\lambda, k)$ at the minimum is about $6m$. The smoothing procedure has resulted in a smoothed minus true standard deviation which is about $1/3$ less than the observational standard deviation.



$RanU^{1/2}$, the randomized version of Unbiased Risk, as a function of $log_{10}(\lambda)$ and k.

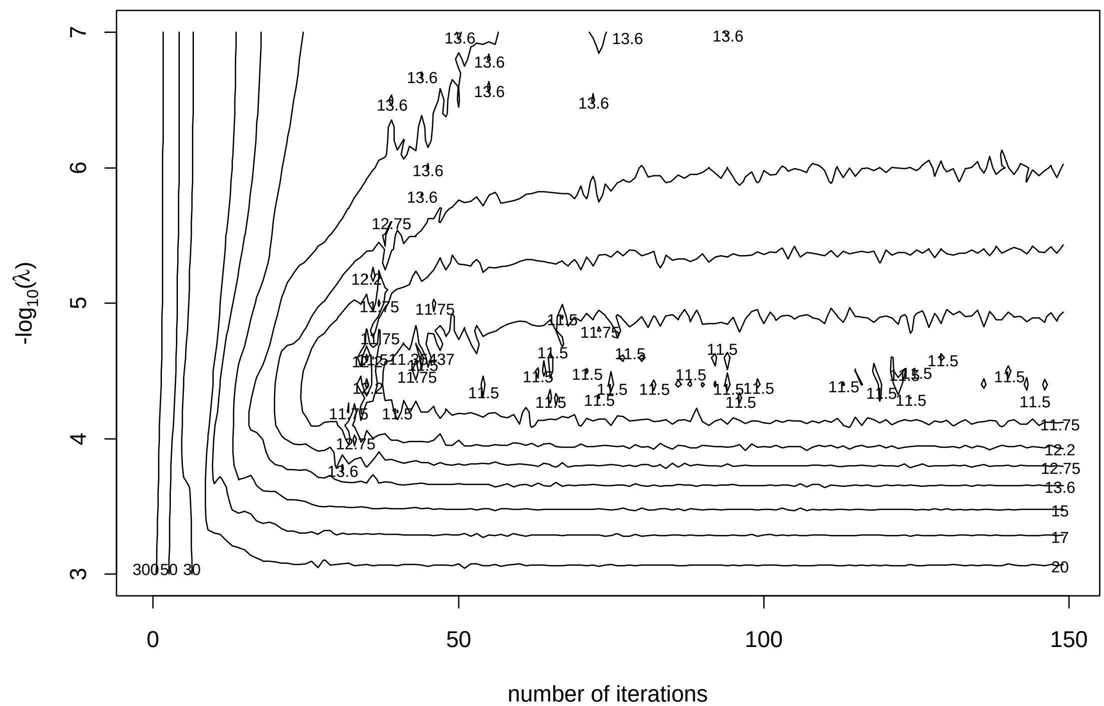
$$RanU(\lambda, k) = \frac{1}{n}\|y - K\hat{x}_\lambda^k\|^2$$

$$+\frac{2\sigma^2}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[K\hat{x}_\lambda^k(y+\xi) - K\hat{x}_\lambda^k(y)]\right\},$$

where ξ came from a random number generator, $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I)$ and the true $\sigma^2 = 9m.$ was used.

[Recall $U(\lambda, k) = \frac{1}{n} RSS(\lambda, k) + \frac{2\sigma^2}{n} trace A(\lambda, k)$.]

The first term in $RanU$ is the mean residual sum of squares and the expression in large brackets is the randomized trace estimate of the influence *operator*. The standard deviation σ_ξ for the random vector ξ should be chosen carefully if the implied influence matrix $A_y^k(\theta)y$ is not linear in y (as it won't be if the conjugate gradient algorithm is used). If σ_ξ is too small, then the calculation of the difference may be unstable, if σ_ξ is too large the behavior at $A_y^k(\theta)$ may not be captured. Trial and error gave a σ_ξ somewhat smaller than the presumed σ of the noise in y, $\sigma_\xi = 3m = \frac{1}{3}\sigma$. $RanU^{1/2}(\lambda, k)$, estimates $R^{1/2}(\lambda, k)$ well. The smallest value of $R^{1/2}(\lambda, k)$ is 5.987. The minimum of $RanU^{1/2}(\lambda, k)$ is located in a region for which the value of $R^{1/2}$ is less than or equal to 6.12 in the R plot, so that if a value of λ and a stopping rule k based on minimizing $RanU$ were used, then the ratio of the resulting predictive mean square error to the minimum possible predictive mean square error (the inefficiency), would be no larger than $6.12/5.987 = 1.022$.



$RanV^{1/2}$ as function of $\log_{10}(\lambda)$ and k.

The randomized GCV function is computed as

$$RanV(\lambda, k) = \frac{\frac{1}{n}\|y - K\hat{x}_\lambda^k\|^2}{\left(\frac{1}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[\xi - (K\hat{x}_\lambda^k(y+\xi) - K\hat{x}_\lambda^k(y))]\right\}\right)^2}$$

[Recall that $V(\lambda, k) = \frac{\frac{1}{n}RSS(\lambda,k)}{\frac{1}{n}(trace(I - A(\lambda,k))^2}$.]

The value of $RanV^{1/2}$ at the minimum (11.354) is roughly an estimate of $\sqrt{\min_{\lambda,k} R(\lambda, k) + \sigma^2} = 10.8$, as predicted by the theory. The minimum GCV score is located in a region for which the PMSE score $R^{1/2}$ is less than or equal to 6.25 so that the inefficiency is no bigger than $6.25/5.987 = 1.044$.

In this experiment, the optimum λ was insensitive to k, but in other experiments with larger noise, it was.