Statistics 860 Lecture 13

BAYESIAN CONFIDENCE INTERVALS

First : Usual(Parametric) Confidence Intervals

$$y = X\beta + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

$$\begin{aligned}\widehat{\beta} = (X'X)^{-1}X'y \;&= (X'X)^{-1}X'(X\beta + \epsilon)\\ &= \beta + (X'X)^{-1}X'\epsilon\end{aligned}$$

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X'X)^{-1})$$

$$\widehat{y} = X\widehat{\beta} = X(X'X)^{-1}X'y = Ay, \text{ say.}$$

$$\begin{aligned}\widehat{y} \;&= (X(X'X)^{-1}X'(X\beta + \epsilon))\\ &= (X\beta) + A\epsilon\\ \widehat{y} \;&\sim \mathcal{N}(X\beta, \sigma^2 A^2)\end{aligned}$$

Cov $(A\epsilon) = \sigma^2 A^2 = \sigma^2 A$, since A is idempotent in this case. Thus, a confidence interval for $(X\beta)_i$, the ith component of $X\beta$, would be

$$\widehat{y}_i \pm z_{.025}\sigma\sqrt{a_{ii}}.$$

Can use $\widehat{\sigma}^2 = \text{RSS}/(n-p) = \text{RSS}/\text{trace}(I - A)$, RSS$=$ residual sum of squares.

Bayesian Confidence Intervals

$$y_i = f(t_i) + \epsilon_i \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

$$f(t_i) - f_\lambda(t_i) = \underbrace{(f(t_i) - Ef_\lambda(t_i))}_{b_i\text{(bias)}} - \underbrace{(f_\lambda(t_i) - Ef_\lambda(t_i))}_{\delta_i\text{(variance)}}$$

$$\begin{pmatrix} f_\lambda(t_1) \\ \vdots \\ f_\lambda(t_n) \end{pmatrix} = A(\lambda)y = A(\lambda)f + A(\lambda)\epsilon$$

bias$= (I - A(\lambda))f \quad \{b_{i\lambda}\}$

variance$= A(\lambda)\epsilon \qquad \{\delta_{i\lambda}\}$

Smoothing spline estimates are biased.

Mean square error

$$\frac{1}{n}\sum_{i=1}^{n}(b_{i\lambda}-\delta_{i\lambda})^2$$

Important remark : $\sum b_{i\lambda} \equiv 0$ if $1 \in \mathcal{H}$ and the null space of the penalty functional since

$$\sum_{i=1}^{n} b_{i\lambda} = f'(I-A(\lambda))\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

$$\text{and } (I-A(\lambda))\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}$$

$Pick$ a random element b (equally likely) from the population $\{b_{i\lambda}\}$ and a random element δ from the population $\{\delta_{i\lambda}\}$ and observe $Eb=0$, $E\delta \approx 0$.

$$Eb^2 = \frac{1}{n}\sum_{i=1}^{n} b_i^2, \quad E\delta^2 = \frac{\sigma^2}{n}\text{tr}A^2(\lambda)$$

and $b+\delta \approx \mathcal{N}(0, \frac{1}{n}ER(\lambda))$ since

$$E_{pick}R(\lambda) = E\frac{1}{n}\sum(f(t_i)-f_\lambda(t_i))^2 = \frac{1}{n}\sum(b_{i\lambda}^2+\delta_{i\lambda}^2).$$

The Bayesian confidence interval is

$$f_\lambda(t_i) \pm z_{\alpha/2}\widehat{\sigma}\sqrt{a_{ii}(\widehat{\lambda})}(*).$$

This confidence interval came from looking at the Bayesian model behind the smoothing spline and observing that the posterior covariance of f_λ is $\sigma^2 A(\lambda)$, when the prior on f is $b\Sigma$ and $\lambda = \sigma^2/nb$. (`wahba.ci.83.pdf`). We take

$$\widehat{\sigma}^2 = \text{RSS}(\widehat{\lambda})/\text{trace}(I - A(\widehat{\lambda})).$$

For this Bayesian confidence interval to work in a frequentist setting, we need $\widehat{\lambda}$ is close to λ^*, the minimizer of $R(\lambda)$: IMPORTANT. The argument (see `wahba.ci.83.pdf, nychka.ci.88.pdf`) is that

$$R(\lambda^*) \approx \alpha\frac{\sigma^2}{n}\sum_{i=1}^{n} a_{ii}(\lambda^*)$$

where α is near 1. This argument needs $R(\widehat{\lambda})$ close to $R(\lambda^*)$ to work in practice, in that case the mean square bias plus variance is near σ^2 times the average $a_{ii}(\widehat{\lambda})$.

For the smoothing spline

$$\alpha \in \Big[\Big(1 + \frac{1}{4m}\Big)\Big(1 - \frac{1}{2m}\Big), 1\Big]$$

For $m = 2$, $\alpha \in \left[\frac{27}{32}, 1\right]$.
The bottom line: Under the $pick$ distribution

$$\begin{aligned} b_{\widehat{\lambda}} + \delta_{\widehat{\lambda}} & \approx \mathcal{N}\Big(0, \frac{1}{n} ER(\widehat{\lambda})\Big) \\ & \approx \mathcal{N}\Big(0, \frac{\sigma^2}{n} \sum_{i=1}^{n} a_{ii}(\widehat{\lambda})\Big) \end{aligned}$$

when $\widehat{\lambda} = \lambda^*$. The Bayesian confidence intervals work “on the average”, NOT pointwise. This means, for example: Suppose $n = 100$. Suppose you pick a number i^* equally likely from $\{1, 2, \cdots, 100\}$. Then the probability that $f_\lambda(t_{i^*}) \pm z_{\alpha/2} \widehat{\sigma} \sqrt{a_{ii}(\widehat{\lambda})}$ covers $f(t_i)$ is about 95%. (are using Z rather than t because $n >> 30$).

To compare the parametric case with the nonparametric case: in the $pick$ context, in the parametric case, since

$$b \equiv 0, \delta = A\epsilon$$

$$E\frac{1}{n}\sum(b_i^2 + \delta_i^2) = \frac{\sigma^2}{n}trA^2 \equiv \frac{\sigma^2}{n}trA$$

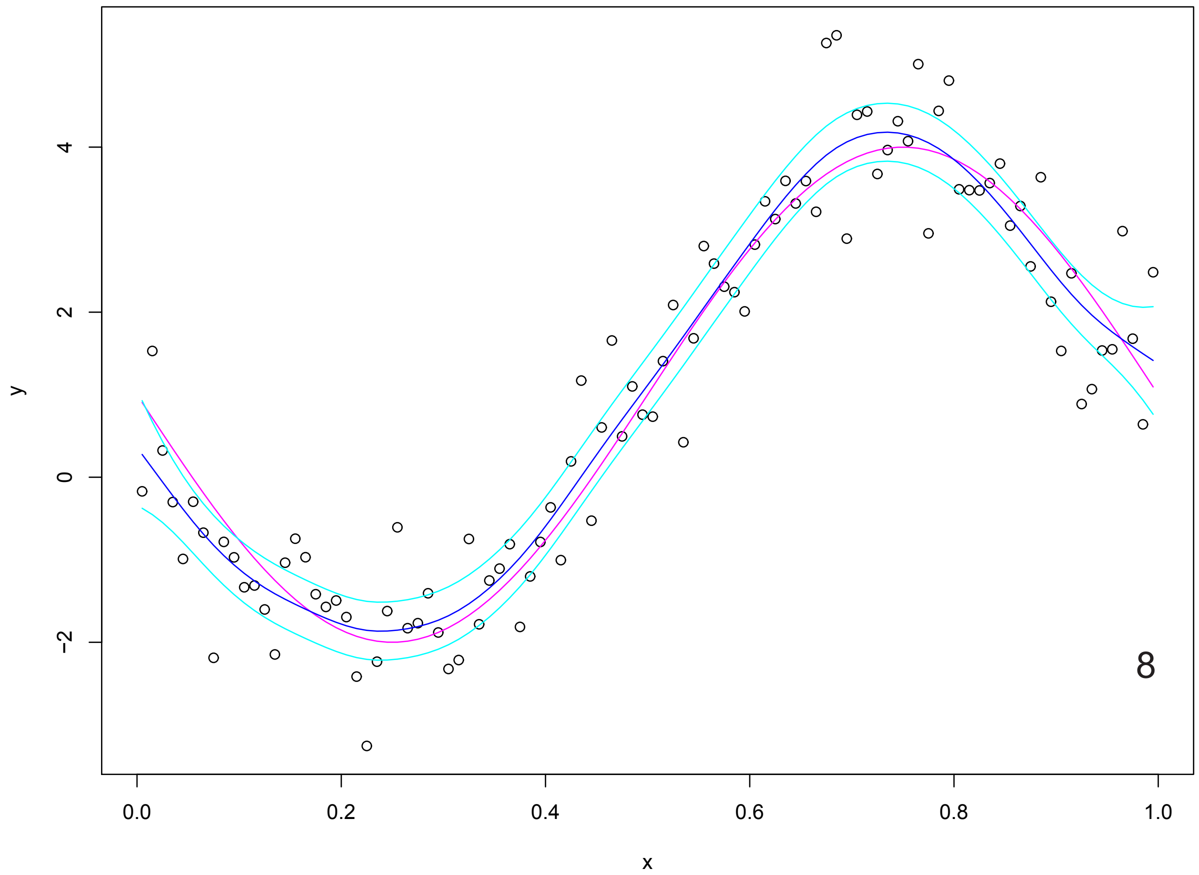
while in the nonparametric case

$$E\frac{1}{n}\sum(b_{i\hat{\lambda}}^2 + \delta_{i\hat{\lambda}}^2) \approx \frac{\sigma^2}{n}trA(\hat{\lambda})$$

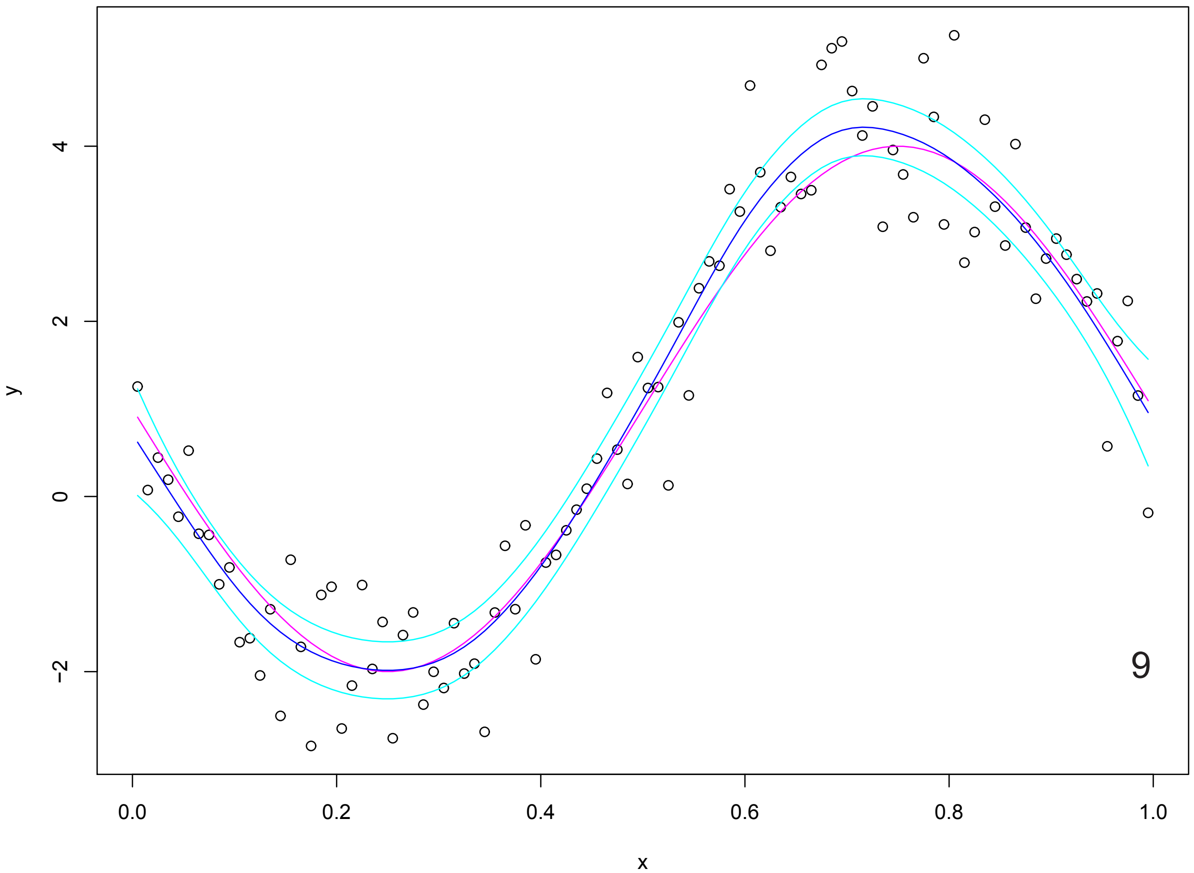
From R/conf.int

```
   nla                   sigrat          count
-3.44771968981658  0.92654288801936        91
-3.20414150430556  0.91939533790345        87
-3.09530446540393  0.86847699330725        91
-3.52345597672399  1.02740449188683        100
-3.0786833109878   1.02004228771482        100
```

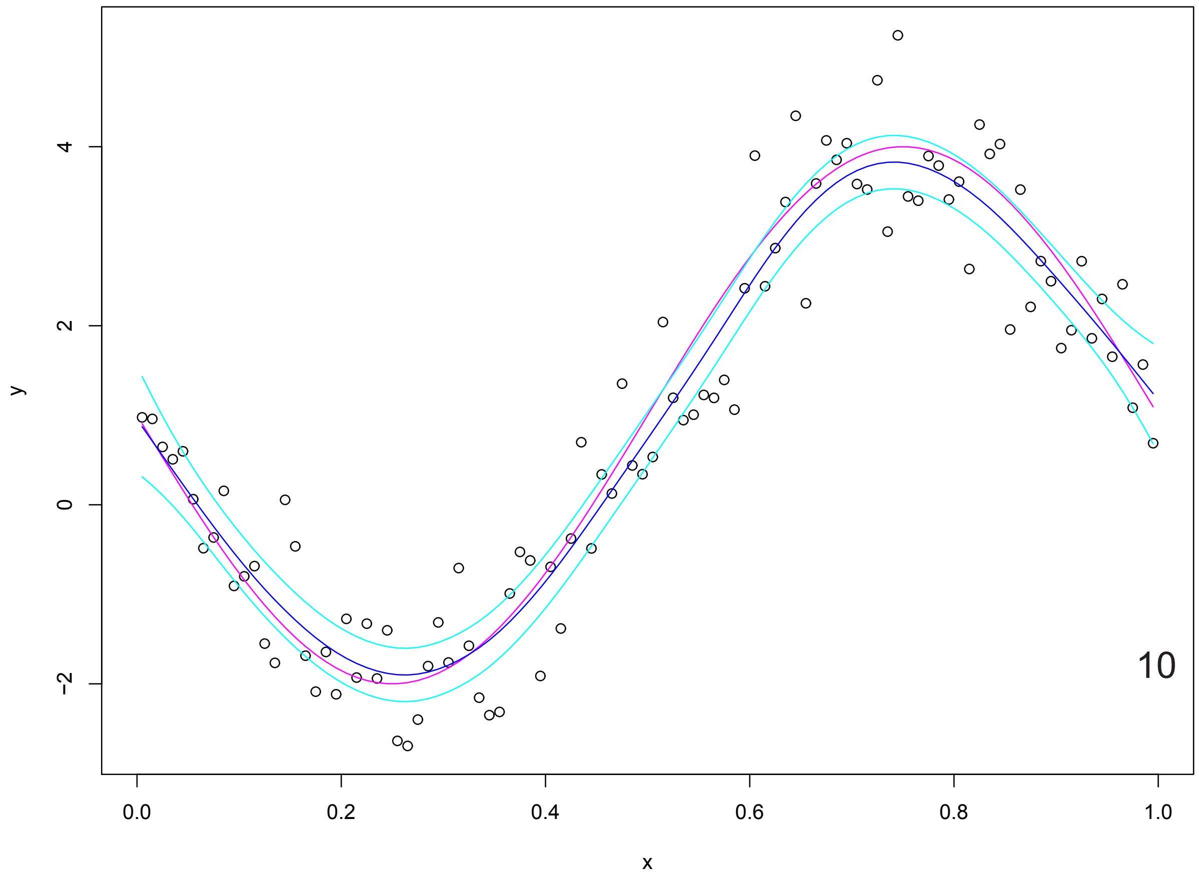
Count = 91



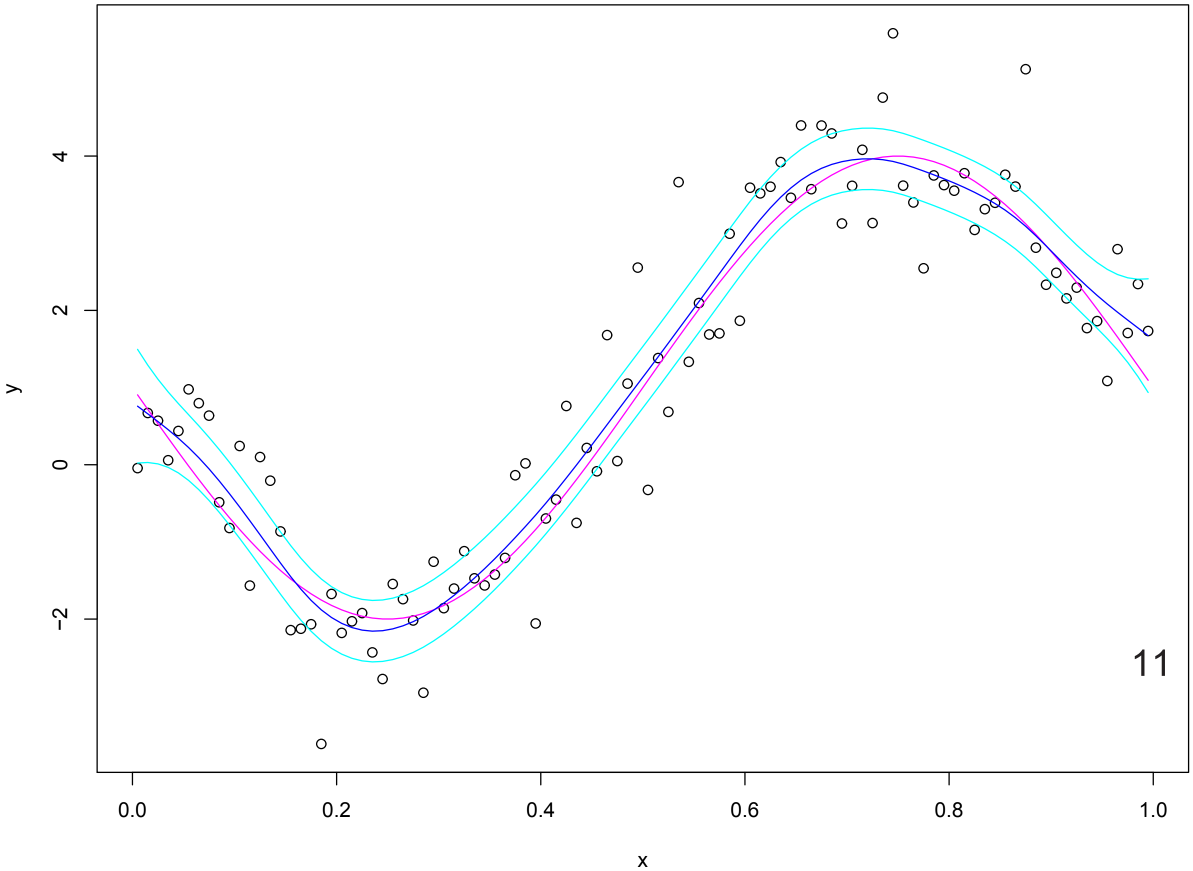
Count = 87



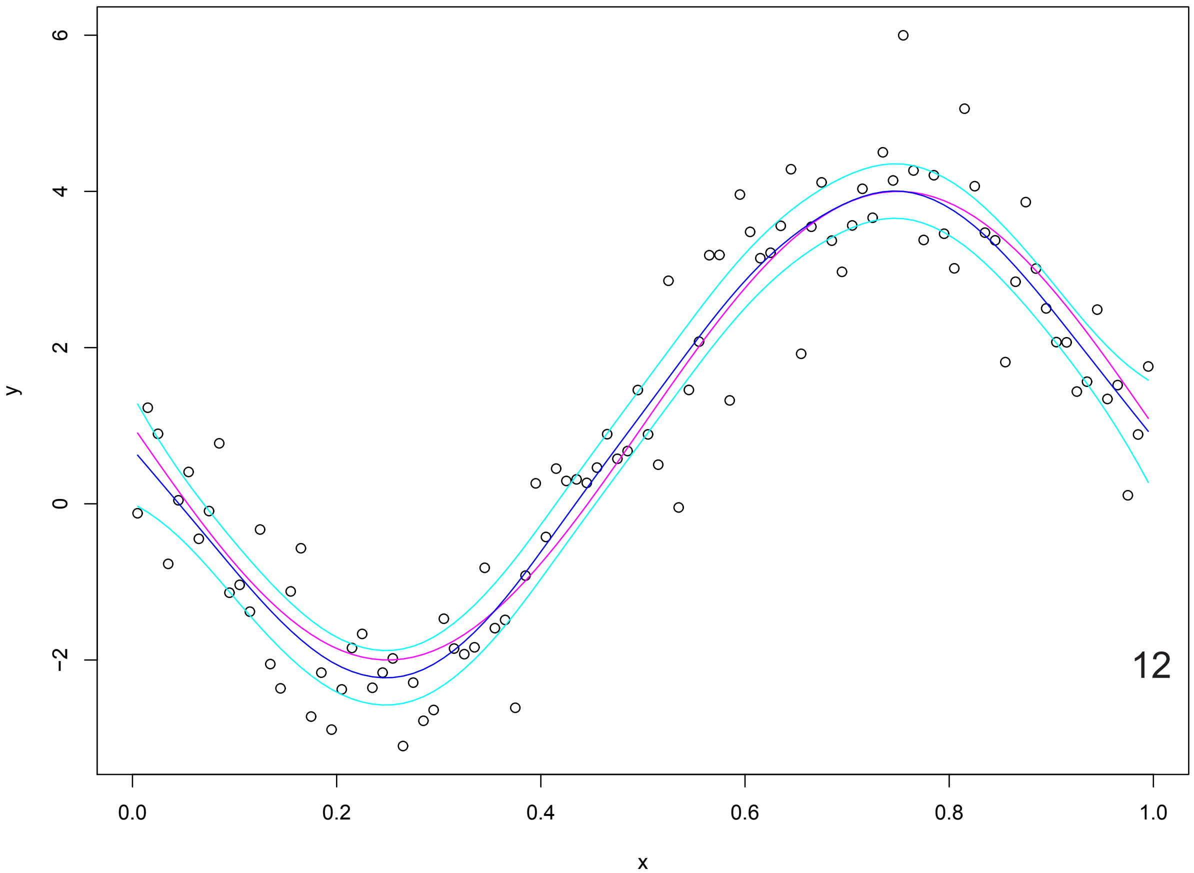
Count = 91



Count = 100



Count = 100



G. Wahba. Bayesian “confidence intervals” for the cross-validated smoothing spline. *J. Roy. Stat. Soc. Ser. B*, 45:133–150, 1983.
`wahba.ci.83.pdf`

X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, 28:1570–1600, 2000
`lin.wahba.xiang.gao.klein.klein.2000.pdf`

illustrate the ‘across the function’ property of property of the Bayesian confidence intervals in a simulation study and also illustrate how they work on a real data set.

Firm(er) theoretical foundation:

D. Nychka, Bayesian Confidence Intervals for Smoothing Splines, J. Amer. Statist. Assoc, 83, 1988, pp 1134-1143. `nychka.ci.88.pdf`

Calculation Tricks

1. Select a subset of the basis functions(representers). `lin.wahba.xiang.gao.klein.klein.2000.pdf` Sect 5.1.

2. Randomized trace method for GCV(Girard, *Ann. Statist.* 1991, Hutchinson, Commun. Statist.) and GACV(X. Lin, Xiang, theses), `lin.wahba.xiang.gao.klein.klein.2000.pdf`.

Let $(\epsilon_1, \cdots, \epsilon_n) \sim \mathcal{N}(0, \sigma_e^2 I)$. Let $A = A(\lambda)$ be a smoother matrix. Then

$$\begin{aligned} E\epsilon' A\epsilon &= E\sum_{i,j=1}^{n} \epsilon_i\epsilon_j a_{ij} \\ &= \sigma_e^2 \sum_{i=1}^{n} a_{ii} = \sigma_e^2 \text{trace} A(\lambda) \end{aligned}$$

Let $t = \frac{1}{n}\epsilon' A\epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I)$. Then the standard deviation of t is $\frac{1}{\sqrt{2n}}\left[\frac{1}{n}\text{trace}A^2\right]^{1/2} \leq \frac{1}{\sqrt{2n}}\left[\frac{1}{n}\text{trace}A\right]^{1/2}$.

Suppose $\epsilon_i = \pm\theta$, what is the variance?

$$\begin{pmatrix} f_\lambda^y(t_1) \\ \vdots \\ f_\lambda^y(t_n) \end{pmatrix} = A(\lambda)y$$

$$\begin{pmatrix} f_\lambda^{y+\epsilon}(t_1) \\ \vdots \\ f_\lambda^{y+\epsilon}(t_n) \end{pmatrix} = A(\lambda)(y+\epsilon)$$

$$\epsilon' \left[\begin{pmatrix} f_\lambda^{y+\epsilon}(t_1) \\ \vdots \\ f_\lambda^{y+\epsilon}(t_n) \end{pmatrix} - \begin{pmatrix} f_\lambda^y(t_1) \\ \vdots \\ f_\lambda^y(t_n) \end{pmatrix} \right] = \epsilon' A(\lambda)\epsilon.$$

Keep the same ϵ as λ varies.

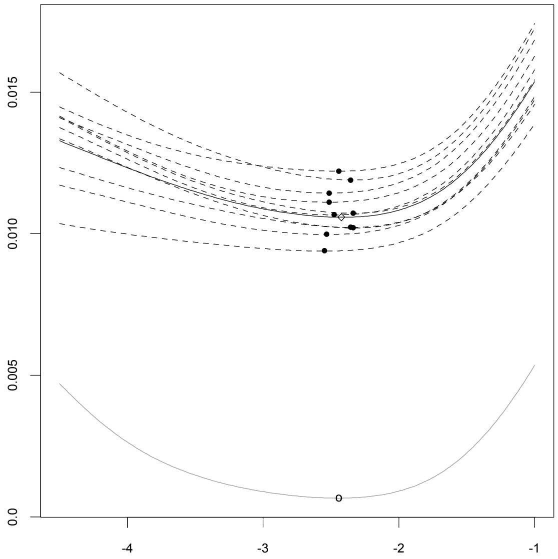
$$\begin{aligned} \text{GCV}(\lambda) &= \frac{\text{RSS}(\lambda)}{(1 - \frac{1}{n}\text{Trace}A(\lambda))^2} \\ \text{UBR}(\lambda) &= \text{RSS}(\lambda) + 2\sigma^2\text{Trace}A(\lambda) \end{aligned}$$

See plots from

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.mwr1995.pdf`

which illustrates what 10 $ranGCV(\lambda)$ curves look like, compared to $GCV(\lambda)$ and $R(\lambda)$

Dashed lines ranGCV, upper solid line exact GCV, bottom solid line mean square error as a function of $\log \lambda$



Also plots from

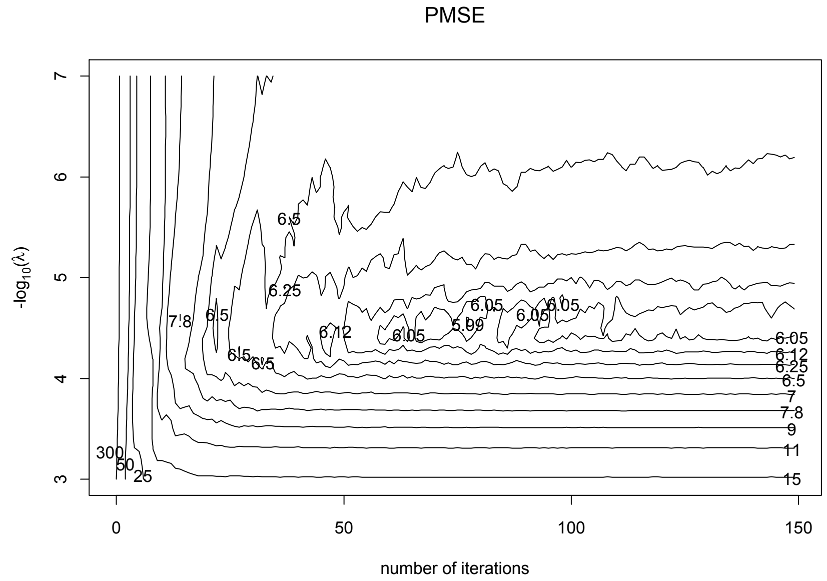
X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, 28:1570–1600, 2000 `lin.wahba.xiang.gao.2000` illustrate what the $ranGACV$ curves look like.

Imputation from nearly regular data sets to regular data sets in the global warming dataset allows the trick of describing the design as a tensor(cartesian!) of two univariate designs, space and time. Then backfitting is used to solve the ensuing simpler equations. See

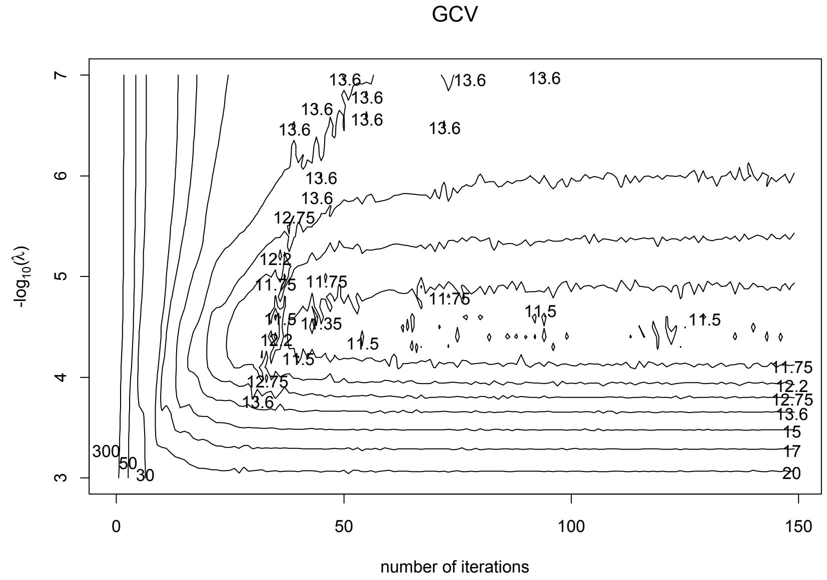
Wahba, G. and Luo, Z. " Smoothing Spline ANOVA Fits for Very Large, Nearly Regular Data Sets, with Application to Historical Global Climate Data" TR 952, October 1995. Slightly revised version in Annals of Numerical Mathematics 4 (1997) 579-598, Festschrift in Honor of Ted Rivlin, C.Micchelli, Ed. `lreg.rev.pdf`.

Early stopping of interative methods for solving large linear systems has a regularization effect. See `wahba.johnson.gao.gong.mwr1995.pdf`. and Wahba, G. ” Three topics in Ill-posed Inverse Problems. ” In ”Inverse and Ill-Posed Problems, M. Engl and G. Groetsch, Eds., Academic Press 1987, pp 37-50. Among other things provides a discussion of how early stopping of iterative methods for solving large linear systems is a form of regularization. `illpose.pdf`.

PMSE as a function of $\log \lambda$ and number of iterations.



GCV as a function of $\log \lambda$ and number of iterations.


GCV
-log10(λ)
number of iterations