Name: Jared Yu Date: May 31, 2018

**1(c)** Suppose for the moment that $\sigma^2$ is known. I want to get an unbiased estimate for the $L_2$-risk:

$$\text{risk}(\lambda) = E||\boldsymbol{f} - \hat{\boldsymbol{f}}_\lambda||^2.$$

Let consider another expression for the risk first.

$$\begin{aligned}
\text{risk}(\lambda) &= E||\boldsymbol{f} - \hat{\boldsymbol{f}}_\lambda||^2 \\
&= E[(\boldsymbol{f} - \hat{\boldsymbol{f}}_\lambda)^T(\boldsymbol{f} - \hat{\boldsymbol{f}}_\lambda)] \\
&= E[(\boldsymbol{f}^T - \boldsymbol{y}^T\boldsymbol{H}_\lambda^T)(\boldsymbol{f} - \boldsymbol{H}_\lambda\boldsymbol{y})] \\
&= E(\boldsymbol{f}^T\boldsymbol{f} - \boldsymbol{f}^T\boldsymbol{H}_\lambda\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{H}_\lambda^T\boldsymbol{f} + \boldsymbol{y}^T\boldsymbol{H}_\lambda^T\boldsymbol{H}_\lambda\boldsymbol{y}) \\
&= \boldsymbol{f}^T\boldsymbol{f} - \boldsymbol{f}^T\boldsymbol{H}_\lambda\boldsymbol{f} - \boldsymbol{f}^T\boldsymbol{H}_\lambda^T\boldsymbol{f} + \text{tr}\left[E(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{H}_\lambda^T\boldsymbol{H}_\lambda\right] \\
&= \boldsymbol{f}^T\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{H}_\lambda\boldsymbol{f} + \text{tr}\left[(\boldsymbol{f}\boldsymbol{f}^T + \sigma^2\boldsymbol{I})\boldsymbol{H}_\lambda^T\boldsymbol{H}_\lambda\right] \\
&= \boldsymbol{f}^T\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{H}_\lambda\boldsymbol{f} + \boldsymbol{f}^T\boldsymbol{H}_\lambda^T\boldsymbol{H}_\lambda\boldsymbol{f} + \sigma^2\text{tr}(\boldsymbol{H}_\lambda^T\boldsymbol{H}_\lambda) \\
&= ||\boldsymbol{f} - \boldsymbol{H}_\lambda\boldsymbol{f}||^2 + \sigma^2\text{tr}(\boldsymbol{H}_\lambda^T\boldsymbol{H}_\lambda).
\end{aligned}$$

Then, I can start to look for an unbiased estimator for the risk using the following expectation.

$$\begin{aligned}
E||\boldsymbol{y} - \hat{\boldsymbol{f}}_\lambda||^2 &= E||\boldsymbol{y} - \boldsymbol{H}_\lambda\boldsymbol{y}||^2 \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + E\left[||\boldsymbol{y} - \boldsymbol{H}_\lambda\boldsymbol{y}||^2 - ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2\right] \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + E\left[\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{y} - \boldsymbol{f}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}\right] \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + E\left\{\text{tr}\left[\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{y}\right] - \text{tr}\left[\boldsymbol{f}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}\right]\right\} \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + E\left\{\text{tr}\left[\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\right] - \text{tr}\left[\boldsymbol{f}\boldsymbol{f}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\right]\right\} \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + \text{tr}\left\{E(\boldsymbol{y}\boldsymbol{y}^T - \boldsymbol{f}\boldsymbol{f}^T)\left[(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\right]\right\} \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + \text{tr}\left\{\sigma^2\left[(\boldsymbol{I} - \boldsymbol{H}_\lambda)^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)\right]\right\} \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}||^2 + \sigma^2\left\{\text{tr}\left(\boldsymbol{I} - \boldsymbol{H}_\lambda^T - \boldsymbol{H}_\lambda + \boldsymbol{H}_\lambda\boldsymbol{H}_\lambda^T\right)\right\} \\
&= ||\boldsymbol{f} - \boldsymbol{H}_\lambda\boldsymbol{f}||^2 + \sigma^2\left\{\text{tr}(\boldsymbol{H}_\lambda\boldsymbol{H}_\lambda^T) - 2\text{tr}(\boldsymbol{H}_\lambda) + n\right\}.
\end{aligned}$$

Use this expression, I can construct an unbiased estimator for risk($\lambda$) (when $\sigma^2$ is known) in the following

$$\widehat{\text{risk}}(\lambda) = ||\boldsymbol{y} - \hat{\boldsymbol{f}}_\lambda||^2 + \sigma^2\left\{2\text{tr}(\boldsymbol{H}_\lambda) - n\right\}.$$

To make our algorithm more efficient, I did some tricks to speed up the computation. When searching for the optimum $\lambda$, I have to fit the model repeatedly for different $\lambda$ and compute the fitted values through

$$\hat{f}_\lambda = X(X^TX + \lambda D)^{-1}X^Ty$$

where X is design matrix$x = \begin{pmatrix} 1 & x_1 & x_1^2 & ... & x_1^p & (x_1-t_1)_+^p & ... & (x_1-t_k)_+^p \\ \vdots & \vdots & \vdots & ... & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & ... & x_n^p & (x_n-t_1)_+^p & ... & (x_n-t_k)_+^p \end{pmatrix}$

$t_1, t_2, \ldots t_k$ are knots

D is a penalty matrix = diag$\{0, \ldots, 0, 1, \ldots, 1\}$

the first p+1 are zero and the last k elements are 1.

So I need to compute the inverse matrix for each $\lambda$ which slows our algorithm down. To keep us from this time consuming step, the cholesky decomposition and eigen decomposition are used. Let the cholesky decomposition of $X^TX$ be $C^TC$, where C is a upper triangular square matrix. Since I use the equally spread knots[1], X is full rank matrix and hence C is invertible. Also, let $C^{-T} = (C^{-1})^T$, then

$$\begin{aligned} X(X^TX + \lambda D)^{-1}X^T &= X(C^TC + \lambda D)^{-1}X^T \\ &= XC^{-1}(I + \lambda C^{-T}DC^{-1})^{-1}C^{-T}X^T \end{aligned}$$

Further, I use the eigen decomposition of $C^{-T}DC^{-1}$, which is $C^{-T}DC^{-1} = PQP^T$, where Q is a diagonal matrix and P is an orthogonal matrix. Then,

$$\begin{aligned} X(X^TX + \lambda D)^{-1}X^T &= XC^{-1}(I + \lambda PQP^T)^{-1}C^{-T}X^T \\ &= XC^{-1}P(I + \lambda Q)^{-1}P^TC^{-T}X^T \end{aligned}$$

Let $G = XC^{-1}P$ and thus $X(X^TX + \lambda D)^{-1}X^T = G(I + \lambda Q)^{-1}G^T$, where $(I + \lambda Q)$ is a diagonal matrix and inverting it just the same as inverting the diagonal elements, this will keep us from the complicating matrix inversion step. And since G is the same for all $\lambda$, I only need to calculate once. Hence, the computation is much more easier than the original setting.

[1]In penalized regression splines, the knots do not necessary to equally spread. However, I place the knots equal-spaced within the domain of the data, ie the knots are placed at the (k/30)th quantiles of the data, for k=1,...,30

**(d)** The simulation study are based on five criteria to do the fitting. They are cross-validation (CV), generalized cross-validation (GCV), corrected AIC ($AIC_c$), unbiased risk estimator with unknown sigma (risk1) and unbiased risk estimator with known sigma (risk2). With the simulation setting given in Thomas C. M. Lee (2001), the simulation results are summarized in Figure 1, 2, 3 and 4. We can see that all criteria provide the similar result. The mean $\log_e r$ value is always less than 0.2. All fittings are acceptable even we have so many different simulation settings.

## Appendix: Tables and Figures

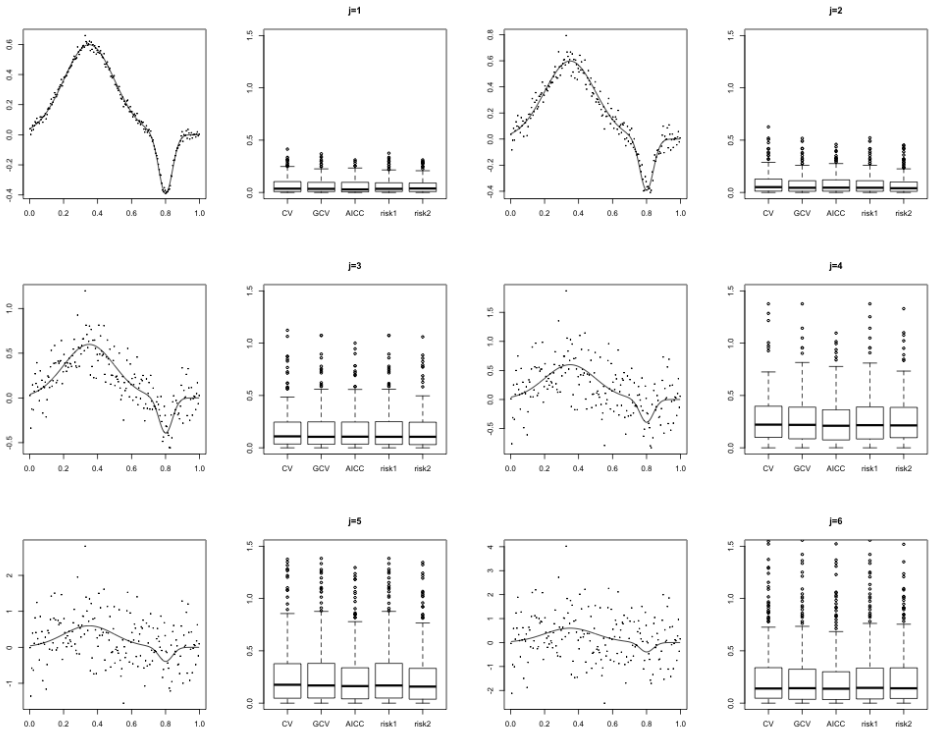


Figure 1: Results for changing the noise level factor. In each pair of panels the left-half displays one typical simulated data set together with the true regression function. The right-half are the boxplots of the $\log_e r$ values for, from left to right, CV, GCV, $AIC_c$, $risk_1$ and $risk_2$.

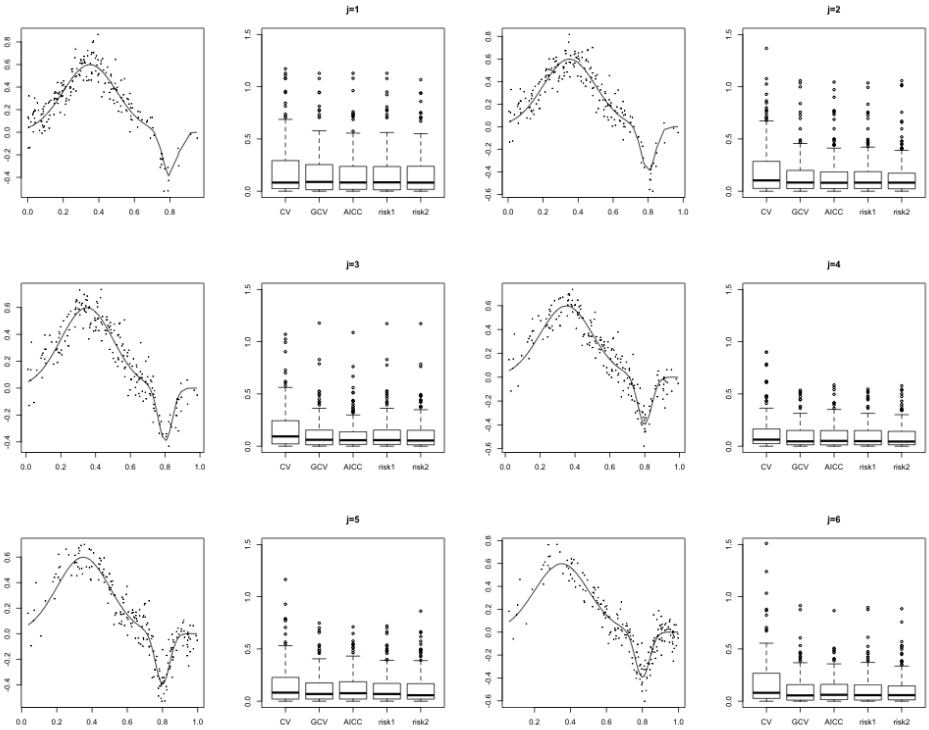


Figure 2: Results for changing the design density level factor. In each pair of panels the left-half displays one typical simulated data set together with the true regression function. The right-half are the boxplots of the $\log_e r$ values for, from left to right, CV, GCV, $AIC_c$, $risk_1$ and $risk_2$.

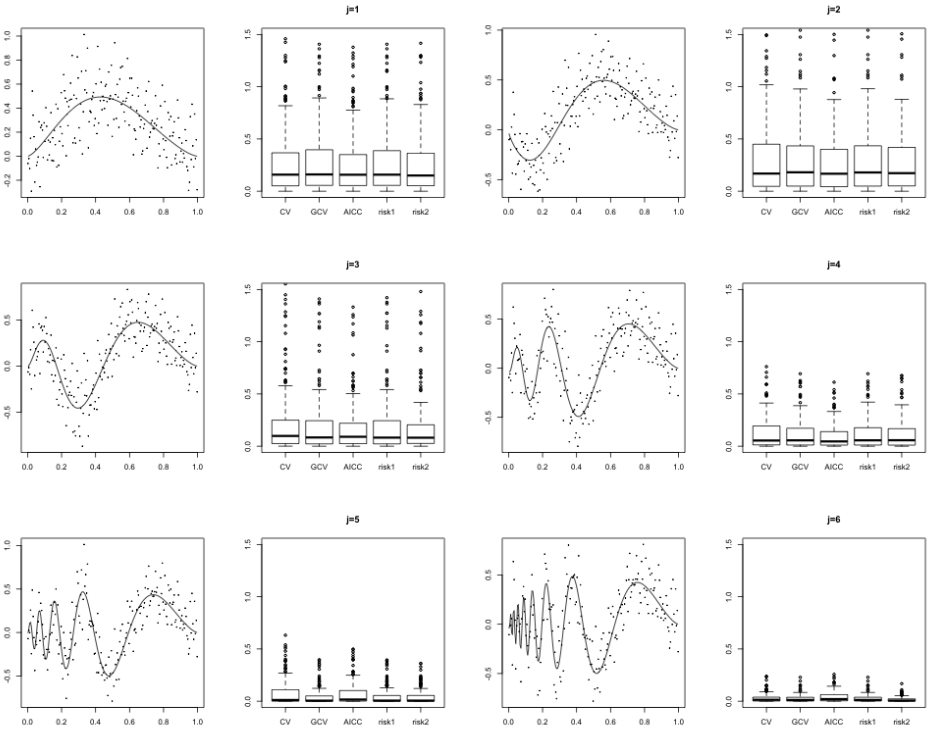


Figure 3: Results for changing the spatial variation level factor. In each pair of panels the left-half displays one typical simulated data set together with the true regression function. The right-half are the boxplots of the $\log_e r$ values for, from left to right, CV, GCV, $AIC_c$, $risk_1$ and $risk_2$.

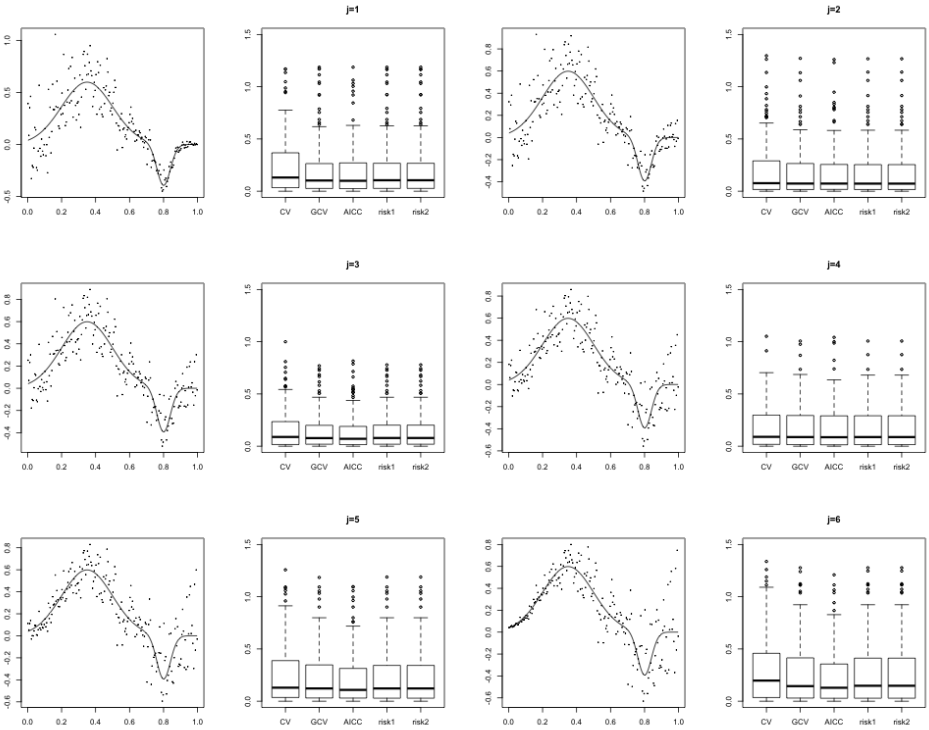


Figure 4: Results for changing the variation function level factor. In each pair of panels the left-half displays one typical simulated data set together with the true regression function. The right-half are the boxplots of the $\log_e r$ values for, from left to right, CV, GCV, $AIC_c$, $risk_1$ and $risk_2$.