STA 243 Assignment 5

Chen Zihao(50%) Man Pan(50%)

We fix the number of knots as 30, and place them equi-spaced within the domain of the data. Let $x_{(1)} = \min(x_i)$ and $x_{(n)} = \max(x_i)$ so that we have the location of the kth knot,$t_k = x_{(1)} + k\frac{x_{(n)} - x_{(1)}}{31}$.

Equi-spaced knots may cause a issue that there may be no data within two knots. In this homework it will not happen because $\{x_i\}$ is a arithmetic series. In this problem, it is equivalent to take all the unique x out and then choose the quantile points which will ensure all the subspace have at least data if the number of knots is appropriate.

the design matrix is as below.

$$X = \begin{bmatrix} 1 & X_1 & X_1^2 & X_1^3 & (X_1 - t_1)_+^3 & (X_1 - t_2)_+^3 & ... & (X_1 - t_{30})_+^3 \\ 1 & X_2 & X_2^2 & X_2^3 & (X_2 - t_1)_+^3 & (X_2 - t_2)_+^3 & ... & (X_2 - t_{30})_+^3 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 1 & X_n & X_n^2 & X_n^3 & (X_n - t_1)_+^3 & (X_n - t_2)_+^3 & ... & (X_n - t_{30})_+^3 \end{bmatrix}$$

we can get

$$\hat{f} = X\hat{\beta} = H_\lambda Y = X(X^\top X + \lambda D)^{-1} X^\top Y$$

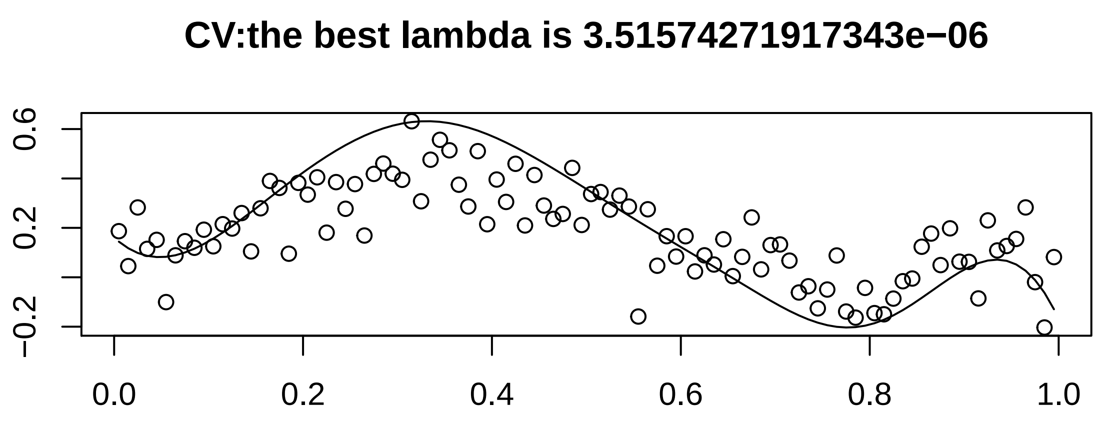
where $D = diag(0, 0, 0, 0, 1, 1, ..., 1)$ is a 34×34 matrix.

(a) Cross- validation (CV)

$$\sum_{i=1}^{n} (y_i - \hat{f}_{-i})^2 \approx \sum_{i=1}^{n} (\frac{y_i - \hat{f}_i}{1 - h_{ii}})^2$$

where $\{h_i i\}$ is the diagonal elements of H_λ

We use $y_i = 15\phi(\frac{x_i - 0.35}{0.15}) - 10\phi(\frac{x_i - 0.8}{0.04}) + \epsilon_i$ similar to the paper to test our algorithm

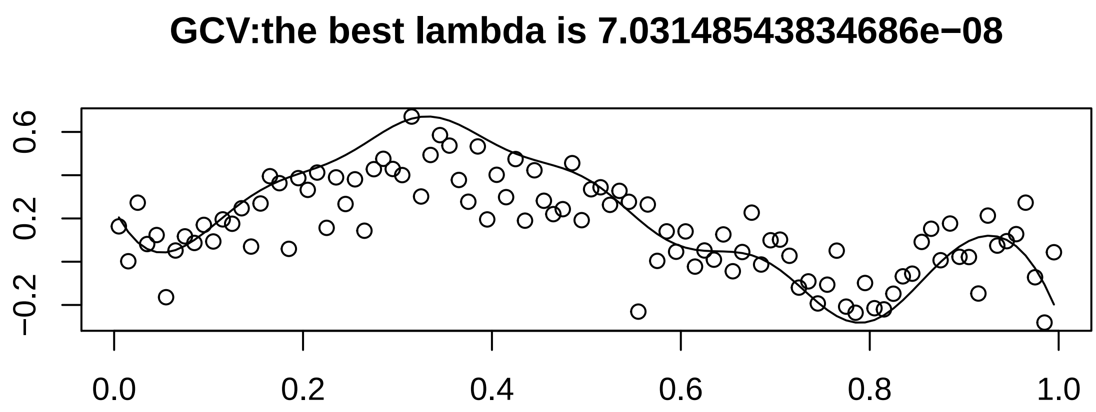


Generalized CV (GCV)

replace h_{ii} by the average of the diagonal elements of H_λ

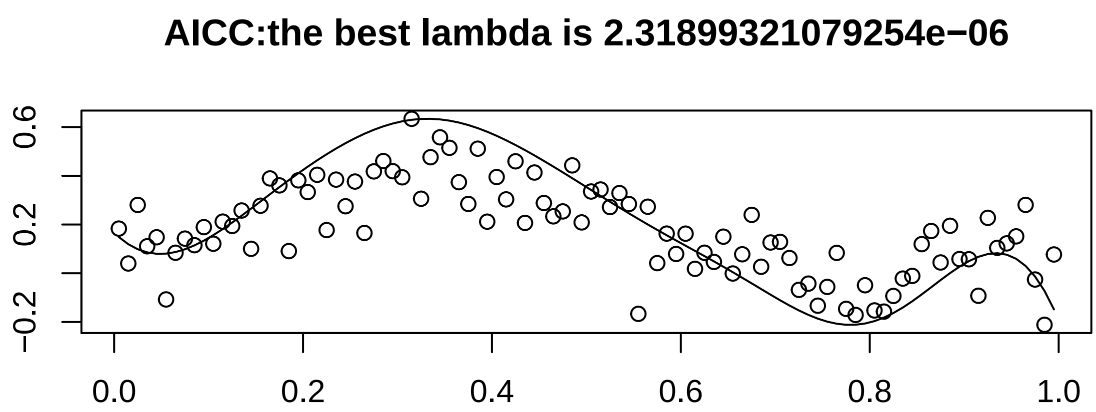
$$GCV = \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}_i}{1 - \frac{tr(H_\lambda)}{n}}\right)^2$$

We use the same data as previous one, to test our algorithm



(b) AICC

We use the same data to test our algorithm



(c) Estimate the expectation of the risk(λ)

$y_i = f(x_i) + \epsilon_i$

$$
\begin{aligned}
E||y - \hat{f}_\lambda||^2 &= E||f + e - H_\lambda(f + e)||^2 \\
&= E||(I - H_\lambda)(f + e)||^2 \\
&= E(((I - H_\lambda)(f + e))^\top((I - H_\lambda)(f + e))) \\
&= E((f + e)^\top(I - H_\lambda)^\top(I - H_\lambda)(f + e)) \\
&= ||(I - H_\lambda)f||^2 + E(f^\top(I - H_\lambda)^\top(I - H_\lambda)e) + E(e^\top(I - H_\lambda)^\top(I - H_\lambda)f) + E(e^\top(I - H_\lambda)^\top(I - H_\lambda)e) \\
&= ||(I - H_\lambda)f||^2 + E(e^\top(I - H_\lambda)^\top(I - H_\lambda)e) \\
&= ||(I - H_\lambda)f||^2 + E(e^\top(I - H_\lambda - H_\lambda^\top + H_\lambda^\top H_\lambda)e) \\
&= ||(I - H_\lambda)f||^2 + E(tr(e^\top(I - 2H_\lambda + H_\lambda^\top H_\lambda)e)) \\
&= ||(I - H_\lambda)f||^2 + E(tr((I - 2H_\lambda + H_\lambda^\top H_\lambda)ee^\top)) \\
&= ||(I - H_\lambda)f||^2 + tr(E((I - 2H_\lambda + H_\lambda^\top H_\lambda)ee^\top)) \\
&= ||(I - H_\lambda)f||^2 + \sigma^2(tr(n - 2tr(H_\lambda) + tr(H_\lambda^\top H_\lambda)) \\
&= ||(I - H_\lambda)f||^2 + \sigma^2\{tr(tr(H_\lambda H_\lambda^\top) - 2tr(H_\lambda) + n\}
\end{aligned}
$$

now consider $E||f - \hat{f}_\lambda||^2$

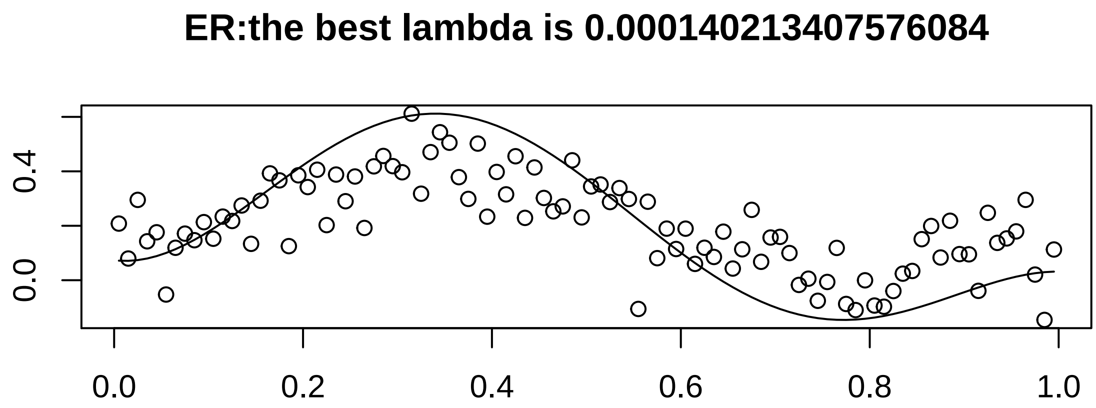
$$
\begin{aligned}
E||f - \hat{f}_\lambda||^2 &= E||f - y + y - H_\lambda y||^2 \\
&= E||f - y||^2 + 2E(f - y)^\top(y - H_\lambda y) + E||y - H_\lambda y||^2 \\
&= n\sigma^2 - 2E\epsilon^\top(I - H_\lambda)y + E||y - H_\lambda y||^2 \\
&= n\sigma^2 - 2E\epsilon^\top(I - H_\lambda)\epsilon + E||y - H_\lambda y||^2 \\
&= n\sigma^2 - 2tr(I - H_\lambda)\sigma^2 + ||y - H_\lambda y||^2 \\
&= E||y - \hat{f}_\lambda||^2 + (2tr(H_\lambda) - n)\sigma^2
\end{aligned}
$$

We use the folowing to estimate the $E||f - \hat{f}_\lambda||^2$

$$
||y - \hat{f}_\lambda||^2 + (2tr(H_\lambda) - n)\hat{\sigma}^2
$$

still need to estimate $\hat{\sigma}^2$,CV is a good estimate so we will use the result from CV.

We use the same data to test our algorithm



(d) using the simulation setting to test.

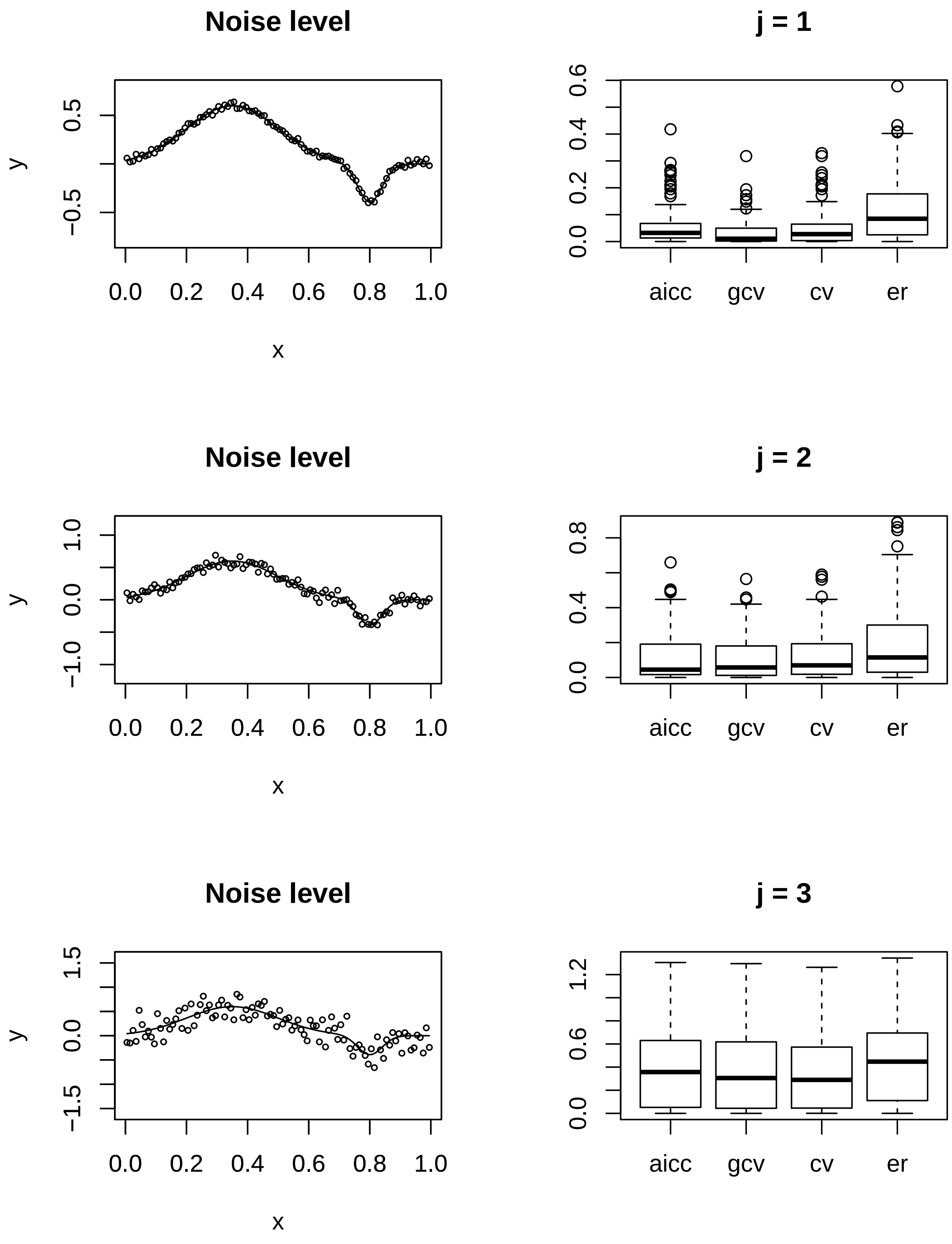
$\log r$,defined as below, is used to measure the performance of the 4 criteria. We simulate 100 times to plot boxplots to do the visulization.

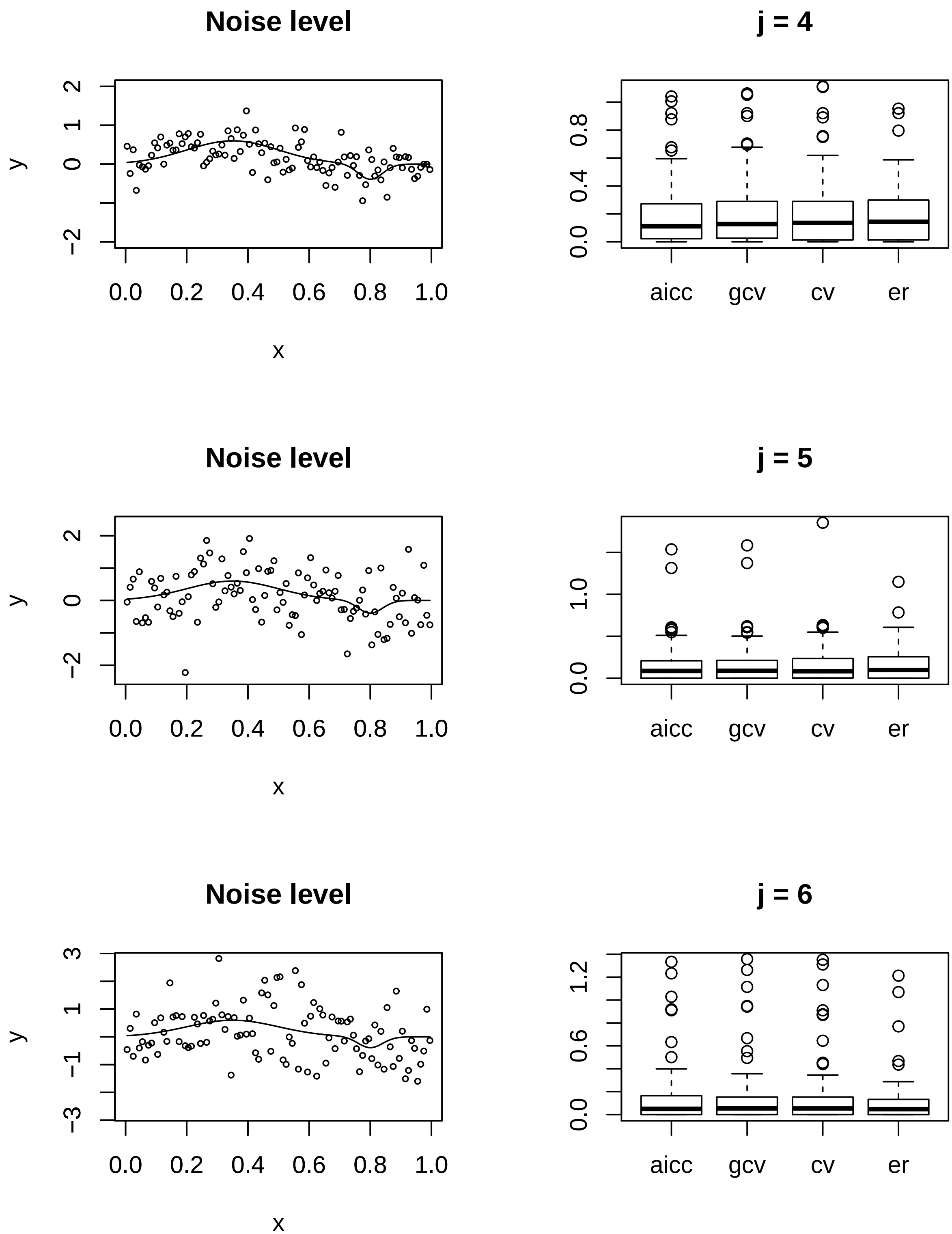
$$\log r = \log \frac{||f - \hat{f}_\lambda||^2}{\min_\lambda ||f - \hat{f}_\lambda||^2}$$

For the denominator, we first use binary search to find the minimizer. Then, we use the λ given by the 4 criteria to do a search in order to make sure our binary search did not traped at the local minimum. Although we can not make sure we get the global minimizer, the λ is still good enough to analyze. According to our algorithm, the λ in the denominator is more acurrate than the λ's given by the criteria.

Noise level

$$y_{ij} = f(x_i) + \sigma_j \epsilon_i \sigma_j = 0.02 + 0.04(j-1)^2$$

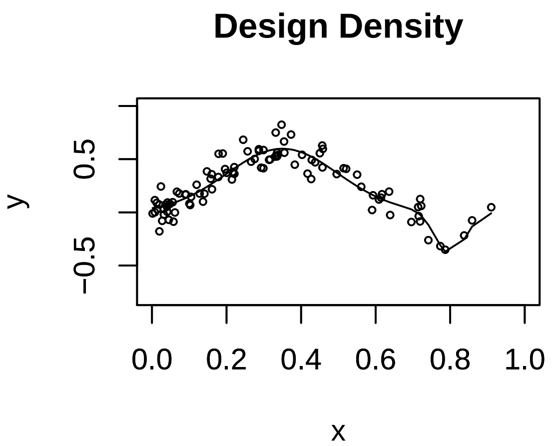

Noise level
j = 1
Noise level
j = 2
Noise level
j = 3
y
x
aicc
gcv
cv
er

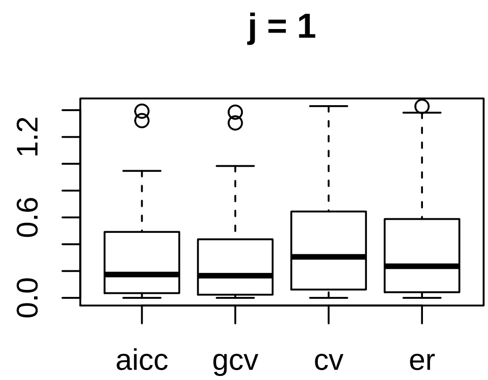

Noise level
j = 4
Noise level
j = 5
Noise level
j = 6
x
y
aicc
gcv
cv
er

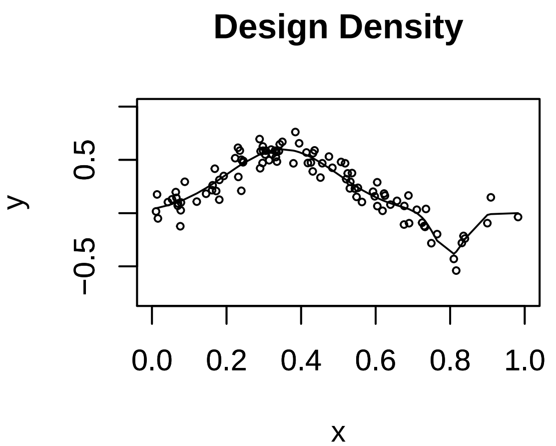
Design Density

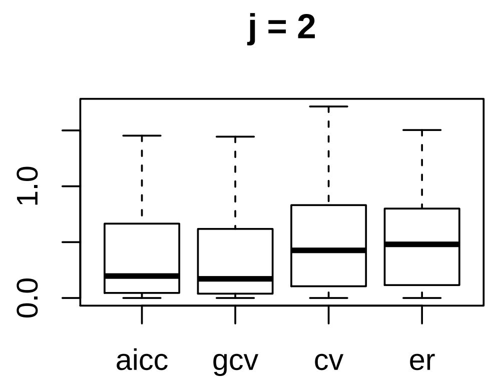
$$y_{ij} = f(X_{ji}) + \sigma\epsilon_i\sigma = 0.1, X_{ji} = F^{-1}(X_i)$$

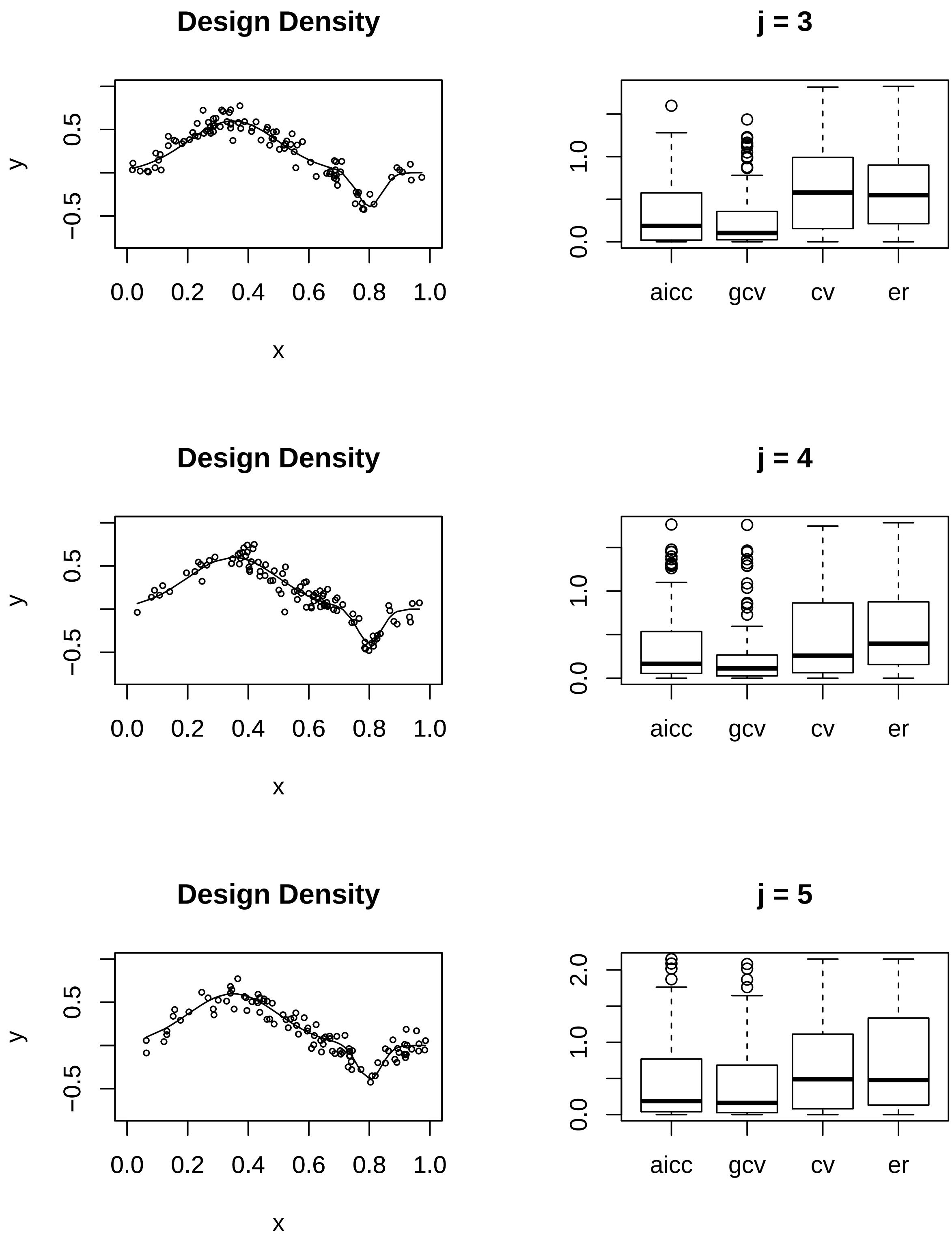
where F_j is the Beta($\frac{j+4}{5}, \frac{11-j}{5}$)

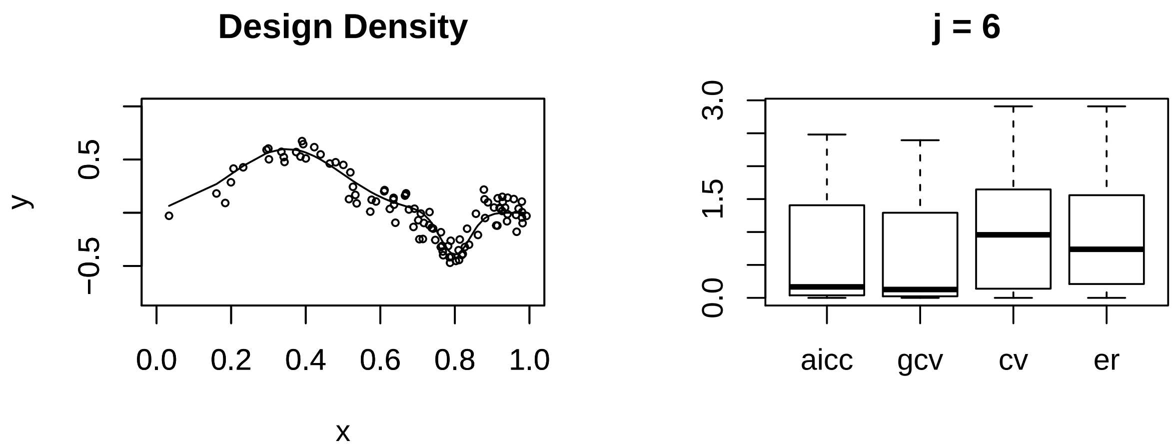




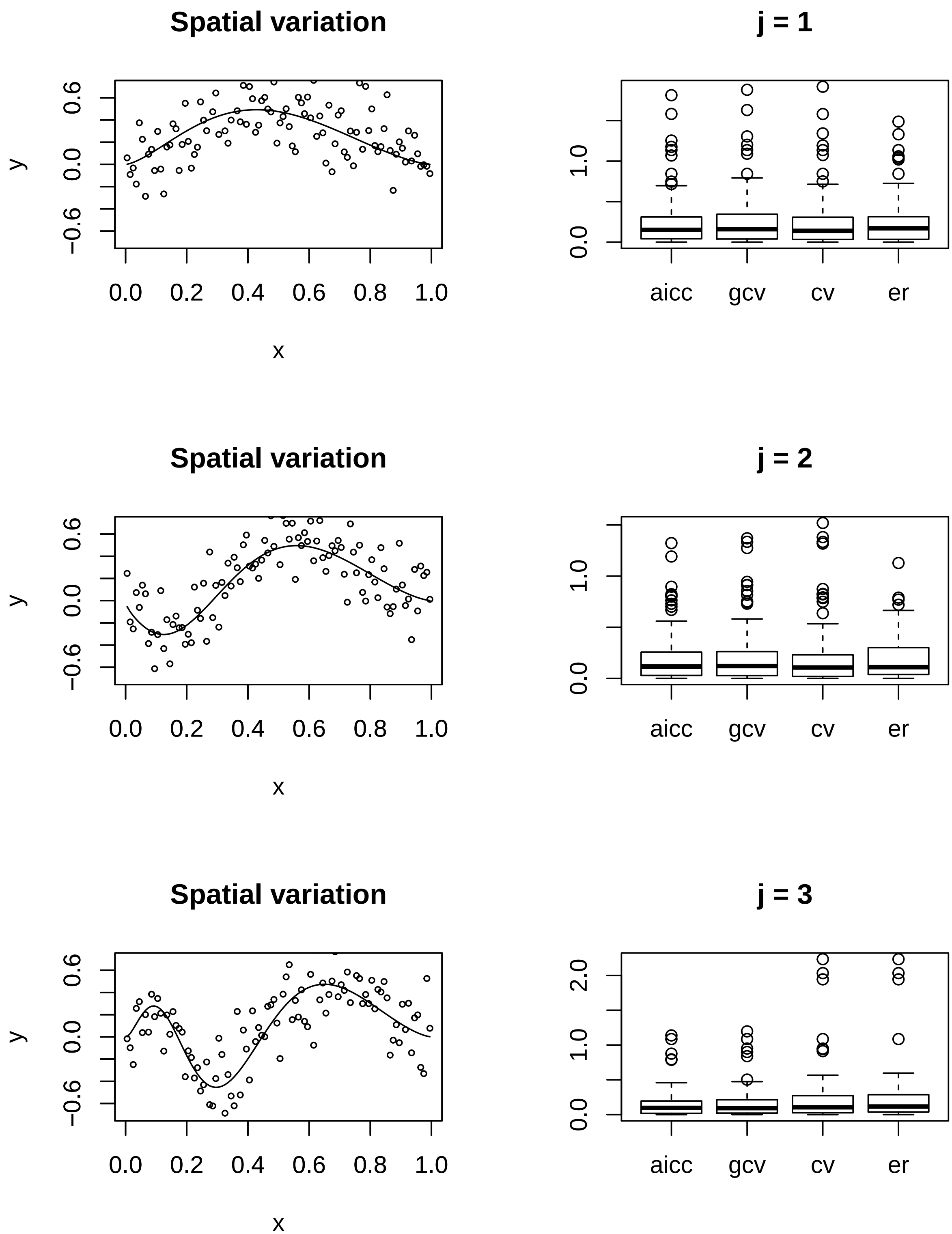


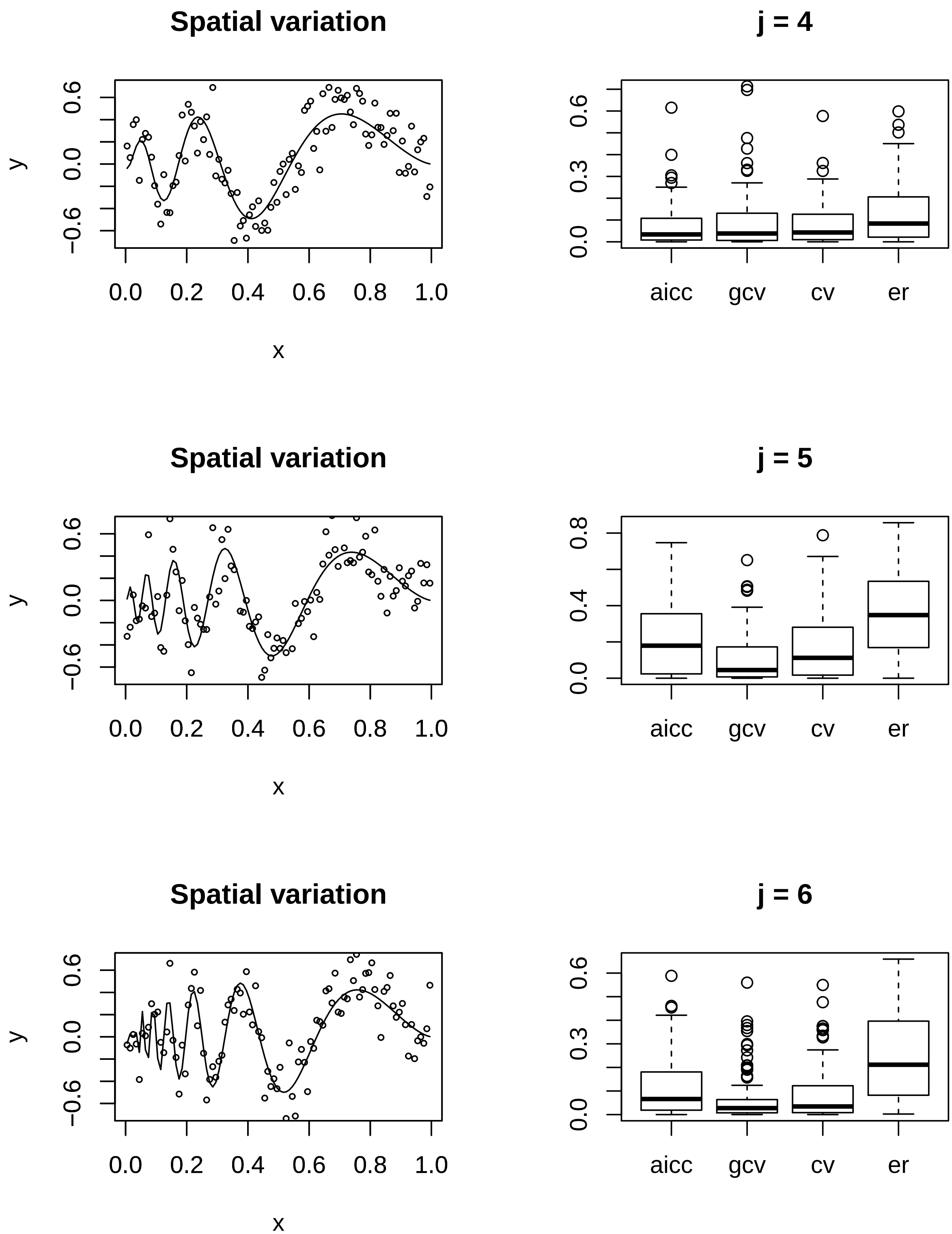



Design Density
j = 3
Design Density
j = 4
Design Density
j = 5
x
y
aicc
gcv
cv
er

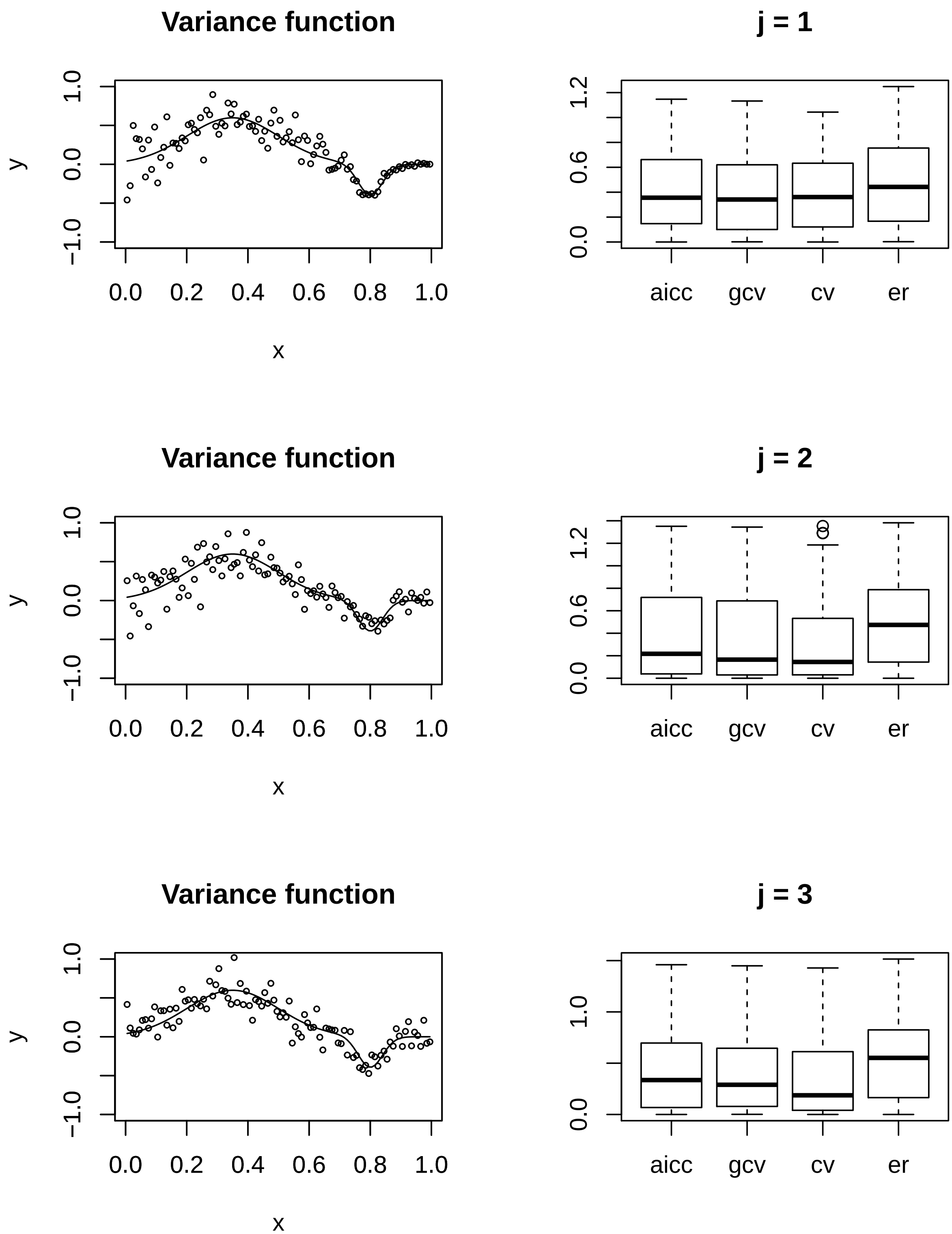

Design Density
y
0.5
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
x
j = 6
3.0
1.5
0.0
aicc
gcv
cv
er

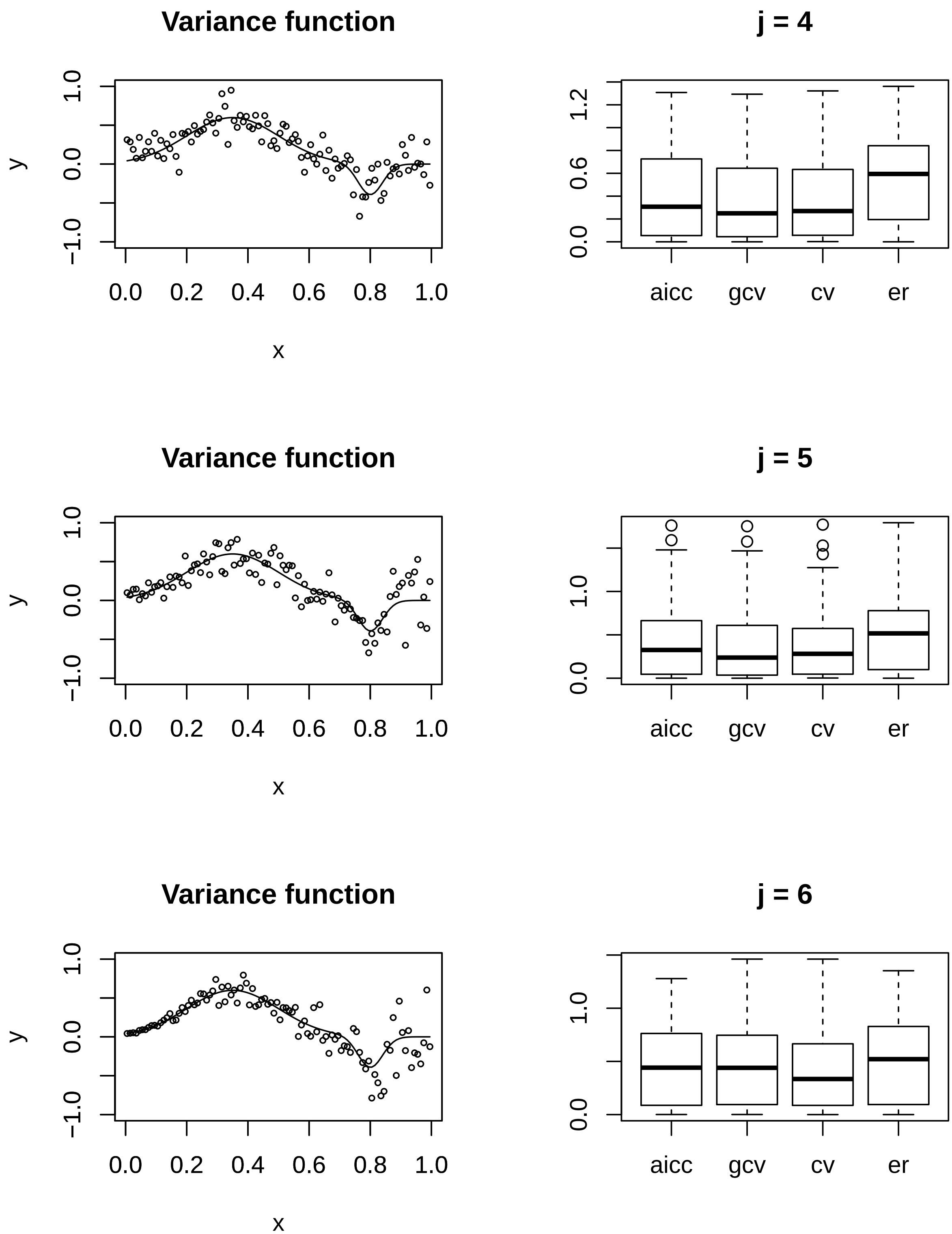
Spatial variation




Spatial variation
j = 4
Spatial variation
j = 5
Spatial variation
j = 6
x
y
0.0
0.2
0.4
0.6
0.8
1.0
−0.6
0.0
0.6
0.3
0.4
0.8
aicc
gcv
cv
er

Variance function




Variance function
j = 4
Variance function
j = 5
Variance function
j = 6
y
x
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
0.0
1.0
1.2
0.6
aicc
gcv
cv
er

Conclusion:

In each pair plot, the left one presents one typical simulated data set together with the true function, the right one is the boxplot of $\log r$. From the plot, we can find there is not a best criterion to choose the smoothing parameter, because the performance of those 4 criteria are so similar. Specially, CV and GCV have very similar results.