On estimation of noise variance in high-dimensional linear models

Ekaterina Krymova
Universität Duisburg-Essen
Yuri Golubev
CNRS, Aix-Marseille Université, I2M

21 December 2017, Marseille

Noise variance estimation in high-dimensional linear model

Ordered spectral regularisations

A family of noise variance estimators

Adaptive smoothing parameter selection for the noise variance estimation

Motivation for an adaptive method and an oracle inequality

An oracle inequality for the adaptive estimator of noise variance

Minimax adaptive estimation of the variance in non-linear regression case

Consider a linear regression model

$$Y = X\beta + \sigma\xi,$$

X is $n \times p$ known real matrix, $\beta \in \mathbb{R}^p$ is an unknown vector, $\xi \in \mathbb{R}^n$ is a vector with i.i.d. standard Gaussian components.

The main goal is to estimate σ^2.

- Efromovich S. and Low M. On optimal adaptive estimation of a quadratic functional. The Annals of Statist. 1996. V. 24. No 3.
- Laurent B. and Massart P. Adaptive estimation of a quadratic functional by model selection. The Annals of Statist. 2000. V. 28. No 5.

In the ideal situation $\beta = 0$, i.e.

$$Y_i = \sigma \xi_i, \quad i = 1, \ldots, n$$

and the best possible estimate is defined as follows

$$\widehat{\sigma}^2_\circ(\xi) \stackrel{\text{def}}{=} \frac{\|\sigma\xi\|^2}{n} = \sigma^2 + \frac{\sigma^2}{n} \sum_{i=1}^{n} (\xi_i^2 - 1).$$

Define the error of estimator $\breve{\sigma}^2(Y)$ as the expectation of

$$\Delta\bigl(\breve{\sigma}^2\bigr) = n\bigl|\breve{\sigma}^2(Y) - \widehat{\sigma}^2_\circ(\xi)\bigr|.$$

Maximum likelihood estimation

$$\widehat{\sigma}^2(Y) = \arg\max_{\sigma^2>0} \max_{\beta\in\mathbb{R}^p} \left\{ -\frac{\|Y - X\beta\|^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) \right\}$$

gives

$$\widetilde{\sigma}_b^2(Y) = \frac{\|Y - X(X^\top X)^{-1}X^\top Y\|^2}{n},$$

which leads to unbiased version

$$\widetilde{\sigma}^2(Y) = \frac{\|Y - X(X^\top X)^{-1}X^\top Y\|^2}{n-p}.$$

Consider Singular Value Decomposition of the matrix $X^\top X$

$$X^\top X \mathbf{e}_k = \lambda_k \mathbf{e}_k, \quad k = 1, \dots, p,$$

where $\mathbf{e}_k \in \mathbb{R}^p$, $k = 1, \dots, p$, $\lambda_1 \geq \lambda_2 \geq, \dots, \geq \lambda_p$. Define (orthonormal) vectors

$$\mathbf{e}_k^* = \frac{X\mathbf{e}_k}{\sqrt{\lambda_k}}, \; k = 1, \dots, p.$$

Complete the basis with $\mathbf{e}_k^*$, $k = p+1, \dots, n$.
Then for $\bar{Y}_k = \langle Y, \mathbf{e}_k^* \rangle$ the initial problem transforms into

$$\begin{aligned} \bar{Y}_k &= \sqrt{\lambda_k}\bar{\beta}_k + \sigma\xi_k', \quad & k = 1, \dots, p, \\ \bar{Y}_k &= \sigma\xi_k', & k = p+1, \dots, n, \end{aligned}$$

where $\bar{\beta}_k = \langle \beta, \mathbf{e}_k \rangle$, ξ_k' are i.i.d. standard Gaussian noise.

The equivalent problem is to estimate σ^2 in the model

$$\begin{aligned} \bar{Y}_k =& \sqrt{\lambda_k}\bar{\beta}_k + \sigma\xi'_k, \quad && k = 1,\dots,p, \\ \bar{Y}_k =& \sigma\xi'_k, && k = p+1,\dots,n, \end{aligned}$$

where $\bar{\beta}_k = \langle \beta, \mathbf{e}_k \rangle$, ξ' is standard Gaussian noise.

$$\widetilde{\sigma}^2(Y) = \frac{\|Y - X(X^\top X)^{-1}X^\top Y\|^2}{n-p} = \frac{1}{n-p}\sum_{k=p+1}^{n} \bar{Y}_k^2.$$

For $p \approx n$ the estimation fails. It is worth using $\bar{Y}_k$, $k = 1,\dots,p$.

Remark. For the estimation we would like to have a method, which is based on the the initial data preferably avoiding SVD (not always possible), and transformation to the equivalent model for the proofs.

Thus one has to "improve" ML estimation of β

$$\widehat{\beta}_\circ(Y) = (X^\top X)^{-1} X^\top Y.$$

Define **spectral regularisations** of $\widehat{\beta}_\circ(Y)$:

$$\widehat{\beta}_\alpha(Y) = H_\alpha(X^\top X)\widehat{\beta}_\circ(Y),$$

where

$$H_\alpha(X^\top X) = \sum_{i=1}^{p} H_\alpha(\lambda_i)\mathbf{e}_i\mathbf{e}_i^\top,$$

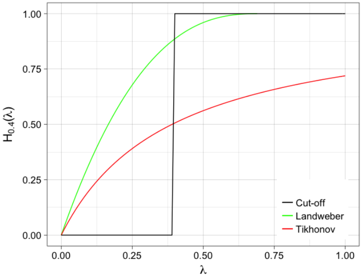
$H_\alpha(\cdot) : \mathbb{R}^+ \to [0,1]$ indexed by $\alpha \in \mathbb{R}^+$ s.t.

$$\lim_{\alpha\to 0} H_\alpha(\lambda) = 1, \quad \text{for all} \quad \lambda > 0;$$
$$\lim_{\lambda\to 0} H_\alpha(\lambda) = 0, \quad \text{for all} \quad \alpha > 0.$$

Ordered smoothers: for all $\alpha, \alpha' \in \mathbb{R}^+$

$$H_\alpha(\lambda) \leq H_{\alpha'}(\lambda) \text{ for all } \lambda > 0$$
or
$$H_{\alpha'}(\lambda) \leq H_\alpha(\lambda) \text{ for all } \lambda > 0.$$

- spectral cut-off: $H_\alpha(\lambda) = 1\{\lambda \geq \alpha\}$,
- Tikhonov rigularization: $H_\alpha(\lambda) = \frac{\lambda}{\lambda+\alpha}$,
    - smoothing splines
- Landweber iterations $H_\alpha(\lambda) = 1 - (1 - \frac{\lambda}{a})^{\alpha^{-1}}$, where $\lambda_1 a < 1$, $\alpha = (k+1)^{-1}$.



+ kernel estimators, etc.:
Engl, H.W., Hanke, M., and Neubauer, A. (1996). Regularization of Inverse Problems. Mathematics and its Applications

Remark. For the estimation we would like to have a method, which is based on the the initial data preferably avoiding SVD (not always possible), and transformation to the equivalent model for the proofs.

For Tikhonov regularization there's no need of SVD:

$$\widehat{\beta}_\alpha^{\mathrm{T}}(Y) = \arg\min_\beta \Big\{ \|Y - X\beta\|^2 + \alpha\|\beta\|^2 \Big\} = (\alpha I + X^\top X)^{-1} X^\top Y.$$

One has just to find a solution to a linear system

$$(\alpha I + X^\top X)\widehat{\beta}_\alpha^{\mathrm{T}}(Y) = X^\top Y$$

and

$$\widehat{\beta}_\alpha^{\mathrm{T}}(Y) = (\alpha I + X^\top X)^{-1} X^\top Y = H_\alpha^{\mathrm{T}}(X^\top X)\big[(X^\top X)^{-1} X^\top Y\big],$$

where $H_\alpha^{\mathrm{T}}(\lambda) = \frac{\lambda}{\alpha+\lambda}$.

Note that for $H_\alpha(\lambda) = \mathbf{1}\{\lambda \geq \alpha\}$ SVD is needed.

For the family $\{H_\alpha(\cdot),\, \alpha \in \mathbb{R}^+\}$ define a family of estimators

$$\widehat{\sigma}^2_\alpha(Y) = \frac{\left\| Y - X\widehat{\beta}_\alpha(Y) \right\|^2}{n - [2\mathrm{tr}\{H_\alpha(X^\top X)\} - \mathrm{tr}\{[H_\alpha(X^\top X)]^2\}]},$$

Equivalently

$$\widehat{\sigma}^2_\alpha(\bar{Y}) = \frac{\sum_{i=1}^n [1 - H_\alpha(\lambda_i)]^2 \bar{Y}_i^2}{\sum_{i=1}^n [1 - H_\alpha(\lambda_i)]^2}.$$

In unbiased version of ML estimator the denominator is $n - p$, thus $\sum_{i=1}^n \left[2H_\alpha(\lambda_i) - [H_\alpha(\lambda_i)]^2\right]$ is the "effective dimension" of the predicted $\widehat{\beta}_\alpha(Y)$.

Denote $G_\alpha(\lambda) = 1 - [1 - H_\alpha(\lambda)]^2 = 2H_\alpha(\lambda) - [H_\alpha(\lambda)]^2$.

To select $\hat{\alpha}$, corresponding to the best estimate in the family $\{\widehat{\sigma}^2_\alpha(Y),\, \alpha \in \mathcal{A} = [\alpha_{\min}, \alpha_{\max}]\}$, we minimise the expectation of the error

$$\Delta\left(\widehat{\sigma}^2_\alpha\right) = n\left|\widehat{\sigma}^2_\alpha(Y) - \widehat{\sigma}^2_\circ(\xi)\right|,$$

in $\alpha \in \mathcal{A} = [\alpha_{\min}, \alpha_{\max}]$, $\widehat{\sigma}^2_\circ(\xi) \overset{\mathrm{def}}{=} \frac{\|\sigma\xi\|^2}{n}$.

Denote

$$G_\alpha(\lambda) = 1 - [1 - H_\alpha(\lambda)]^2 = 2H_\alpha(\lambda) - [H_\alpha(\lambda)]^2.$$

Condition A: $\frac{1}{n}\sum_{i=1}^n G_\alpha(\lambda_i) \ll 1$

$$\widehat{\sigma}_\alpha^2(\bar{Y}) \approx \frac{1}{n}\left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right]\sum_{i=1}^n [1 - H_\alpha(\lambda_i)]^2\,\bar{Y}_i^2.$$

Condition B $\exists K\,\forall\alpha > 0$

$$\sum_{i=1}^n G_\alpha(\lambda_i) \le K \sum_{i=1}^n G_\alpha^2(\lambda_i)$$

$$
\begin{aligned}
\Delta\left(\widehat{\sigma}_\alpha^2\right) =& n\left|\widehat{\sigma}_\alpha^2(\bar{Y}) - \widehat{\sigma}_\circ^2(\xi')\right| \\
\approx& \Bigg|\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2\lambda_i\bar{\beta}_i^2 \\
&+\sigma^2\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{k=1}^{n} G_\alpha(\lambda_k)(1 - \xi_k'^2) \\
&+\sigma^2\frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\sum_{i=1}^{n}(\xi_i'^2 - 1) \\
&+2\sigma\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2\xi_i'\sqrt{\lambda_i}\bar{\beta}_i\Bigg|.
\end{aligned}
$$

$$
\begin{aligned}
\Delta\left(\widehat{\sigma}^2_\alpha\right) =& n\left|\widehat{\sigma}^2_\alpha(\bar{Y}) - \widehat{\sigma}^2_\circ(\xi')\right| \\
\approx& \Bigg|\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2\lambda_i\bar{\beta}_i^2 \\
&+\sigma^2\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{k=1}^{n} G_\alpha(\lambda_k)(1-\xi_k'^2) \qquad \sim \sqrt{\textstyle\sum_{i=1}^{n} G_\alpha^2(\lambda_i)} \\
&+\sigma^2\frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\sum_{i=1}^{n}(\xi_i'^2 - 1) \\
&+2\sigma\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2\xi_i'\sqrt{\lambda_i}\bar{\beta}_i\Bigg|.
\end{aligned}
$$

$$\begin{aligned}
\Delta\left(\widehat{\sigma}^2_\alpha\right) =& n\left|\widehat{\sigma}^2_\alpha(\bar{Y}) - \widehat{\sigma}^2_\circ(\xi')\right| \\
\approx& \Bigg|\left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right]\sum_{i=1}^n \left[1 - H_\alpha(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2 \\
&+\sigma^2\left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right]\sum_{k=1}^n G_\alpha(\lambda_k)(1 - \xi_k'^2) \qquad \sim \sqrt{\textstyle\sum_{i=1}^n G_\alpha^2(\lambda_i)} \\
&+\sigma^2\frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\sum_{i=1}^n(\xi_i'^2 - 1) \qquad \sim n^{-1/2}\textstyle\sum_{i=1}^n G_\alpha(\lambda_i) \\
&+2\sigma\left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right]\sum_{i=1}^n \left[1 - H_\alpha(\lambda_i)\right]^2 \xi_i'\sqrt{\lambda_i}\bar{\beta}_i\Bigg|.
\end{aligned}$$

$$
\begin{aligned}
\Delta\left(\widehat{\sigma}_\alpha^2\right) =& n\left|\widehat{\sigma}_\alpha^2(\bar{Y}) - \widehat{\sigma}_\circ^2(\xi')\right| \\
\approx& \Bigg| \left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right] \sum_{i=1}^n \big[1 - H_\alpha(\lambda_i)\big]^2 \lambda_i \bar{\beta}_i^2 \\
&+\sigma^2 \left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right] \sum_{k=1}^n G_\alpha(\lambda_k)(1 - \xi_k'^2) \qquad \sim \sqrt{\textstyle\sum_{i=1}^n G_\alpha^2(\lambda_i)} \\
&\cancel{+\sigma^2 \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k) \sum_{i=1}^n (\xi_i'^2 - 1)} \qquad \sim n^{-1/2} \textstyle\sum_{i=1}^n G_\alpha(\lambda_i) \\
&+2\sigma \left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right] \sum_{i=1}^n \big[1 - H_\alpha(\lambda_i)\big]^2 \xi_i' \sqrt{\lambda_i} \bar{\beta}_i \Bigg|.
\end{aligned}
$$

$$
\begin{aligned}
\Delta\left(\widehat{\sigma}_\alpha^2\right) =& n\left|\widehat{\sigma}_\alpha^2(\bar{Y}) - \widehat{\sigma}_\circ^2(\xi')\right| \\
\approx& \Bigg| \left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right] \sum_{i=1}^n \left[1 - H_\alpha(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2 \\
&+\sigma^2 \left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right] \sum_{k=1}^n G_\alpha(\lambda_k)(1 - \xi_k'^2) \qquad \sim \sqrt{\textstyle\sum_{i=1}^n G_\alpha^2(\lambda_i)} \\
&\cancel{+\sigma^2 \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k) \sum_{i=1}^n (\xi_i'^2 - 1)} \qquad \sim n^{-1/2} \textstyle\sum_{i=1}^n G_\alpha(\lambda_i) \\
&+2\sigma \left[1 + \frac{1}{n}\sum_{k=1}^n G_\alpha(\lambda_k)\right] \sum_{i=1}^n \left[1 - H_\alpha(\lambda_i)\right]^2 \xi_i' \sqrt{\lambda_i}\bar{\beta}_i \Bigg|
\end{aligned}
$$

small compared to the first term

$$
\begin{aligned}
\Delta\left(\widehat{\sigma}_\alpha^2\right) =& n\left|\widehat{\sigma}_\alpha^2(\bar{Y}) - \widehat{\sigma}_\circ^2(\xi')\right| \\
\approx& \Bigg|\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2\lambda_i\bar{\beta}_i^2 \\
&+\sigma^2\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{k=1}^{n} G_\alpha(\lambda_k)(1-\xi_k'^2) \qquad \sim \sqrt{\textstyle\sum_{i=1}^n G_\alpha^2(\lambda_i)} \\
&\cancel{+\sigma^2\frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\sum_{i=1}^{n}(\xi_i'^2-1)} \qquad \sim n^{-1/2}\textstyle\sum_{i=1}^n G_\alpha(\lambda_i) \\
&\cancel{+2\sigma\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right]\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2\xi_i'\sqrt{\lambda_i}\bar{\beta}_i}\Bigg|
\end{aligned}
$$

small compared to the first term

$$\Delta\big(\widehat{\sigma}^2_\alpha\big) \approx \left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right] \left|\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 \sum_{k=1}^{n} G_\alpha(\lambda_k)(1 - \xi_k'^2)\right|.$$

To minimize $\mathbf{E}\Delta\big(\widehat{\sigma}^2_\alpha\big)$ we need to

1. construct a deterministic bound for the process

$$\zeta(\alpha) = \sum_{k=1}^{n} G_\alpha(\lambda_k)(1 - \xi_k'^2), \quad \alpha \in \mathbb{R}^+,$$

namely one needs to find a minimal deterministic function $V(\alpha)$ s.t.

$$\mathbf{E} \sup_{\alpha \le \alpha_{\max}} \big[|\zeta(\alpha)| - V(\alpha)\big]_+ \le C\mathbf{E}\big[|\zeta(\alpha_{\max})| - V(\alpha_{\max})\big]_+ \le C\sqrt{\mathbf{E}\zeta^2(\alpha_{\max})};$$

2. estimate $\lambda_i \bar{\beta}_i^2$ and σ^2 .

Theorem 1 (Deterministic bound)

For each $\epsilon \in (0,1]$

$$\mathbf{E} \sup_{\alpha \le \alpha_{\max}} \left[|\zeta(\alpha)| - V_\epsilon(\alpha) \right]_+ \le C\epsilon^{-1} \sqrt{D(\alpha_{\max})},$$

where

$$D(\alpha) = \sum_{k=1}^{n} G_\alpha^2(\lambda_k),$$

$$V_\epsilon(\alpha) = (1+\epsilon)\sqrt{2D(\alpha)} \left\{ \log \frac{D(\alpha)}{D(\alpha_{\max})} + 2(1+\epsilon) \log \left[\frac{Q}{\epsilon^2} \log \frac{D(\alpha)}{D(\alpha_{\max})} \right] \right\}^{1/2},$$

$$Q = \frac{4}{(\sqrt{2}-1)^2}.$$

$$\Delta\left(\widehat{\sigma}^2_\alpha\right) \approx \left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right] \left|\sum_{i=1}^{n}\left[1 - H_\alpha(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 \sum_{k=1}^{n} G_\alpha(\lambda_k)(1 - \xi_k'^2)\right|.$$

1. Applying Theorem 1, for any $\widetilde{\alpha}(Y)$

$$\mathbf{E}\Delta\left(\widehat{\sigma}^2_{\widetilde{\alpha}}\right) \lesssim \mathbf{E}\left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right] \left\{\sum_{i=1}^{n}\left[1 - H_{\widetilde{\alpha}}(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 V_\epsilon(\widetilde{\alpha})\right\} + C\sigma^2 \frac{\sqrt{D(\alpha_{\max})}}{\epsilon},$$

which leads to

$$\widetilde{\alpha}(\beta) = \arg\min_{\alpha\in\mathcal{A}} \left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right] \left\{\sum_{i=1}^{n}\left[1 - H_\alpha(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 V_\epsilon(\alpha)\right\}.$$

$$\widetilde{\alpha}(\beta) = \arg\min_{\alpha \in \mathcal{A}} \left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right] \left\{\sum_{i=1}^{n} \big[1 - H_\alpha(\lambda_i)\big]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 V_\epsilon(\alpha)\right\}.$$

2. Substitute

- $\lambda_i \bar{\beta}_i^2$ with $\bar{Y}_i^2 - \sigma^2$;
- σ^2 with $\frac{\left\|Y - X\widehat{\beta}_\alpha(Y)\right\|^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\big[1 - H_\alpha(\lambda_i)\big]^2 Y_i^2$.

Thus

$$\widehat{\alpha}(\bar{Y}) = \arg\min_{\alpha \in \mathcal{A}} \left[1 + \frac{1}{n}\sum_{k=1}^{n} G_\alpha(\lambda_k)\right] \left\{\sum_{i=1}^{n} \big[1 - H_\alpha(\lambda_i)\big]^2 \bar{Y}_i^2 \left[1 + \frac{V_\epsilon(\alpha)}{n}\right]\right\}.$$

The data-driven smoothing parameter selection procedure:

$$\widehat{\alpha}(Y) \approx \arg\min_{\alpha \in \mathcal{A}} \left\{ \widehat{\sigma}^2_{\alpha}(Y) \left[1 + \frac{V_{\epsilon}(\alpha)}{n} \right] \right\}.$$

The corresponding estimator:

$$\widehat{\sigma}^2_{\widehat{\alpha}}(Y) = \frac{1}{n} \left\| Y - X \widehat{\beta}_{\widehat{\alpha}}(Y) \right\|^2 \left[1 - \frac{2\mathrm{tr}\left\{ H_{\widehat{\alpha}}(X^\top X) \right\} - \mathrm{tr}\left\{ \left[H_{\widehat{\alpha}}(X^\top X) \right]^2 \right\}}{n} \right]^{-1}.$$

Denote

$$R_\epsilon(\alpha,\beta) \stackrel{\text{def}}{=} \left[1+\frac{V_\epsilon(\alpha)}{n}\right]\left\{\frac{\sum_{i=1}^n \left[1-H_\alpha(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2}{\sum_{k=1}^n [1-H_\alpha(\lambda_k)]^2} + \sigma^2 V_\epsilon(\alpha)\right\},$$

$$r_{\mathcal{A},\epsilon}(\beta) \stackrel{\text{def}}{=} \min_{\alpha\in\mathcal{A}} R_\epsilon(\alpha,\beta), \quad \rho_{\mathcal{A},\epsilon}(\beta) \stackrel{\text{def}}{=} \frac{\sigma^2\sqrt{D(\alpha_{\max})}}{r_{\mathcal{A},\epsilon}(\beta)}.$$

Theorem 2

Under conditions A,B and $\alpha_{\min}$, $\alpha_{\max}$ s.t. $\lim_{n\to\infty} \frac{D(\alpha_{\min})}{n} = 0$, $D(\alpha_{\max}) \geq 5$ for each $\gamma \in \big(0, \epsilon/(1+\epsilon)\big)$, and all $n \geq n_\gamma$

$$\mathbf{E}\Delta(\widehat{\sigma}^2_{\widehat{\alpha}}) \leq \frac{r_{\mathcal{A},\epsilon}(\beta)}{\gamma}\left\{1 + \frac{\log^{-1/8}\left[\rho^{-1}_{\mathcal{A},\epsilon}(\beta)\right]}{\sqrt[4]{D(\alpha_{\max})}} + \left[\frac{C\rho_{\mathcal{A},\epsilon}(\beta)}{(\epsilon-\gamma-\gamma\epsilon)}\right]^{1/2}\right\}^2.$$

Denote

$$R_\epsilon(\alpha,\beta) \stackrel{\text{def}}{=} \left[1+\frac{V_\epsilon(\alpha)}{n}\right]\left\{\frac{\sum_{i=1}^n \left[1-H_\alpha(\lambda_i)\right]^2 \lambda_i \bar{\beta}_i^2}{\sum_{k=1}^n [1-H_\alpha(\lambda_k)]^2} + \sigma^2 V_\epsilon(\alpha)\right\},$$

$$r_{\mathcal{A},\epsilon}(\beta) \stackrel{\text{def}}{=} \min_{\alpha\in\mathcal{A}} R_\epsilon(\alpha,\beta), \quad \rho_{\mathcal{A},\epsilon}(\beta) \stackrel{\text{def}}{=} \frac{\sigma^2\sqrt{D(\alpha_{\max})}}{r_{\mathcal{A},\epsilon}(\beta)}.$$

small

Theorem 2

Under conditions A,B and $\alpha_{\min}$, $\alpha_{\max}$ *s.t.* $\lim_{n\to\infty} \frac{D(\alpha_{\min})}{n} = 0$, $D(\alpha_{\max}) \geq 5$ *for each* $\gamma \in \big(0, \epsilon/(1+\epsilon)\big)$, *and all* $n \geq n_\gamma$

$$\mathbf{E}\Delta(\widehat{\sigma}^2_{\widehat{\alpha}}) \leq \frac{r_{\mathcal{A},\epsilon}(\beta)}{\gamma}\left\{1+\frac{\log^{-1/8}\left[\rho^{-1}_{\mathcal{A},\epsilon}(\beta)\right]}{\sqrt[4]{D(\alpha_{\max})}} + \left[\frac{C\rho_{\mathcal{A},\epsilon}(\beta)}{(\epsilon-\gamma-\gamma\epsilon)}\right]^{1/2}\right\}^2.$$

Remark

For the selection of best smoothing parameter for the unknown β estimation:

$$\arg\min_{\alpha \in \mathcal{A}} \left\{ \sum_{i=1}^{n} [1 - H_\alpha(\lambda_i)]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 \sum_{i=1}^{n} H_\alpha^2(\lambda_i) \right\}$$

and for the unknown σ^2 estimation:

$$\arg\min_{\alpha \in \mathcal{A}} \left\{ \sum_{i=1}^{n} [1 - H_\alpha(\lambda_i)]^2 \lambda_i \bar{\beta}_i^2 + \sigma^2 V_\epsilon(\alpha) \right\}.$$

For the small α (the models with the high effective dimension)

$$V_\epsilon(\alpha) \ll \sum_{i=1}^{n} H_\alpha^2(\lambda_i).$$

Let us apply the Theorem 2 to minimax adaptive estimation of the noise variance given the noisy observations of a smooth non-linear regression function

$$Y_i = f(X_i) + \sigma\xi_i, \quad i = 1, \dots, n,$$

where ξ is standard Gaussian, $X_i \in [0, 1]$, $f(x), x \in [0, 1]$ is a function from the class

$$\mathcal{W}_2^m = \left\{ f : \int_0^1 [f^{(m)}(x)]^2 \le L \right\};$$

Consider the smoothing spline method

$$\widehat{f}_\alpha(\cdot, Y) = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^n \bigl[Y_i - f(X_i) \bigr]^2 + \alpha \int_0^1 [f^{(m)}(x)]^2 \, dx \right\}.$$

Denote

$$\bar{\Delta}(\widetilde{\sigma}^2, \mathcal{W}_2^m) = \max_{f \in \mathcal{W}_2^m} n\mathbf{E} \left| \widetilde{\sigma}^2(Y) - n^{-1}\sigma^2 \|\xi\|^2 \right|,$$

where $\widetilde{\sigma}^2(Y)$ is some estimator of the noise variance.

In Demmler-Reinsch basis $\{\phi_k\}_{k=1,\dots,n}$,

$$\frac{1}{n}\sum_{i=1}^{n}\phi_k(X_i)\phi_s(X_i)=\delta_{sk},\quad \int_0^1 \phi_k^{(m)}(x)\phi_s^{(m)}(x)=\nu_k^n\delta_{ks},$$

with eigenvalues

$$\nu_1^n\le\nu_2^n\le\cdots\le\nu_n^n.$$

Asymptotically

$$\nu_k^n=(1+o(1))(\pi k)^{2m}.$$

$$\bar{Y}_k=\bar{f}_k+\frac{\sigma}{\sqrt{n}}\xi_k',\quad k=1,\dots,n,$$

where

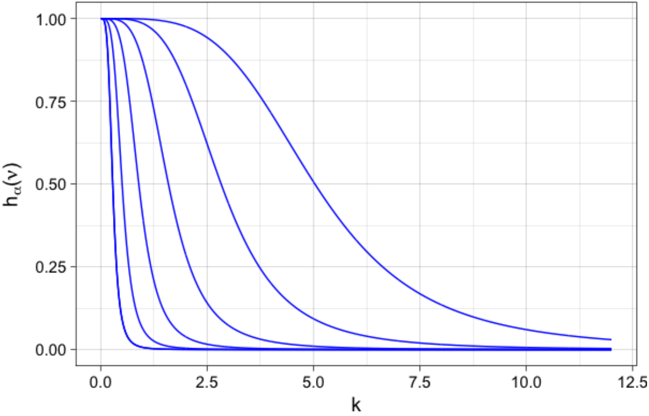
$$\bar{f}_k=\frac{1}{n}\sum_{i=1}^{n}f(X_i)\phi_k(X_i)\quad \bar{Y}_k=\frac{1}{n}\sum_{i=1}^{n}Y_i\phi_k(X_i).$$

Thus we transformed the regression model into the sequence space model with

$$\sigma:=\frac{\sigma}{\sqrt{n}},\quad \lambda_k=\frac{1}{\nu_k^n},\quad \beta_k=\sqrt{\nu_k^n}\bar{f}_k.$$

And the spline estimate is defined as

$$\widehat{f}_\alpha(x, Y) = \sum_{k=1}^{n} h_\alpha(\nu_k^n)\bar{Y}_k\phi_k(x), \quad \text{where } h_\alpha(z) = \frac{1}{1+\alpha z}, \quad z > 0.$$



The bounding function

$$V_\epsilon(\alpha) = (1 + \epsilon + o(1))\alpha^{-1/(4m)} \sqrt{\frac{K(m)}{\pi m} \log \frac{\alpha_{\max}}{\alpha}}.$$

The maximal bias on the class $\mathcal{W}_2^m = \left\{ \bar{f}_k : \sum_{k=1}^n \nu_k^n \bar{f}_k^2 \leq L \right\}$

$$\sup_{f \in \mathcal{W}_2^m} \sum_{k=1}^n [1 - h_\alpha(\nu_k^n)]^2 \bar{f}_k^2 = L \max_k \frac{[1 - h_\alpha(\nu_k^n)]^2}{\nu_k^n} \leq L\alpha.$$

Thus for any $f \in \mathcal{W}_2^m$ for $n \to \infty$

$$r_{\mathcal{A},\epsilon}(f) \asymp \left(\frac{\sigma^2}{n}\right)^{4m/(4m+1)} L^{1/(4m+1)} \left[\log\left(\frac{Ln}{\sigma^2}\right)\right]^{2m/(4m+1)}.$$

Thus

$$\rho_{\mathcal{A},\epsilon}(f) \asymp \left(\frac{nL}{\sigma^2}\right)^{-1/(4m+1)} \left[\log\left(\frac{Ln}{\sigma^2}\right)\right]^{-2m/(4m+1)} \to 0.$$

Denote

$$W(\alpha) = \sum_{k=1}^{n} [2h_\alpha(\nu_k^2) - h_\alpha^2(\nu_k^2)].$$

The noise variance estimator is

$$\widehat{\sigma}^2_{\widehat{\alpha}}(Y) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \widehat{f}_{\widehat{\alpha}}(X_i, Y)]^2 \left[1 - \frac{W(\widehat{\alpha})}{n}\right]^{-1},$$

where

$$\widehat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \left\{ \sum_{i=1}^{n} [Y_i - \widehat{f}_\alpha(X_i, Y)]^2 \left[1 - \frac{W(\alpha)}{n}\right]^{-1} \left[1 + \frac{V_\epsilon(\alpha)}{n}\right] \right\}.$$

From Theorem 2 the upper bound on the maximal error in the class $\mathcal{W}_2^m$ is

$$\Delta(\widehat{\sigma}^2_{\widehat{\alpha}}, \mathcal{W}_2^m) \overset{\text{asymp}}{\leq} \frac{C(m) L^{\frac{1}{4m+1}}}{\gamma} \left(\frac{\sigma^2}{n}\right)^{\frac{4m}{4m+1}} \left[\log\left(\frac{Ln}{\sigma^2}\right)\right]^{\frac{2m}{4m+1}},$$

where $\gamma < \epsilon/(1+\epsilon)$.

Remark

The bound is not optimal

$$\Delta(\widehat{\sigma}^2_{\widehat{\alpha}}, \mathcal{W}_2^m) \overset{\text{asymp}}{\leq} \frac{C(m)L^{\frac{1}{4m+1}}}{\gamma} \left(\frac{\sigma^2}{n}\right)^{\frac{4m}{4m+1}} \left[\log\left(\frac{Ln}{\sigma^2}\right)\right]^{\frac{2m}{4m+1}},$$

probably up to $\frac{1}{\gamma}$.

- Efromovich S. and Low M. On optimal adaptive estimation of a quadratic functional. The Annals of Statist. 1996. V. 24. No 3.
- Laurent B. and Massart P. Adaptive estimation of a quadratic functional by model selection. The Annals of Statist. 2000. V. 28. No 5.

Problem of quadratic functional $\sum_i \bar{s}_i^2$ estimation in the model

$$\bar{Y}_i = \bar{s}_i + \sigma\xi_i, \quad i = 1, \ldots, n,$$

where ξ_i are i.i.d. $\mathcal{N}(0,1)$.
In our case as an estimate of the quadratic functional one might use $\sum_{i=1}^n Y_i^2 - n\widehat{\sigma}^2_{\widehat{\alpha}}(Y)$.

Thank you for your attention!

$$\widetilde{V}_\epsilon(\alpha) = \sqrt{2D(\alpha)}\left\{\log \frac{D(\alpha)}{D(\alpha_{\mathsf{max}})} + 2(1+\epsilon)\left[\log\log \frac{D(\alpha)}{D(\alpha_{\mathsf{max}})} + \log\frac{1}{\epsilon}\right]\right\}^{1/2}.$$