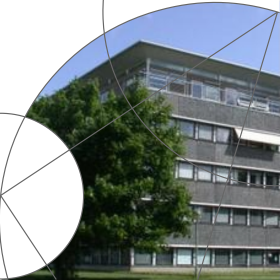
Faculty of Science

Post-selection inference: risk estimation after data-driven model selection

Code for simulations

Niels Richard Hansen, Frederik Riis Mikkelsen &
Alexander Sokol
Department of Mathematical Sciences



July 14, 2020

AIC for Gaussian regression (fixed variance)

If $Y \sim \mathcal{N}(\xi, \sigma^2 I)$ then

$$\mathrm{AIC} = \|Y - \hat{\xi}\|^2/\sigma^2 + 2d$$

when σ^2 is fixed and $\hat{\xi}$ is the least squares estimator in a subset of dimension d.

Fix $\sigma^2 = 1$ from hereon, and for $\lambda \in \Lambda$ (an index set) let

$$\mathrm{AIC}(\lambda) = \|Y - \hat{\xi}^\lambda\|^2 + 2d(\lambda).$$

Example: $\xi = X\beta$ for X and $n \times p$ matrix and $\beta \in \mathbb{R}^p$.

AIC as a test error and risk estimate

Let $Y \perp\!\!\!\perp Y^{\mathrm{New}}$ and $Y \overset{\mathcal{D}}{=} Y^{\mathrm{New}}$.

If $Y \sim (\xi, I)$, $\hat{\xi}^{\lambda} = S_{\lambda} Y$ and $d(\lambda) = \mathrm{tr}(S_{\lambda})$ then

$$E(\mathrm{AIC}(\lambda)) = E\|Y^{\mathrm{New}} - \hat{\xi}^{\lambda}\|^2 = n + \underbrace{E\|\xi - \hat{\xi}^{\lambda}\|^2}_{\mathrm{MSE}}.$$

Thus

$$\mathrm{AIC}(\lambda) - n = \|Y - S_{\lambda} Y\|^2 + 2d(\lambda) - n$$

is an unbiased estimate of MSE.

Forward stepwise variable selection

If S_λ is a fixed projection then $d(\lambda) = \mathrm{rank}(S_\lambda)$.

Forward stepwise variable selection results in a sequence of projections

$$S_0, \ldots, S_p$$

onto nested subspaces of dimensions $0 < 1 < 2 \ldots < p$.

Note: S_d is selected in a data dependent way.

Model weights and model averaging

Introduce weights

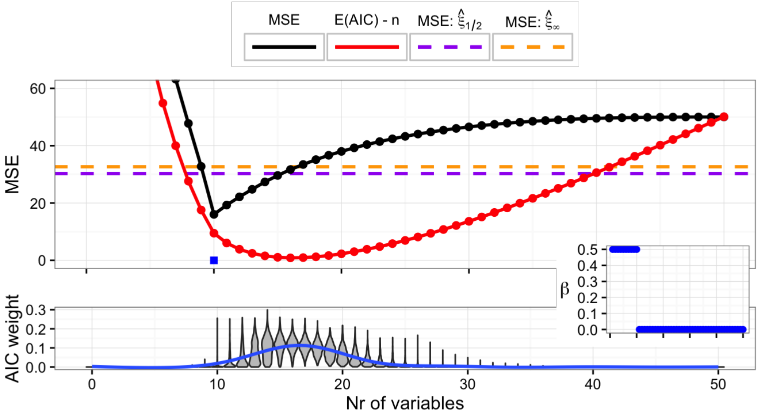
$$w_\gamma(\lambda) = \frac{\exp(-\gamma \mathrm{IC}(\lambda))}{\int \exp(-\gamma \mathrm{IC}(\lambda)) \pi(\mathrm{d}\lambda)}.$$

- $\gamma = 1/2$ has a Bayes interpretation
- $\gamma \to 0$ gives all models the same weight
- $\gamma \to \infty$ concentrates the weights on models with minimal IC.

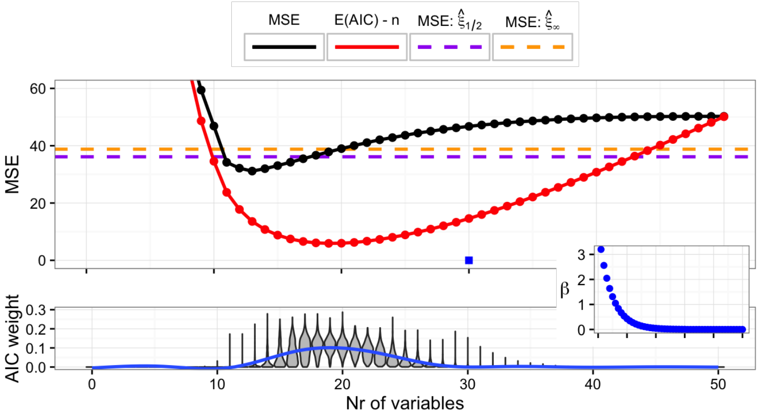
$$\hat{\xi}_\gamma = \int \hat{\xi}(\lambda) w_\gamma(\lambda) \pi(\mathrm{d}\lambda)$$

is the model averaging estimator.

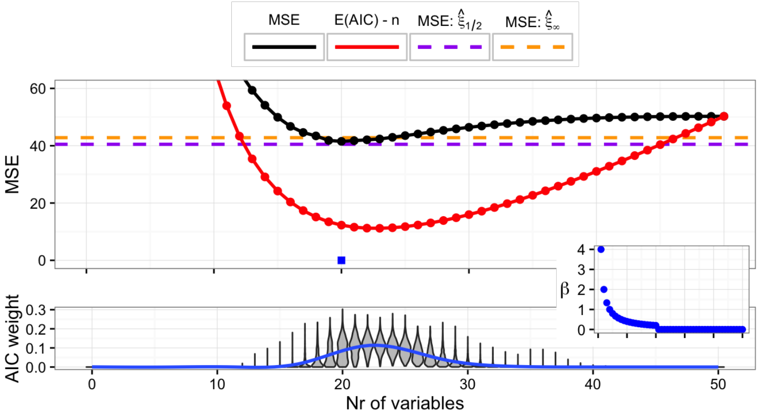
$n = 100,\ p = 50$


MSE
E(AIC) - n
MSE: $\hat{\xi}_{1/2}$
MSE: $\hat{\xi}_{\infty}$
60
40
20
0
MSE
0.5
0.4
0.3
0.2
0.1
0.0
β
0.3
0.2
0.1
0.0
AIC weight
0
10
20
30
40
50
Nr of variables

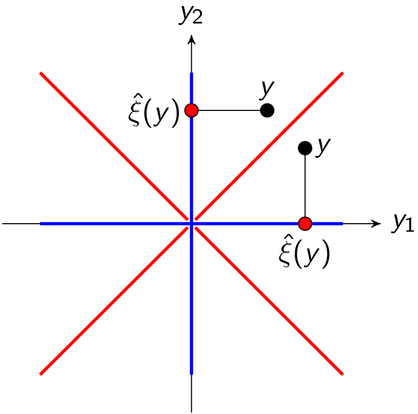
$n = 100,\ p = 50$


MSE
E(AIC) - n
MSE: $\hat{\xi}_{1/2}$
MSE: $\hat{\xi}_{\infty}$
60
40
20
0
MSE
AIC weight
0.3
0.2
0.1
0.0
0
10
20
30
40
50
Nr of variables
β
3
2
1
0

$n = 100, \ p = 50$


MSE
E(AIC) - n
MSE: $\hat{\xi}_{1/2}$
MSE: $\hat{\xi}_{\infty}$
60
40
20
0
MSE
4
3
2
1
0
β
0.3
0.2
0.1
0.0
AIC weight
0
10
20
30
40
50
Nr of variables

Best subspace selection



Best subspace selection is the projection onto the union of subspaces. The estimator is discontinous on the union of the diagonals.

Fundamental identity

Recall the fundamental AIC identity

$$\underbrace{E\|Y^{\text{New}} - \hat{\xi}\|^2}_{\text{expected test error}} = \underbrace{E\|Y - \hat{\xi}\|^2}_{\text{expected training error}} + 2d,$$

which justifies

$$\text{AIC} = \|Y - \hat{\xi}\|^2 + 2d$$

as a prediction error estimate and $\text{AIC} - n$ as a risk estimate.

For Lipschitz continuous estimators (Stein's lemma)

$$d = E(\nabla \cdot \hat{\xi}).$$

Theorem (NRH, Mikkelsen, Sokol)

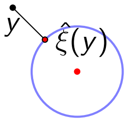
In general

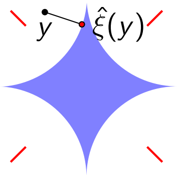
$$d = E(\nabla \cdot \hat{\xi}) + \frac{1}{(2\pi)^{n/2}} \int e^{-\frac{\|y-\xi\|^2}{2}} \mathrm{d}\nu(\mathrm{d}y)$$

with ν a measure singular w.r.t. Lebesgue measure.

The singular measure ν

$\mathcal{H}^{n-1}$ is Hausdorff measure.





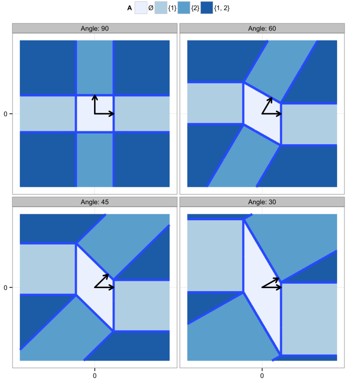
- If $E \subseteq \mathbb{R}^n$ is closed and $\hat{\xi} : E^c \to \mathbb{R}^n$ is continuously differentiable then $\nu = 0$ if $\mathcal{H}^{n-1}(E) = 0$. (Reduced rank estimators, NRH (2018) *Stat. Prob. Letters.*)
- If $\hat{\xi}$ is a metric projection onto a closed subset of $\mathbb{R}^n$ then ν is a positive measure (NRH & Sokol, arXiv:1402.2997).
- If $\hat{\xi} = \sum_i \hat{\xi}^i 1_{U_i}$ then

$$\nu = \frac{1}{2} \sum_{i \neq j} 1_{\overline{U}_i \cap \overline{U}_j} \langle \hat{\xi}^j - \hat{\xi}^i, \eta_i \rangle \cdot \mathcal{H}^{n-1}$$

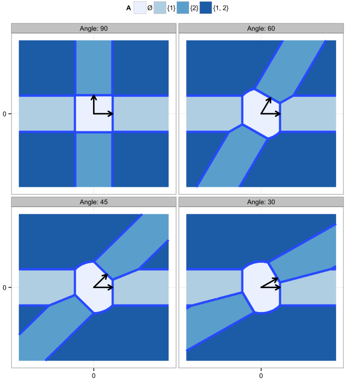
with η_i the outer unit normal to the boundary of U_i.
(Mikkelsen & NRH (2018), *Ann. Inst. H. Poincaré Probab. Statist.*)

Two examples

Lasso-OLS



Best subset selection



http://web.math.ku.dk/~richard/selectionAnimation.html

The correction term

Suppose that $U_i^t = F(t, U_i^0)$ and $\hat{\xi}^{i,t}$ are parametrized by $t \in \mathbb{R}$ and F is a flow.

Example: Lasso gives $U_i^t = e^t U_i^0$ for penalty $\lambda = e^t$.

Theorem (Mikkelsen & NRH, in preparation)

There is a statistic $H(t, Y)$ such that under technical conditions

$$\frac{1}{(2\pi)^{n/2}} \int e^{-\frac{\|y-\xi\|^2}{2}} \mathrm{d}\nu^t(\mathrm{d}y) = \partial_t E(H(t, Y)).$$

Example: Lasso-OLS gives $H(t, Y) = -\nabla \cdot \hat{\xi}^t(Y)$.

Applies to: marginal screening, relaxed lasso, best subset selection, some smoothing-selection algorithms and greedy basis pursuit.

A refined information criterion

We propose

$$\mathrm{IC}(t) = \|Y - \hat{\xi}^t(Y)\|^2 + 2\left(\nabla \cdot \hat{\xi}^t(Y) + \partial_t \mathrm{smooth}(H(t, Y))\right)$$

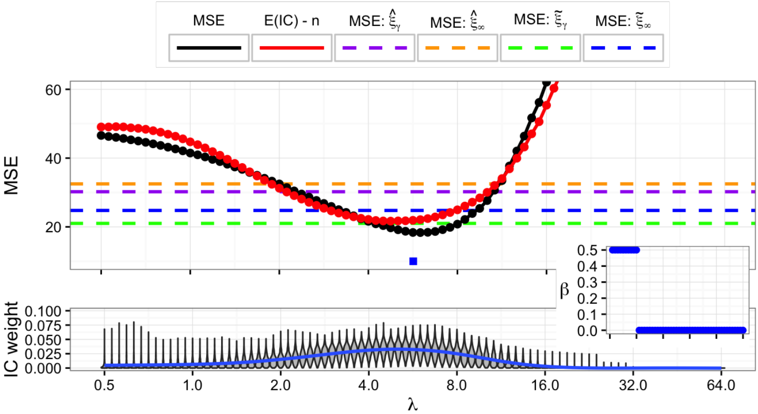
with $\mathrm{smooth}(H(t, Y))$ denoting a t-smoothing of the stochastic jump function $t \mapsto H(t, Y)$.

Forward stepwise variable selection can be recast in flow-form as

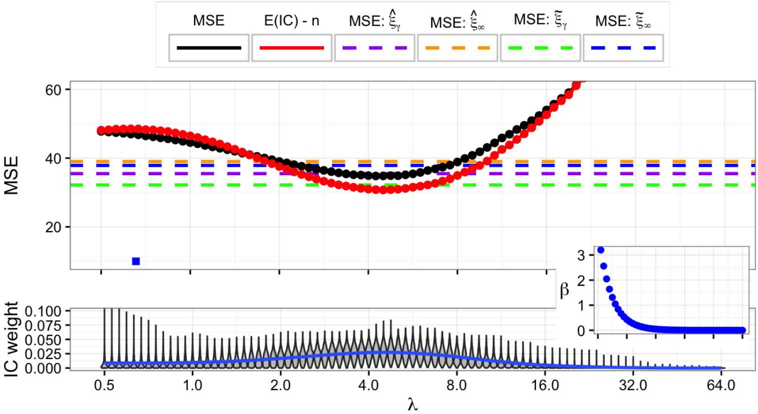
$$d(t) = \underset{d=0,\ldots,p}{\arg\min} ||Y - S_d Y||^2 + \underbrace{e^{2t}}_{\lambda} d$$

with $\nabla \cdot \hat{\xi}^t(Y) = d(t)$.

$n = 100,\ p = 50,\ \gamma = 0.1$



$n = 100,\ p = 50,\ \gamma = 0.1$



$n = 100,\ p = 50,\ \gamma = 0.1$