Post-Selection Distributions, Model Averaging, Bagging, Inference

focustat
FOCUS DRIVEN STATISTICAL INFERENCE WITH COMPLEX DATA

Nils Lid Hjort

Department of Mathematics, University of Oslo

Post-Selection Workshop, Leuven, August 2016

The problem: selection, averaging, post-inference

Typical setup: data $(x_1, y_1), \ldots, (x_n, y_n)$, with different candidate regression models (also: time series, spatial models, survival analysis models, etc.). Focus parameter μ, e.g.

$$\mathrm{E}\,(Y \,|\, x_0) \quad \text{or} \quad \mathrm{Pr}\{Y \geq \text{threshold} \,|\, x_0\} \quad \text{or} \quad F^{-1}(0.99 \,|\, x_0).$$

We select among, or average over, candidate $\widehat{\mu}_S$:

$$\widehat{\mu}^* = \sum_{\text{models } S} c(S \,|\, \text{data}) \widehat{\mu}_S.$$

E.g.

$$c_{\text{aic}}(S \,|\, \text{data}) = \begin{cases} 1 \text{ for winning AIC model,} \\ 0 \text{ for the other models,} \end{cases}$$

or $c_{\text{sm}-\text{fic}}(S \,|\, \text{data}) \propto \exp\{-\lambda\, \text{fic}(S)\}$.

Choice of weights?; distribution of $\widehat{\mu}^*$?; inference?

Plan

Model selection, model averaging, post-selection and post-averaging inference, bagging ...

- A Local large-sample framework
- B Distributions for general model-averaging estimators
- C AIC and FIC (and relatives)
- D Optimal weights (with estimates)
- E The Quiet Scandal of Statistics
- F Bagging
- G Better post-selection and post-averaging confidence intervals
- H Concluding remarks

A: Large-sample framework

Wide model: $f(y, \theta, \gamma)$, of dimension $p + q$.

Narrow model: $f(y, \theta, \gamma_0)$, of dimension p, where γ_0 is a known null value.

Here $\theta = (\theta_1, \ldots, \theta_p)$ is protected, $\gamma = (\gamma_1, \ldots, \gamma_q)$ is open.

Candidate models: for each $S \in \{1, \ldots, q\}$, work with the model where γ_j is free for $j \in S$, but $\gamma_j = \gamma_{0,j}$ for $j \notin S$. Focus parameter: $\mu = \mu(\theta, \gamma)$.

Estimate based on model S: find maximum likelihood estimates $(\widehat{\theta}_S, \widehat{\gamma}_S)$ in model S, and then

$$\widehat{\mu}_S = \mu(\widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c}).$$

These range from

$$\widehat{\mu}_{\rm narr} = \mu(\widehat{\theta}_{\rm narr}, \gamma_0) \quad \text{to} \quad \widehat{\mu}_{\rm wide} = \mu(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide}).$$

For each candidate model S, wish to assess (understand, approximate, estimate)

$$\text{distribution of } \sqrt{n}(\widehat{\mu}_S - \mu)$$

and

$$\text{risk}(S) = \text{mse}_S(\theta, \gamma) = n\,\text{E}\,\{\widehat{\mu}_S - \mu(\theta, \gamma\}^2.$$

Variances are $O(1/n)$, biases are fixed, type $\mu(\theta, \gamma) - \mu(\theta_{\text{l.f.}}, \gamma_0)$. But variances and squared biases become exchangeable currencies in a local large-sample framework where

$$f_{\text{true}}(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}).$$

Here $\delta = \sqrt{n}(\gamma - \gamma_0)$ is relative distance from narrow model to the real model.

May now work out precise limit distributions for $\sqrt{n}(\widehat{\mu}_S - \mu_{\text{true}})$ and so on.

B: Limit distributions and mse approximations

Master Theorem One (next page) gives limit distribution for each $\widehat{\mu}_S$. Some quantities & notation are needed. Let

$$J = \mathrm{Var}\begin{pmatrix} \partial \log f(Y, \theta_0, \gamma_0)/\partial\theta \\ \partial \log f(Y, \theta_0, \gamma_0)/\partial\gamma \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$$

be the Fisher information matrix at (θ_0, γ_0), with inverse

$$J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}, \quad \text{with } Q = J^{11}.$$

Also, let

$$\omega = J_{10}J_{00}^{-1}\frac{\partial\mu}{\partial\theta} - \frac{\partial\mu}{\partial\gamma} \quad \text{and} \quad \tau_0^2 = (\tfrac{\partial\mu}{\partial\theta})^{\mathrm{t}} J_{00}^{-1}\frac{\partial\mu}{\partial\theta},$$

with derivatives at null point. Crucial ingredient:

$$D_n = \sqrt{n}(\widehat{\gamma}_{\mathrm{wide}} - \gamma_0) \to_d D \sim \mathrm{N}_q(\delta, Q).$$

Let finally $Q_S = (\pi_S Q^{-1} \pi_S^{\mathrm{t}})^{-1}$ and $G_S = \pi_S^{\mathrm{t}} Q_S \pi_S Q^{-1}$; these are $q \times q$ matrices with $\mathrm{Tr}(G_S) = |S|$.

Master Theorem One gives orthogonal decomposition:

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_S = \Lambda_0 + \omega^{\rm t}(\delta - G_S D),$$

where $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ is independent of $D \sim {\rm N}_q(\delta, Q)$.

Narrow model: $G_\emptyset = 0$, limit is $\Lambda_0 + \omega^{\rm t}\delta$.

Wide model: $G_{\rm wide} = I$, limit is $\Lambda_0 + \omega^{\rm t}(\delta - D)$.

Narrow better than wide model, for fixed focus parameter: when

$$(\omega^{\rm t}\delta)^2 \le \omega^{\rm t} Q\omega, \quad \text{or} \quad |\omega^{\rm t}(\gamma - \gamma_0)| \le \{\omega^{\rm t} Q\omega\}^{1/2}/\sqrt{n},$$

which is an infinite strip. Narrow better than wide model, for all focus parameters:

$$\delta^{\rm t} Q^{-1}\delta \le 1 \quad \text{or} \quad (\gamma - \gamma_0)^{\rm t} Q^{-1}(\gamma - \gamma_0) \le 1/n.$$

Can do similar analyses for all submodels.

Another Master Lemma says that there is joint convergence in distribution of the 2^q variables $\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true})$ and $D_n = \sqrt{n}(\widehat{\gamma}_{\rm wide} - \gamma_0)$ to the appropriate $(\Lambda_{\rm narr}, \ldots, \Lambda_{\rm wide}, D)$, each element a function of $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ and $D \sim {\rm N}_q(\delta, Q)$.

Consider a model averaging operation

$$\widehat{\mu}^* = \sum_S c(S \,|\, D_n)\widehat{\mu}_S,$$

with weights summing to 1. Master Theorem Two says

$$\sqrt{n}(\widehat{\mu}^* - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D)\},$$

where

$$\widehat{\delta}(D) = \sum_S c(S \,|\, D) G_S D.$$

This holds even when $c(S \,|\, D)$ has a finite number of discontinuities in D (as with AIC, FIC etc.), and with $c_n(S \,|\, D_n) \to_d c(S \,|\, D)$, etc.

Master Theorem Two implies that distribution and performance of general post-selection or model-average estimator $\widehat{\mu}^*$ for μ (rather complicated) – is large-sample equivalent to studying distribution and performance of

$$\omega^{\rm t}\widehat{\delta}(D) = \omega^{\rm t}\sum_S c(S\,|\,D)G_S D \quad \text{for estimating} \quad \omega^{\rm t}\delta,$$

based on $D \sim \mathrm{N}_q(\delta, Q)$ (which is non-trivial, but much simpler).

This amounts to studying risk functions

$$\mathrm{risk}(\delta) = \mathrm{E}\{\omega^{\rm t}\widehat{\delta}(D) - \omega^{\rm t}\delta\}^2.$$

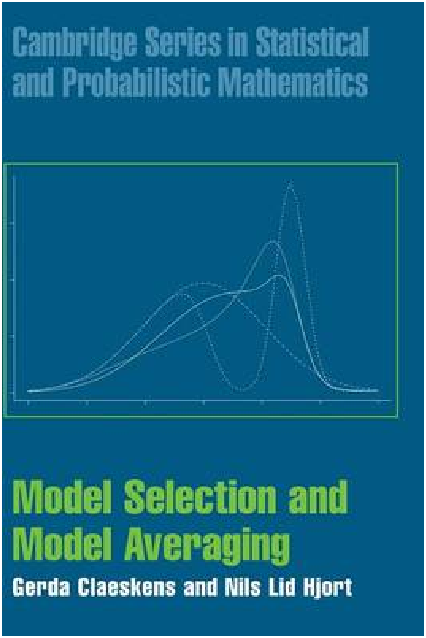
Using narrow model: $\widehat{\delta}(D) = 0$, $\mathrm{risk}_{\rm narr}(\delta) = (\omega^{\rm t}\delta)^2$.

Using wide model: $\widehat{\delta}(D) = D$, $\mathrm{risk}_{\rm wide}(\delta) \equiv \omega^{\rm t}Q\omega$.

Using FIC is typically better than AIC: $\mathrm{risk}_{\rm fic}(\delta) < \mathrm{risk}_{\rm aic}(\delta)$ in big parts of the parameter space.

Two-way bridge: Finite-n-model problems $\longleftrightarrow$ Limit Experiment.

The distributions of post-selection and model-average estimators are captured by $\sum_S c(S \mid D) G_S D$, are typically very non-linear mixtures of normals, and hence not normal.



Choosing weights, when averaging over models:
Suppose $q = 3$ with diagonal Q. General model averaging estimator, weighting across 8 models:

$$\begin{aligned}\widehat{\mu}^* = {} & c_{000}\widehat{\mu}_{\mathrm{narr}} + c_{100}\widehat{\mu}_{100} + c_{010}\widehat{\mu}_{010} + c_{001}\widehat{\mu}_{001} \\ & + c_{110}\widehat{\mu}_{110} + c_{101}\widehat{\mu}_{101} + c_{011}\widehat{\mu}_{001} + c_{111}\widehat{\mu}_{111},\end{aligned}$$

where weights $c_{i,j,k} = c_{i,j,k}(D_n)$ may depend on data. Then matters are determined by

$$\begin{aligned}\widehat{\delta}(D) &= \sum_{8\,\mathrm{models}} c_{i,j,k}(D) G_{i,j,k} D \\ &= \begin{pmatrix} \{c_{100}(D) + c_{110}(D) + c_{101}(D) + c_{111}(D)\} D_1 \\ \{c_{010}(D) + c_{110}(D) + c_{011}(D) + c_{111}(D)\} D_2 \\ \{c_{001}(D) + c_{011}(D) + c_{101}(D) + c_{111}(D)\} D_3 \end{pmatrix}\end{aligned}$$

and

$$\mathrm{E}\{\omega^{\mathrm{t}}\delta - \omega^{\mathrm{t}}\widehat{\delta}(D)\}^2.$$

There's overrepresentation – we learn that many different model average operations are equivalent. We don't need 7 weights here, only 3, averaging over the 3 singletons.

C: AIC, FIC, and relatives

From Master Theorem One: limit risk of $n\,\mathrm{E}(\widehat\mu_S - \mu_{\rm true})^2$ when using model S is

$$\begin{aligned}\mathrm{mse}(S) &= \mathrm{E}\{\Lambda_0 + \omega^{\rm t}(\delta - G_S D)\}^2 \\ &= \tau_0^2 + \omega^{\rm t} G_S Q G_S^{\rm t}\omega + \omega^{\rm t}(I - G_S)^{\rm t}\delta\delta^{\rm t}(I - G_S)\omega.\end{aligned}$$

In the Limit Experiment, all quantities are known apart from δ, for which we have $D \sim \mathrm{N}_q(\delta, Q)$.

Since $DD^{\rm t}$ estimates $\delta\delta^{\rm t} + Q$, we use

$$\mathrm{fic}(S) = \tau_0^2 + \omega^{\rm t} G_S Q G_S^{\rm t}\omega + \max\{\omega^{\rm t}(I - G_S)^{\rm t}(DD^{\rm t} - Q)(I - G_S)\omega, 0\}.$$

For real data (and finite n), we insert consistent estimators for τ_0, ω, G_S, Q, and $D_n = \sqrt{n}(\widehat\gamma_{\rm wide} - \gamma_0)$ for D.

Large-sample analysis of AIC: with $\mathrm{aic}_{n,S} = 2\ell_{n,\max,S} - 2(p + |S|)$, we have

$$\mathrm{aic}_{n,S} - \mathrm{aic}_{n,\emptyset} \to_d \mathrm{aic}(S, D) = D^{\mathrm{t}} Q^{-1} \pi_S^{\mathrm{t}} Q_S \pi_S Q^{-1} D - 2|S|.$$

Via $D \sim \mathrm{N}_q(\delta, Q)$, this gives clear limits for

$$\mathrm{Pr}\{\text{AIC selects } S\} \to \mathrm{Pr}_\delta\{\mathrm{aic}(S, D) > \text{all other } \mathrm{aic}(S', D)\}.$$

Can compare these probabilities with

$$\mathrm{Pr}\{\text{FIC selects } S\} \to \mathrm{Pr}_\delta\{\mathrm{fic}(S, D) < \text{all other } \mathrm{fic}(S', D)\}.$$

Typically (but not uniformly), FIC has a bigger chance of finding $S_{\mathrm{opt}}(\delta)$, the model where $\mathrm{mse}(S, \delta)$ is smallest.

Also, $\widehat{\mu}_{\mathrm{fic,final}} = \widehat{\mu}_{S_{\mathrm{fic}}}$ is typically (but not uniformly) better than $\widehat{\mu}_{\mathrm{aic,final}} = \widehat{\mu}_{S_{\mathrm{aic}}}$:

$$\mathrm{E}\,\{\omega^{\mathrm{t}}(\delta - G_{\widehat{S},\mathrm{fic}} D)\}^2 < \mathrm{E}\,\{\omega^{\mathrm{t}}(\delta - G_{\widehat{S},\mathrm{aic}} D)\}^2 \quad \text{for big space of } \delta.$$

FIC is set up to work for one given focus parameter at the time.

May generalise to AFIC, average-weighted FIC, when we have a list of foci, $\{\mu(u)\colon u \in \mathcal{U}\}$, along with importance function $w(u)$:

$$\mathrm{risk}_n(S) = n\,\mathrm{E}\sum_{u\in\mathcal{U}} w(u)\{\widehat{\mu}_S(u) - \mu(u)\}^2.$$

Details in Claeskens and Hjort (2008a, 2008b).

Message: AIC is (approximately) same as AFIC, when we're equally interested in everything.

For regression models, $f(y_i \,|\, x_i, z_i)$, use AFIC for $\mathrm{E}(Y_i \,|\, x_i, z_i)$, with same weight of importance for all (x_i, z_i): then we're back to AIC.

Model selection is an un-smooth operation – and is inadmissible in the decision theoretic sense. Complete class theorems (1950ies to 1970ies) $\Longrightarrow$ all admissible estimators $\widehat{\mu}$ must be Bayes or generalised Bayes:

$$\widetilde{\delta}(D) = \mathrm{E}(\delta \,|\, D) = \frac{\int \delta\phi(\delta - D)\,\mathrm{d}\pi(\delta)}{\int \phi(\delta - D)\,\mathrm{d}\pi(\delta)} \quad \text{for some } \mathrm{d}\pi(\delta). \qquad (*)$$

Prototype example: $y_1, \ldots, y_n$ are i.i.d. $\mathrm{N}(\mu, 1)$. Model 0: $\mu = 0$. Model 1: $\mu \in \mathbb{R}$. Then

$$\widehat{\mu}_{\mathrm{aic}} = \bar{y}\, I\{|\sqrt{n}\bar{y}| \geq \sqrt{2}\} = \begin{cases} \bar{y} & \text{if } |\sqrt{n}\bar{y}| \geq \sqrt{2}, \\ 0 & \text{if } |\sqrt{n}\bar{y}| < \sqrt{2}. \end{cases}$$

This is an ok estimator of μ – but can be uniformly improved upon.

I translate the situation to canonical form: with $\mu = \delta/\sqrt{n}$ and $D = \sqrt{n}\bar{y} \sim \mathrm{N}(\delta, 1)$,

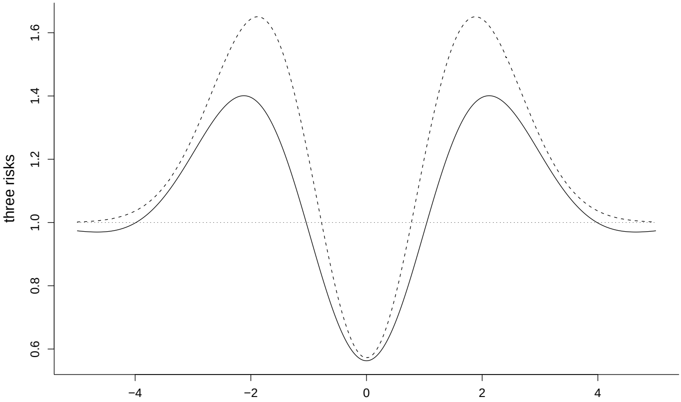
$$\sqrt{n}(\widehat{\mu}_{\mathrm{aic}} - \mu) = \widehat{\delta}_{\mathrm{aic}}(D) - \delta = D\, I\{|D| \geq \sqrt{2}\} - \delta.$$

I shall exhibit a generalised Bayes estimator $(*)$ which is uniformly better than $\widehat{\delta}_{\mathrm{aic}}(D)$.

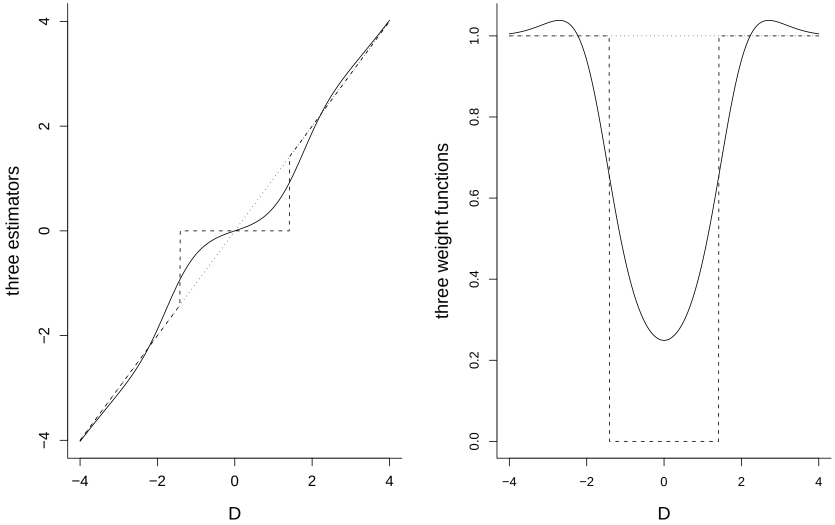
I work with the generalised prior having point mass k at zero and otherwise uniform on $(-\infty, u) \cup (u, \infty)$. The generalised Bayes estimator takes the form

$$\widetilde{\delta}(D) = D + \frac{-kD\phi(D) + \phi(D-u) - \phi(D+u)}{k\phi(D) + 1 + \Phi(D+u) - \Phi(D-u)}.$$

With $k = 4.4$ and $u = 1.6$, the AIC is uniformly beaten:



With $D \sim \mathrm{N}(\delta, 1)$, three estimators, $\widehat{\delta}(D) = c(D)D$: AIC is beaten; it pays to smooth.



D: Optimal weights (and estimates thereof)

Candidate models $M_1, \ldots, M_k$, estimators $\widehat\mu_1, \ldots, \widehat\mu_k$ for focus parameter μ: Which weights should be used?

Suppose $\mathrm{E}\,\widehat\mu_j = \mu + b_j$, and variance matrix Σ. The linear combination $\widehat\mu^* = c^{\mathrm{t}}\widehat\mu$, with $\sum_{j=1}^m c_j = 1$, has

$$\mathrm{E}(\widehat\mu^* - \mu)^2 = c^{\mathrm{t}}\Sigma c + (c^{\mathrm{t}}b)^2 = c^{\mathrm{t}}(\Sigma + bb^{\mathrm{t}})c.$$

This is minimised by $\widehat\mu^* = (c^*)^{\mathrm{t}}\widehat\mu$, with

$$\widehat\mu^* = \frac{\mathbf{1}^{\mathrm{t}}(\Sigma + bb^{\mathrm{t}})^{-1}\widehat\mu}{\mathbf{1}^{\mathrm{t}}(\Sigma + bb^{\mathrm{t}})^{-1}\mathbf{1}}, \quad \text{where } c^* \propto (\Sigma + bb^{\mathrm{t}})^{-1}\mathbf{1}.$$

In our setup, with

$$\sqrt{n}(\widehat\mu_S - \mu_{\mathrm{true}}) \to_d \Lambda_0 + \omega^{\mathrm{t}}(\delta - G_S D),$$

can read off biases, variances, covariances, for any set of candidate S models.

For candidate models $S_0, S_1, \ldots, S_m$, where S_0 is narrow model, consider $\widehat{\mu}^* = \sum_{j=0}^m c_j \widehat{\mu}_j$. Limiting risk is

$$r(\delta, c) = \tau_0^2 + c^{\rm t}(\Sigma + bb^{\rm t})c,$$

with biases $b_j = \omega^{\rm t}(I - G_j)\delta$, and Σ with elements $\omega^{\rm t} G_j Q G_k^{\rm t} \omega$.

Minimising the risk: let Σ_{11} $(m \times m)$ have elements $\omega^{\rm t} G_j Q G_k^{\rm t} \omega$, and z $(m \times 1)$ have $\omega^{\rm t} G_j \delta$. Then

$$c_0^* = 1 - \omega^{\rm t}\delta \frac{\mathbf{1}^{\rm t}\Sigma_{11}^{-1} z}{1 + z^{\rm t}\Sigma_{11}^{-1} z}, \quad \begin{pmatrix} c_1^* \\ \vdots \\ c_m^* \end{pmatrix} = \omega^{\rm t}\delta \frac{\Sigma_{11}^{-1} z}{1 + z^{\rm t}\Sigma_{11}^{-1} z}.$$

The weights contain various $\delta\delta^{\rm t}$ terms, and data information is $D_n \to_d D \sim \mathrm{N}_q(\delta, Q)$. Choices include: (i) inserting D_n for δ; (ii) inserting $D_n D_n^{\rm t} - \widehat{Q}$ for δ, with truncation; (iii) estimating all relevant terms in $r(\delta, c)$ and then minimising.

Example 1: weighting between narrow and wide only,

$$\widehat{\mu}^* = (1-c)\widehat{\mu}_{\rm narr} + c\widehat{\mu}_{\rm wide}.$$

Optimal (oracle) weight:

$$c^* = \frac{(\omega^{\rm t}\delta)^2}{(\omega^{\rm t}\delta)^2 + \omega^{\rm t}Q\omega}.$$

May use

$$\widehat{c}^* = \frac{(\omega^{\rm t}D)^2}{(\omega^{\rm t}D)^2 + \omega^{\rm t}Q\omega}$$

or

$$\begin{aligned}\widehat{c}^* &= \frac{\max\{(\omega^{\rm t}D)^2 - \omega^{\rm t}Q\omega, 0\}}{\max\{(\omega^{\rm t}D)^2 - \omega^{\rm t}Q\omega, 0\} + \omega^{\rm t}Q\omega} \\ &= \begin{cases} 0 & \text{if } (\omega^{\rm t}D)^2 \le \omega^{\rm t}Q\omega, \\ 1 - (\omega^{\rm t}Q\omega)/(\omega^{\rm t}D)^2 & \text{if } (\omega^{\rm t}D)^2 > \omega^{\rm t}Q\omega. \end{cases}\end{aligned}$$

Example 2: weighting across the singletons,

$$\widehat{\mu}^* = c_0\widehat{\mu}_{\rm narr} + \sum_{j=1}^{q} c_j\widehat{\mu}_j,$$

where $\widehat{\mu}_j$ is from the model having only γ_j on board.

Optimal weights are readily found. For the case of Q diagonal $(\kappa_1^2, \ldots, \kappa_q^2)$:

$$c_j^* = \omega^{\rm t}\delta \frac{\delta_j/(\omega_j\kappa_j^2)}{1 + \sum_{j'=1}^{q} \delta_{j'}^2/\kappa_{j'}^2}.$$

At least three natural choices for estimating these.

Example 3: Can find the best weight for

$$\widehat{\mu}^* = (1-\widehat{\rho})\widehat{\mu}_{\rm narr} + \widehat{\rho}\frac{1}{q}\sum_{j=1}^{q} \widehat{\mu}_{{\rm singleton}\ j}.$$

Also: various empirical Bayes like model averaging procedures.

E: The Quiet Scandal of Statistics

Given that data follow a model M, one may typically construct a confidence interval

$$\mathrm{Pr}\{\mathrm{low}(M) \le \mu \le \mathrm{up}(M) \,|\, \mathrm{model}\ M\ \mathrm{holds}\} \doteq 0.95.$$

Typically,

$$[\mathrm{low}(M), \mathrm{up}(M)] = \widehat{\mu}_M \pm 1.96\widehat{\tau}_M/\sqrt{n},$$

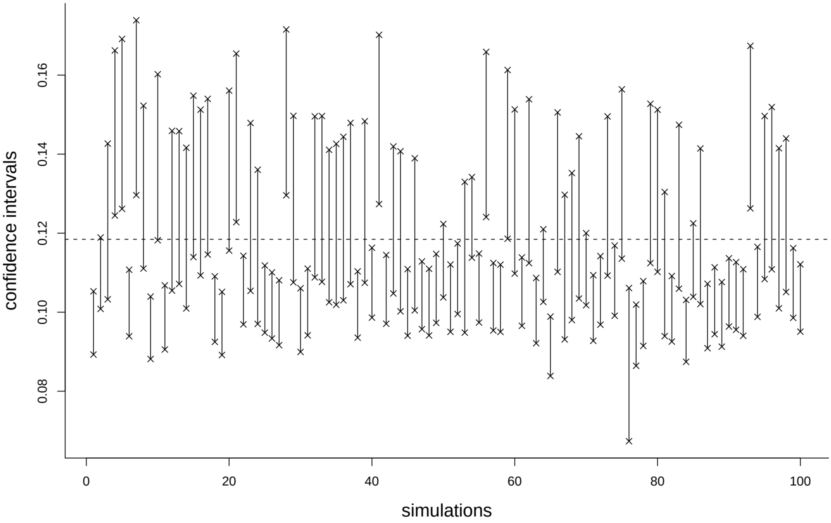
or something first-order equivalent – and in our framework, $\tau_M = (\tau_0^2 + \omega^{\mathrm{t}} G_M Q G_M^{\mathrm{t}} \omega)^{1/2}$.

This is 'textbook material' (and 'textbook modus').

Suppose model M has been selected among various competitors, using AIC or FIC or BIC – then reporting $[\mathrm{low}(M), \mathrm{up}(M)]$ is too simplistic and overly optimistic. (1) The model might still have a bias; (2) the initial model selection phase is ignored.

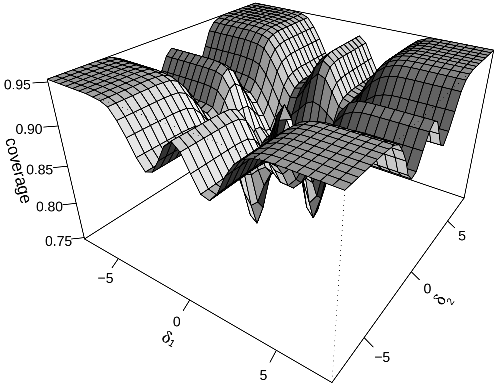
This hiding (or ignoring, or forgetting) uncertainty is called the Quiet Scandal of Statistics (Leo Breiman).

Illustration: 100 intended 90% confidence intervals for $\mu = F^{-1}(0.10)$; M_1 exponential, M_2 Weibull; truth = a little bit away from M_1; method = AIC. Half of the intervals are ok; the other half too short and also biased.



May use Master Theorems One and Two to understand and assess the degree of overoptimism. How much smaller than 0.95 is $\Pr\{\mathrm{low}(\widehat{M}) \le \mu \le \mathrm{up}(\widehat{M})\}$, when $\widehat{M}$ is selected by e.g. AIC?

$$\Pr\big[a \le \sqrt{n}\{\widehat{\mu}(\widehat{S}) - \mu_{\mathrm{true}}\}/\widehat{\tau}(\widehat{M}) \le b\big] \to \mathrm{clear\ limit}(\delta).$$



The Smaller Scandal of Statistics: a clever statistician works out clever weights for model averaging,

$$\widehat{\mu}^* = \sum_S c(S \mid D_n)\widehat{\mu}_S,$$

but does the rest of the analysis pretending (i.e. ignoring the difficulties) that the weights are nonrandom.

But distributions (and limit distributions) of

$$\sqrt{n}\Big\{\sum_S c(S)\widehat{\mu}_S - \mu_{\rm true}\Big\} \quad \text{and} \quad \sqrt{n}\Big\{\sum_S \widehat{c}(S)\widehat{\mu}_S - \mu_{\rm true}\Big\}$$

are very different, particularly in parts of the parameter space where models are bumping into each other.

Clear theory for both covered by Master Theorem Two. Sometimes cleverness doesn't pay off – the variability in $\widehat{c}(S)$ might mess up the benefits of hunting for clever weights.

F: Bagging

Suppose $\widehat{\mu}$ is some post-selection or model-average (or otherwise complicated) estimator of μ:

$$\widehat{\mu} = \sum_S c(S \mid D_n)\widehat{\mu}_S.$$

An alternative to $\widehat{\mu}$ is its bagging version, or averaging over bootstraps. Bootstrapped data, say $(x_1, y_1^*), \ldots, (x_n, y_n^*)$ with the y_i^* sampled from $f(y_i \mid x_i, \widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$, yield

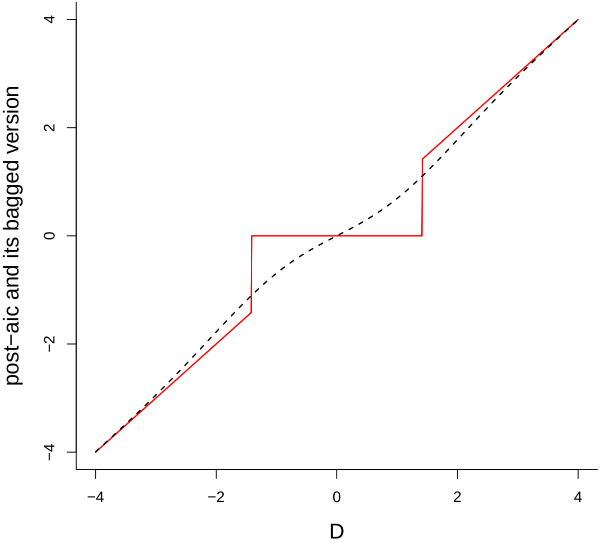
$$\widehat{\mu}^* = \sum_S c(S \mid D_n^*)\widehat{\mu}_S^*.$$

I do this $B = 1000$ times:

$$\widehat{\mu}_{\rm bagg} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_b^*.$$

This smooths out the sharp decisions of model selection etc.

Prototype situation: $y_1, \ldots, y_n$ are i.i.d. $\mathrm{N}(\mu, 1)$, AIC gives $\widehat{\mu} = \bar{y}\, I\{|\sqrt{n}\bar{y}| \geq \sqrt{2}\}$, equivalent to using $\widehat{\delta}(D) = D\, I\{|D| \geq \sqrt{2}\}$ when $D \sim \mathrm{N}(\delta, 1)$. Bagging smooths out:
$\widehat{\delta}_{\mathrm{bagg}}(D) = \mathrm{E}_* D^*\, I\{|D^*| \geq \sqrt{2}\}$, with $D^* \sim \mathrm{N}(D, 1)$.



Recall from Master Theorem Two that

$$\sqrt{n}(\widehat{\mu} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D)\}$$

with $\widehat{\delta}(D) = \sum_S c(S \mid D) G_S D$. We have

$$\begin{aligned}\sqrt{n}(\widehat{\mu}_{\rm bagg} - \mu_{\rm true}) &= \frac{1}{B}\sum_{b=1}^{B} \sqrt{n}(\widehat{\mu}^*_b - \mu_{\rm true}) \\ &\doteq_d \frac{1}{B}\sum_{b=1}^{B} \big[\Lambda^*_{0,b} + \omega^{\rm t}\{\delta - \widehat{\delta}(D^*_b)\}\big],\end{aligned}$$

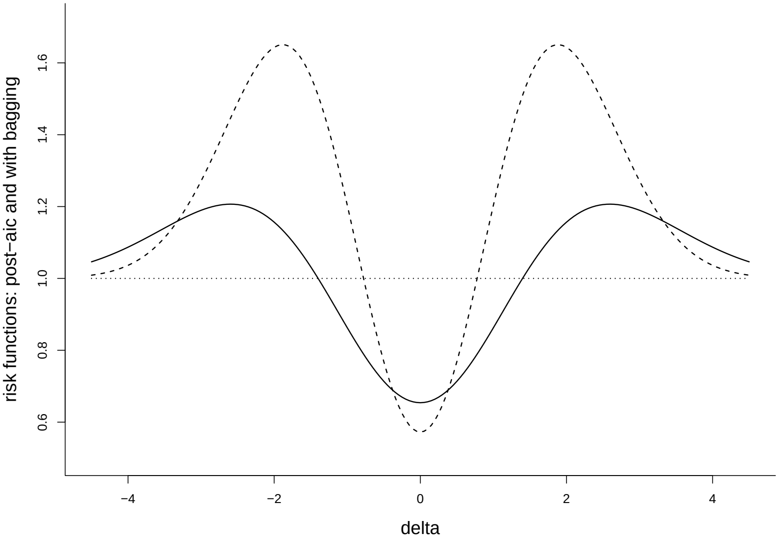
where $\Lambda^*_{0,b} \sim {\rm N}(0, \tau_0^2)$ and $D^*_b \sim {\rm N}_q(D_n, Q)$. Limit operation gives Master Theorem Three:

$$\sqrt{n}(\widehat{\mu}_{\rm bagg} - \mu_{\rm true}) \to_d {\rm E}_{\rm boot}\big[\Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D^*)\}\big] = \Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}_{\rm bagg}(D)\},$$

with

$$\widehat{\delta}_{\rm bagg}(D) = {\rm E}_{\rm boot}\{\widehat{\delta}(D^*) \mid D\} = \int \widehat{\delta}(x)\phi(x - D, Q)\,{\rm d}x.$$

Bagging the post-selection methods lowers max-risk, without losing much close to the narrow model:



Even very complicated procedures (post-selection, model-averaging, etc.) can be bagged. Can also attempt double-bagging (but in some simple cases I've been through it doesn't help much).

Bridging from start-method to bagged-method, via shrunken bags: For any $\rho \in [0,1]$, I can construct a method

$$\widehat{\mu}_{\rm shrunken\ bag} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mu}({\rm data}_b^*)$$

corresponding to $(D^* \mid D) \sim {\rm N}_q(D, \rho Q)$, and

$$\widehat{\delta}_{\rho\text{-bagg}} = {\rm E}_{\rho\text{-bagg}}\{\widehat{\delta}(D^*) \mid D\} = \int \widehat{\delta}(x)\phi(x - D, \rho Q)\,{\rm d}x.$$

For $\rho = 0$: the start-method itself, no additional smoothing.
For $\rho = 1$: (usual) bagging.
For $\rho = 0.5$: a shrunken bag.

G: Post-selection and post-averaging inference

Consider any post-selection or post-averaging estimator

$$\widehat{\mu} = \sum_S c_n(S \mid D_n)\widehat{\mu}_S.$$

Wish to construct $[\mathrm{low}, \mathrm{up}]$ such that

$$\Pr\{\mathrm{low} \le \mu_{\mathrm{true}} \le \mathrm{up}\} \doteq 0.90.$$

This is a tall order, as the distribution of $\widehat{\mu}$ is (very) complicated, depending also on smaller local variations in the parameter space.

Attempts at constructing approximate pivots

$$T_n = \sqrt{n}(\widehat{\mu} - \mu_{\mathrm{true}})/\widehat{\kappa}$$

do not quite succeed:

$$T_n \to_d \frac{\Lambda_0 + \omega^{\mathrm{t}}\{\delta - \widehat{\delta}(D)\}}{\kappa(D, \delta)}.$$

This is fine, the distribution is precise (and can be precisely simulated), for any given δ – but hard to use.

We have

$$\sqrt{n}(\widehat{\mu}^* - \mu_{\rm true}) \to_d \Lambda(\delta) = \Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D)\}$$

and may simulate this limit distribution for each given δ:

$$\Pr_\delta\{{\rm low}(\delta) \le \Lambda(\delta) \le {\rm up}(\delta)\} = 0.95.$$

So the 'oracle interval' is

$$CI = [\widehat{\mu}^* - {\rm up}(\delta)/\sqrt{n}, \widehat{\mu}^* - {\rm low}(\delta)/\sqrt{n}].$$

A simple attempt: Insert estimate D for δ:

$$CI = [\widehat{\mu}^* - {\rm up}(D)/\sqrt{n}, \widehat{\mu}^* - {\rm low}(D)/\sqrt{n}].$$

But this typically doesn't work, and coverage is off.

Safer (yields guaranteed conservative 0.90 intervals for each parameter we're interested in): consider

$$E_n = \{\delta\colon (\delta - D_n)^{\rm t}\widehat{Q}^{-1}(\delta - D_n) = n(\gamma - \widehat{\gamma}_{\rm wide})^{\rm t}\widehat{Q}^{-1}(\gamma - \widehat{\gamma}_{\rm wide}) \le z_{q,0.95}\}.$$

Then $\mathrm{Pr}_\delta\{\delta \in E_n\} \to 0.95$. Also, from

$$\mathrm{Pr}_\delta\{\mathrm{low}(\delta) \le \Lambda(\delta) \le \mathrm{up}(\delta)\} = 0.95.$$

form the wider

$$\begin{aligned} \mathrm{low}^* &= \min\{\mathrm{low}(\delta)\colon \delta \in E_n\}, \\ \mathrm{up}^* &= \max\{\mathrm{up}(\delta)\colon \delta \in E_n\}, \end{aligned}$$

To be safe, we need to pass from oracle intervals

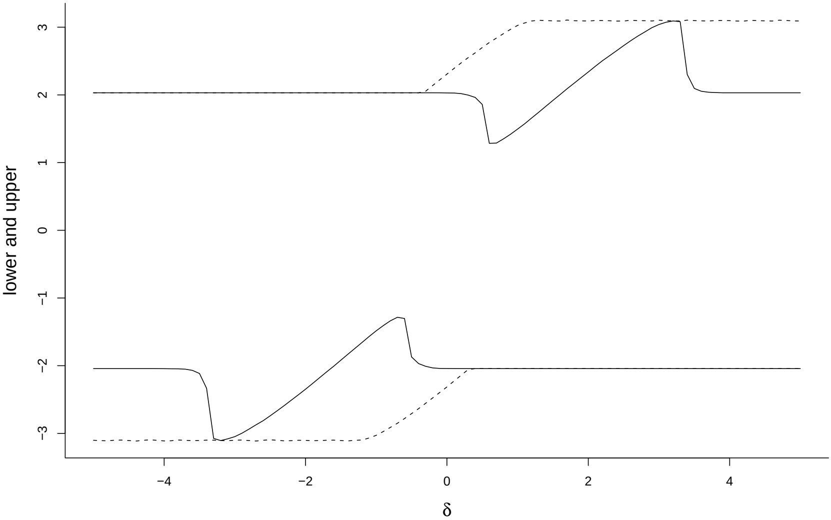
$$CI = [\widehat{\mu}^* - \mathrm{up}(\delta)/\sqrt{n}, \widehat{\mu}^* - \mathrm{low}(\delta)/\sqrt{n}]$$

to the wider

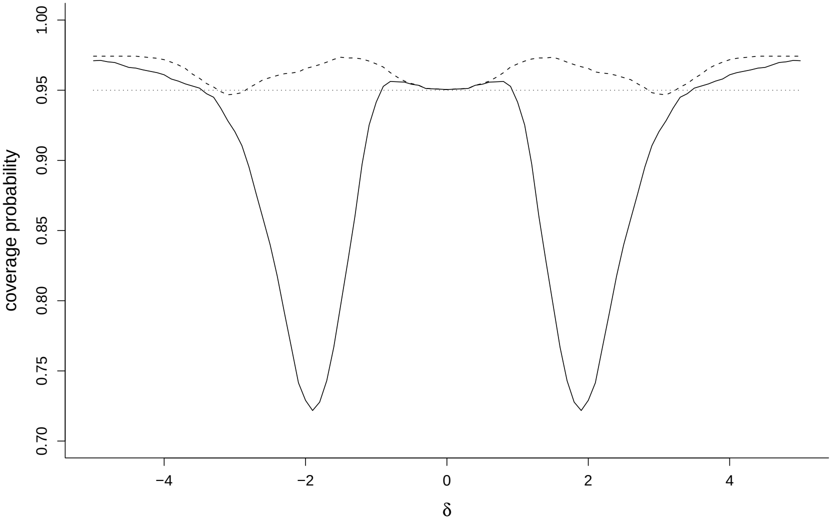
$$CI^* = [\widehat{\mu}^* - \mathrm{up}^*/\sqrt{n}, \widehat{\mu}^* - \mathrm{low}^*/\sqrt{n}].$$

Here $\mathrm{Pr}_\delta\{\mu_{\rm true} \in CI^*\} \to p(\delta) \ge 0.90$.

Illustration, one-dimensional case: $\mathrm{low}(\delta), \mathrm{up}(\delta)$, along with $\mathrm{low}^*(D), \mathrm{up}^*(D)$.



Checking coverage, of simple method (via oracle interval, and plug-in of D_n for δ), and of safer method:



H: Concluding remarks

1. Model selection: used in (at least) two conceptually different ways. To explain or to predict? Answer: it depends (on context, situation, problem, end users).

Often, selection and averaging are used to form $\widehat\mu^* = \sum_S c(S \,|\, \mathrm{data})\widehat\mu_S$, or prediction – it's a (clever) black box.

Sometimes, interpretation of final model is more important.

Scylla & Charybdis: if you insist on

$$\Pr\{\widehat S = S_{\rm true}\} \to 1 \quad \text{as } n \text{ grows,}$$

then

$$\mathrm{risk}(\theta,\gamma) = \mathrm{E}_{\theta,\gamma}\{\widehat\mu_{\rm final} - \mu(\theta,\gamma)\}^2$$

will typically be much worse (in parts of the parameter space).

2. Other loss and risk functions (and other Focused Information Criteria): We may use

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - G_S D)$$

to assess and estimate other risk functions

$${\rm risk}_S(\delta) = {\rm E}\, L\big(\Lambda_0 + \omega^{\rm t}(\delta - G_S D)\big)$$

than for squared error loss $L(z) = z^2$.

In particular, clear FIC formulae available for linex loss

$$L_a(z) = \{\exp(az) - 1 - az\}/a^2,$$

and for hit-or-miss loss

$$L(z) = \begin{cases} 1 & \text{if } |z| \ge \varepsilon, \\ 0 & \text{if } |z| < \varepsilon. \end{cases}$$

We wish to select model S with highest hit probability

$$p_n(S) = \Pr\{\sqrt{n}|\widehat{\mu}_S - \mu_{\rm true}| \le \varepsilon\}.$$

Here

$$p_n(S) \to p(S,\delta) = \Pr\{|\Lambda_0 + \omega^{\rm t}(\delta - G_S D)| \le \varepsilon\}.$$

and high $p(S,\delta)$ is seen to be the same as small

$$\lambda(S,\delta) = \log(\tau_0^2 + \omega^{\rm t} G_S Q G_S^{\rm t}\omega) + \frac{\omega^{\rm t}(I - G_S)\delta\delta^{\rm t}(I - G_S)^{\rm t}\omega}{\tau_0^2 + \omega^{\rm t} G_S Q G_S^{\rm t}\omega}.$$

This leads to hit-FIC formulae, upon using $DD^{\rm t} - Q$ for $\delta\delta^{\rm t}$ (with truncation), etc.

3. New-FIC: Suppose $y_1, \ldots, y_n$ are i.i.d. from distribution G. Focus parameter $\mu = \mu(G)$. Consider $k+1$ competing models – parametric models $1, \ldots, k$ and the nonparametric $\widehat\mu_{\rm nonpara} = \mu(G_n)$. May work with biases and variances of $\widehat\mu_{\rm para}$ to form ${\rm fic}_{\rm para}$ and ${\rm fic}_{\rm nonpara}$, without $O(1/\sqrt{n})$ framework.

Doable also for e.g. regression: should I use $a+bx$, or $a+bx+cx^2$, or a nonparametric smoother $\widehat m(x)$, for estimating ${\rm E}(Y\,|\,x)$, on a given interval $[x_{\rm low}, x_{\rm up}]$? But it's a bit messy, requiring bandwidths for different purposes, etc.

See Jullum and Hjort (Sinica, 2016), Hermansen, Hjort, Jullum (2016, for time series).

4. FIC selection and averaging with 3^q choices: Aalen's linear hazard regression model,

$$h_i(s) = x_{i,1}\alpha_1(s) + \cdots + x_{i,q}\alpha_q(s) \quad \text{for } i = 1, \ldots, n.$$

Choose, for each covariate, between (i) zero, (ii) constant, (iii) nonparametric.

FIC plot for the PBC data set: all $3^6 = 729$ estimates of cumulative hazard rate at time = 1 yr, for a 70 yr old high-risk man.