Conditional Predictive Inference Post Model Selection

Hannes Leeb

Department of Statistics
Yale University

Model Selection Workshop, Vienna, July 25, 2008

Problem statment

Predictive inference post model selection in setting with large dimension and (comparatively) small sample size.

Example: Stenbakken & Souders (1987, 1991): Predict performance of D/A converters. Select 64 explanatory variables from a total of 8,192 based on a sample of size 88.

Features of this example:

- Large number of candidate models
- Selected model is complex in relation to sample size
- Focus on predictive performance and inference, not on correctness
- Model is selected and fitted to the data once and then used repeatedly for prediction

Problem statment

Predictive inference post model selection in setting with large dimension and (comparatively) small sample size.

Example: Stenbakken & Souders (1987, 1991): Predict performance of D/A converters. Select 64 explanatory variables from a total of 8,192 based on a sample of size 88.

Features of this example:

- Large number of candidate models
- Selected model is complex in relation to sample size
- Focus on predictive performance and inference, not on correctness
- Model is selected and fitted to the data once and then used repeatedly for prediction

Problem statment

Predictive inference post model selection in setting with large dimension and (comparatively) small sample size.

Problem studied here:

Given a training sample of size n and a collection $\mathcal{M}$ of candidate models, find a 'good' model $m \in \mathcal{M}$ and conduct predictive inference based on selected model, conditional on the training sample. Features:

- $\#\mathcal{M} \gg n$, i.e., potentially many candidate models
- $|m| \sim n$, i.e., potentially complex candidate models
- no strong regularity conditions

Overview of results

We consider a model selector and a prediction interval post model selection (that are based on a variant of generalized cross-validation) in linear regression with random design.

For Gaussian data we show:

The prediction interval is 'approximately valid and short' conditional on the training sample, except on an event whose probability is less than

$$C_1 \ \#\mathcal{M} \ \ \exp\left[-C_2(n-|\mathcal{M}|)\right],$$

where $\#\mathcal{M}$ denotes the number of candidate models, and $|\mathcal{M}|$ denotes the number of parameters in the most complex candidate model.
This finite-sample result holds uniformly over all data-generating processes that we consider.

Overview of results

We consider a model selector and a prediction interval post model selection (that are based on a variant of generalized cross-validation) in linear regression with random design.

For Gaussian data we show:

The prediction interval is 'approximately valid and short' conditional on the training sample, except on an event whose probability is less than

$$C_1 \;\; \#\mathcal{M} \;\; \exp\left[-C_2(n - |\mathcal{M}|)\right],$$

where $\#\mathcal{M}$ denotes the number of candidate models, and $|\mathcal{M}|$ denotes the number of parameters in the most complex candidate model.

This finite-sample result holds uniformly over all data-generating processes that we consider.

Overview of results

We consider a model selector and a prediction interval post model selection (that are based on a variant of generalized cross-validation) in linear regression with random design.

For Gaussian data we show:

The prediction interval is 'approximately valid and short' conditional on the training sample, except on an event whose probability is less than

$$C_1 \ \#\mathcal{M} \ \exp\left[-C_2(n - |\mathcal{M}|)\right],$$

where $\#\mathcal{M}$ denotes the number of candidate models, and $|\mathcal{M}|$ denotes the number of parameters in the most complex candidate model.
This finite-sample result holds uniformly over all data-generating processes that we consider.

The data-generating process

Gaussian linear model with random design

Consider a response y that is related to a (possibly infinite) number of explanatory variables x_j, $j \geq 1$, by

$$y = \sum_{j=1}^{\infty} x_j \theta_j + u \qquad (1)$$

with $x_1 = 1$. Assume that u has mean zero and is uncorrelated with the x_j's. Moreover, assume that the x_j's for $j > 1$ and u are jointly non-degenerate Gaussian, such that the sum converges in L_2.

The data-generating process

Gaussian linear model with random design

Consider a response y that is related to a (possibly infinite) number of explanatory variables x_j, $j \geq 1$, by

$$y \quad = \quad \sum_{j=1}^{\infty} x_j \theta_j + u \tag{1}$$

with $x_1 = 1$. Assume that u has mean zero and is uncorrelated with the x_j's. Moreover, assume that the x_j's for $j > 1$ and u are jointly non-degenerate Gaussian, such that the sum converges in L_2.

The unknown parameters here are θ, the variance of u, as well as the means and the variance/covariance structure of the x_j's.

The data-generating process

Gaussian linear model with random design

Consider a response y that is related to a (possibly infinite) number of explanatory variables x_j, $j \geq 1$, by

$$y \quad = \quad \sum_{j=1}^{\infty} x_j \theta_j + u \tag{1}$$

with $x_1 = 1$. Assume that u has mean zero and is uncorrelated with the x_j's. Moreover, assume that the x_j's for $j > 1$ and u are jointly non-degenerate Gaussian, such that the sum converges in L_2.

No further regularity conditions are imposed.

The candidate models and predictors

The candidate models and predictors

Consider a sample (X, Y) of n independent realizations of (x, y) as in (1), and a collection $\mathcal{M}$ of candidate models. Each model $m \in \mathcal{M}$ is assumed to satisfy $|m| < n - 1$. Each model m is fit to the data by least-squares. Given a new set of explanatory variables $x^{(f)}$, the corresponding response $y^{(f)}$ is predicted by

$$\hat{y}^{(f)}(m) \quad = \quad \sum_{j=1}^{\infty} x_j^{(f)} \tilde{\theta}_j(m)$$

when using model m. Here, $x^{(f)}, y^{(f)}$ is another independent realization from (1), and $\tilde{\theta}(m)$ is the restricted least-squares estimator corresponding to m.

Two goals

(i) Select a 'good' model from $\mathcal{M}$ for prediction out-of-sample, and (ii) conduct predictive inference based on the selected model, both conditional on the training sample.

Two Quantities of Interest

For $m \in \mathcal{M}$, let $\rho^2(m)$ denote the conditional mean-squared error of the predictor $\hat{y}^{(f)}(m)$ given the training sample, i.e.,

$$\rho^2(m) \quad = \quad E\left[\left(y^{(f)} - \hat{y}^{(f)}(m)\right)^2 \middle\| X, Y\right].$$

For $m \in \mathcal{M}$, the conditional distribution of the prediction error $\hat{y}^{(f)}(m) - y^{(f)}$ given the training sample is

$$\hat{y}^{(f)}(m) - y^{(f)} \,\Big\|\, X, Y \quad \sim \quad N(\nu(m), \delta^2(m)) \quad \equiv \quad \mathbb{L}(m).$$

Note that $\rho^2(m) = \nu^2(m) + \delta^2(m)$.

Two goals

(i) Select a 'good' model from $\mathcal{M}$ for prediction out-of-sample, and (ii) conduct predictive inference based on the selected model, both conditional on the training sample.

Two Quantities of Interest

For $m \in \mathcal{M}$, let $\rho^2(m)$ denote the conditional mean-squared error of the predictor $\hat{y}^{(f)}(m)$ given the training sample, i.e.,

$$\rho^2(m) \quad = \quad E\left[\left(y^{(f)} - \hat{y}^{(f)}(m) \right)^2 \middle\| X, Y \right].$$

For $m \in \mathcal{M}$, the conditional distribution of the prediction error $\hat{y}^{(f)}(m) - y^{(f)}$ given the training sample is

$$\hat{y}^{(f)}(m) - y^{(f)} \, \Big\| \, X, Y \quad \sim \quad N(\nu(m), \delta^2(m)) \quad \equiv \quad \mathbb{L}(m).$$

Note that $\rho^2(m) = \nu^2(m) + \delta^2(m)$.

Two goals

(i) Select a 'good' model from $\mathcal{M}$ for prediction out-of-sample, and (ii) conduct predictive inference based on the selected model, both conditional on the training sample.

Two Quantities of Interest

For $m \in \mathcal{M}$, let $\rho^2(m)$ denote the conditional mean-squared error of the predictor $\hat{y}^{(f)}(m)$ given the training sample, i.e.,

$$\rho^2(m) \quad = \quad E\left[\left(y^{(f)} - \hat{y}^{(f)}(m) \right)^2 \Big\| X, Y \right].$$

For $m \in \mathcal{M}$, the conditional distribution of the prediction error $\hat{y}^{(f)}(m) - y^{(f)}$ given the training sample is

$$\hat{y}^{(f)}(m) - y^{(f)} \Big\| X, Y \quad \sim \quad N(\nu(m), \delta^2(m)) \quad \equiv \quad \mathbb{L}(m).$$

Note that $\rho^2(m) = \nu^2(m) + \delta^2(m)$.

Two goals

(i) Select a 'good' model from $\mathcal{M}$ for prediction out-of-sample, and (ii) conduct predictive inference based on the selected model, both conditional on the training sample.

Two Quantities of Interest

For $m \in \mathcal{M}$, let $\rho^2(m)$ denote the conditional mean-squared error of the predictor $\hat{y}^{(f)}(m)$ given the training sample, i.e.,

$$\rho^2(m) \quad = \quad E\left[\left(y^{(f)} - \hat{y}^{(f)}(m) \right)^2 \middle\| X, Y \right].$$

For $m \in \mathcal{M}$, the conditional distribution of the prediction error $\hat{y}^{(f)}(m) - y^{(f)}$ given the training sample is

$$\hat{y}^{(f)}(m) - y^{(f)} \ \Big\| \ X, Y \quad \sim \quad N(\nu(m), \delta^2(m)) \quad \equiv \quad \mathbb{L}(m).$$

Note that $\rho^2(m) = \nu^2(m) + \delta^2(m)$.

A useful observation

Write $\sigma^2(m)$ for the conditional variance of the response given those explanatory variables that are included in model m, i.e.,

$$\sigma^2(m) \quad = \quad \mathrm{Var}[y \,||\, x_j \text{ included in model } m,\ j \geq 1].$$

Lemma

$$\delta^2(m) \quad \sim \quad \sigma^2(m)\left(1 + \frac{\chi^2_{|m|-1}}{\chi^2_{n-|m|+1}}\right),$$

where the χ^2-random variables are independent. Similarly,

$$\nu^2(m) \quad \sim \quad \frac{1}{n}\sigma^2(m)\left(1 + \frac{\chi^2_{|m|-1}}{\chi^2_{n-|m|+1}}\right),$$

and $\hat{\sigma}^2(m) = \mathrm{RSS}(m)/(n-|m|) \sim \sigma^2(m)\chi^2_{n-|m|}/(n-|m|)$.

[The Lemma extends Theorem 1.3 of Breiman & Friedman (1983).]

A useful observation

Write $\sigma^2(m)$ for the conditional variance of the response given those explanatory variables that are included in model m, i.e.,

$$\sigma^2(m) \quad = \quad \mathrm{Var}[y \mid\mid x_j \text{ included in model } m,\ j \geq 1].$$

Lemma

$$\delta^2(m) \quad \sim \quad \sigma^2(m)\left(1 + \frac{\chi^2_{|m|-1}}{\chi^2_{n-|m|+1}}\right),$$

where the χ^2-random variables are independent. Similarly,

$$\nu^2(m) \quad \sim \quad \frac{1}{n}\sigma^2(m)\left(1 + \frac{\chi^2_{|m|-1}}{\chi^2_{n-|m|+1}}\right),$$

and $\hat{\sigma}^2(m) = \mathsf{RSS}(m)/(n-|m|) \sim \sigma^2(m)\chi^2_{n-|m|}/(n-|m|)$.

[The Lemma extends Theorem 1.3 of Breiman & Friedman (1983).]

Estimators for $\rho^2(m)$

Note that

$$E\left[\rho^2(m)\right] \quad = \quad \sigma^2(m)\frac{n-2}{n-1-|m|}\left(1+\frac{1}{n}\right).$$

The S_p criterion (Tukey, 1967):

$$S_p(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n-2}{n-1-|m|}.$$

The GCV-criterion (Craven & Wahba, 1978):

$$\mathrm{GCV}(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n-|m|}.$$

An auxiliary criterion:

$$\hat{\rho}^2(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n+1-|m|}.$$

Estimators for $\rho^2(m)$

Note that

$$E\left[\rho^2(m)\right] \quad = \quad \sigma^2(m)\frac{n-2}{n-1-|m|}\left(1+\frac{1}{n}\right).$$

The S_p criterion (Tukey, 1967):

$$S_p(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n-2}{n-1-|m|}.$$

The GCV-criterion (Craven & Wahba, 1978):

$$\mathrm{GCV}(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n-|m|}.$$

An auxiliary criterion:

$$\hat{\rho}^2(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n+1-|m|}.$$

Estimators for $\rho^2(m)$

Note that

$$E\left[\rho^2(m)\right] \quad = \quad \sigma^2(m)\frac{n-2}{n-1-|m|}\left(1+\frac{1}{n}\right).$$

The S_p criterion (Tukey, 1967):

$$S_p(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n-2}{n-1-|m|}.$$

The GCV-criterion (Craven & Wahba, 1978):

$$\mathrm{GCV}(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n-|m|}.$$

An auxiliary criterion:

$$\hat{\rho}^2(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n+1-|m|}.$$

Estimators for $\rho^2(m)$

Note that

$$E\left[\rho^2(m)\right] \quad = \quad \sigma^2(m)\frac{n-2}{n-1-|m|}\left(1+\frac{1}{n}\right).$$

The S_p criterion (Tukey, 1967):

$$S_p(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n-2}{n-1-|m|}.$$

The GCV-criterion (Craven & Wahba, 1978):

$$\mathrm{GCV}(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n-|m|}.$$

An auxiliary criterion:

$$\hat{\rho}^2(m) \quad = \quad \hat{\sigma}^2(m)\,\frac{n}{n+1-|m|}.$$

Performance of $\hat{\rho}^2(m)$

Want: $\hat{\rho}^2(m)/\rho^2(m) \approx 1$ or, equivalently, $\log \hat{\rho}^2(m)/\rho^2(m) \approx 0$ with high probability.

Theorem

For each $\epsilon > 0$, we have

$$P\left(\left|\log \frac{\hat{\rho}^2(m)}{\rho^2(m)}\right| > \epsilon\right) \quad \leq \quad 6 \exp\left[-\frac{n-|m|}{8}\frac{\epsilon^2}{\epsilon+8}\right],$$

for each sample size n and uniformly over all data-generating processes as in (1).

Performance of $\hat{\rho}^2(m)$

Want: $\hat{\rho}^2(m)/\rho^2(m) \approx 1$ or, equivalently, $\log \hat{\rho}^2(m)/\rho^2(m) \approx 0$ with high probability.

Theorem

For each $\epsilon > 0$, we have

$$P\left(\left|\log \frac{\hat{\rho}^2(m)}{\rho^2(m)}\right| > \epsilon\right) \quad \leq \quad 6\exp\left[-\frac{n-|m|}{8}\,\frac{\epsilon^2}{\epsilon+8}\right],$$

for each sample size n and uniformly over all data-generating processes as in (1).

Performance of $\hat\rho^2(m)$

Want: $\hat\rho^2(m)/\rho^2(m) \approx 1$ or, equivalently, $\log\hat\rho^2(m)/\rho^2(m) \approx 0$ with high probability.

Theorem

For each $\epsilon > 0$, we have

$$P\left(\left|\log\frac{\hat\rho^2(m)}{\rho^2(m)}\right| > \epsilon\right) \quad\leq\quad 6\exp\left[-\frac{n-|m|}{8}\,\frac{\epsilon^2}{\epsilon+8}\right],$$

for each sample size n and uniformly over all data-generating processes as in (1).

A similar result holds for the absolute difference $|\hat\rho^2(m) - \rho^2(m)|$, uniformly over all data-generating processes with bounded variance, i.e., where $\operatorname{Var}[y] \leq s^2$ (with an upper bound of the form $C_1\exp[-(n-|m|)\,C(\epsilon,s^2)]$; here s^2 is a fixed constant).

Performance of $\hat{\rho}^2(m)$

Want: $\hat{\rho}^2(m)/\rho^2(m) \approx 1$ or, equivalently, $\log \hat{\rho}^2(m)/\rho^2(m) \approx 0$ with high probability.

Theorem

For each $\epsilon > 0$, we have

$$P\left(\left|\log \frac{\hat{\rho}^2(m)}{\rho^2(m)}\right| > \epsilon\right) \quad \leq \quad 6\exp\left[-\frac{n-|m|}{8}\frac{\epsilon^2}{\epsilon+8}\right],$$

for each sample size n and uniformly over all data-generating processes as in (1).

Method of proof: Chernoff's method or variations thereof (Gaussian case); Marčenko-Pastur law (non-Gaussian case).

Selecting the empirically best model

Write m_* and $\hat{m}$ for the truly best and the empirically best candidate model, i.e.,

$$m_* = \operatorname{argmin}_{\mathcal{M}} \rho^2(m) \quad \text{and} \quad \hat{m} = \operatorname{argmin}_{\mathcal{M}} \hat{\rho}^2(m).$$

Moreover, write $|\mathcal{M}|$ for the number of parameters in the most complex candidate model.

Corollary

For each fixed sample size n and uniformly over all data-generating processes as in (1), we have

$$\begin{aligned} P\left(\log \frac{\rho^2(\hat{m})}{\rho^2(m_*)} > \epsilon\right) &\leq 6\exp\left[\log \#\mathcal{M} - \frac{n-|\mathcal{M}|}{16}\frac{\epsilon^2}{\epsilon+16}\right], \\ P\left(\left|\log \frac{\hat{\rho}^2(\hat{m})}{\rho^2(\hat{m})}\right| > \epsilon\right) &\leq 6\exp\left[\log \#\mathcal{M} - \frac{n-|\mathcal{M}|}{8}\frac{\epsilon^2}{\epsilon+8}\right], \end{aligned}$$

for each $\epsilon > 0$.

Selecting the empirically best model

Write m_* and $\hat{m}$ for the truly best and the empirically best candidate model, i.e.,

$$m_* = \text{argmin}_{\mathcal{M}} \rho^2(m) \quad \text{and} \quad \hat{m} = \text{argmin}_{\mathcal{M}} \hat{\rho}^2(m).$$

Moreover, write $|\mathcal{M}|$ for the number of parameters in the most complex candidate model.

Corollary

For each fixed sample size n and uniformly over all data-generating processes as in (1), we have

$$\begin{aligned} P\left(\log \frac{\rho^2(\hat{m})}{\rho^2(m_*)} > \epsilon\right) &\leq 6 \exp\left[\log \#\mathcal{M} - \frac{n - |\mathcal{M}|}{16} \frac{\epsilon^2}{\epsilon + 16}\right], \\ P\left(\left|\log \frac{\hat{\rho}^2(\hat{m})}{\rho^2(\hat{m})}\right| > \epsilon\right) &\leq 6 \exp\left[\log \#\mathcal{M} - \frac{n - |\mathcal{M}|}{8} \frac{\epsilon^2}{\epsilon + 8}\right], \end{aligned}$$

for each $\epsilon > 0$.

Other model selectors

Consider AIC (Akaike, 1969), AICc (Hurvich & Tsai, 1989), FPE (Akaike, 1970), and BIC (Schwarz, 1978). Taking the exponential of the objective functions of AIC, AICc and BIC, and using the fact that $\mathrm{GCV}(m) = \frac{1}{n}\mathrm{RSS}(m)/(1-|m|/n)^2 \approx \rho(m)$, we get

$$
\begin{aligned}
\mathrm{AIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|}{n}} &&\approx \rho(m)e^{2\frac{|m|}{n}}(1-|m|/n)^2\\
\mathrm{AICc}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|-1}{n-|m|-2}} &&\approx \rho(m)e^{2\frac{|m|-1}{n-|m|-2}}(1-|m|/n)^2\\
\mathrm{FPE}(m) &= \frac{1}{n}\mathrm{RSS}(m)\frac{1+|m|/n}{1-|m|/n} &&\approx \rho(m)(1+|m|/n)(1-|m|/n)\\
\mathrm{BIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{\log(n)\frac{|m|}{n}} &&\approx \rho(m)n^{|m|/n}(1-|m|/n)^2.
\end{aligned}
$$

Other model selectors

Consider AIC (Akaike, 1969), AICc (Hurvich & Tsai, 1989), FPE (Akaike, 1970), and BIC (Schwarz, 1978). Taking the exponential of the objective functions of AIC, AICc and BIC, and using the fact that $\mathrm{GCV}(m) = \frac{1}{n}\mathrm{RSS}(m)/(1-|m|/n)^2 \approx \rho(m)$, we get

$$\begin{aligned}
\mathrm{AIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|}{n}} &&\approx \rho(m)e^{2\frac{|m|}{n}}(1-|m|/n)^2 \\
\mathrm{AICc}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|-1}{n-|m|-2}} &&\approx \rho(m)e^{2\frac{|m|-1}{n-|m|-2}}(1-|m|/n)^2 \\
\mathrm{FPE}(m) &= \frac{1}{n}\mathrm{RSS}(m)\frac{1+|m|/n}{1-|m|/n} &&\approx \rho(m)(1+|m|/n)(1-|m|/n) \\
\mathrm{BIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{\log(n)\frac{|m|}{n}} &&\approx \rho(m)n^{|m|/n}(1-|m|/n)^2.
\end{aligned}$$

Other model selectors

Consider AIC (Akaike, 1969), AICc (Hurvich & Tsai, 1989), FPE (Akaike, 1970), and BIC (Schwarz, 1978). Taking the exponential of the objective functions of AIC, AICc and BIC, and using the fact that $\mathrm{GCV}(m) = \frac{1}{n}\mathrm{RSS}(m)/(1-|m|/n)^2 \approx \rho(m)$, we get

$$\begin{aligned}
\mathrm{AIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|}{n}} &&\approx \rho(m)e^{2\frac{|m|}{n}}(1-|m|/n)^2\\
\mathrm{AICc}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|-1}{n-|m|-2}} &&\approx \rho(m)e^{2\frac{|m|-1}{n-|m|-2}}(1-|m|/n)^2\\
\mathrm{FPE}(m) &= \frac{1}{n}\mathrm{RSS}(m)\frac{1+|m|/n}{1-|m|/n} &&\approx \rho(m)(1+|m|/n)(1-|m|/n)\\
\mathrm{BIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{\log(n)\frac{|m|}{n}} &&\approx \rho(m)n^{|m|/n}(1-|m|/n)^2.
\end{aligned}$$

Other model selectors

Consider AIC (Akaike, 1969), AICc (Hurvich & Tsai, 1989), FPE (Akaike, 1970), and BIC (Schwarz, 1978). Taking the exponential of the objective functions of AIC, AICc and BIC, and using the fact that $\mathrm{GCV}(m) = \frac{1}{n}\mathrm{RSS}(m)/(1-|m|/n)^2 \approx \rho(m)$, we get

$$\begin{aligned}
\mathrm{AIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|}{n}} &&\approx \rho(m)e^{2\frac{|m|}{n}}(1-|m|/n)^2\\
\mathrm{AICc}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|-1}{n-|m|-2}} &&\approx \rho(m)e^{2\frac{|m|-1}{n-|m|-2}}(1-|m|/n)^2\\
\mathrm{FPE}(m) &= \frac{1}{n}\mathrm{RSS}(m)\frac{1+|m|/n}{1-|m|/n} &&\approx \rho(m)(1+|m|/n)(1-|m|/n)\\
\mathrm{BIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{\log(n)\frac{|m|}{n}} &&\approx \rho(m)n^{|m|/n}(1-|m|/n)^2.
\end{aligned}$$

Other model selectors

Consider AIC (Akaike, 1969), AICc (Hurvich & Tsai, 1989), FPE (Akaike, 1970), and BIC (Schwarz, 1978). Taking the exponential of the objective functions of AIC, AICc and BIC, and using the fact that $\mathrm{GCV}(m) = \frac{1}{n}\mathrm{RSS}(m)/(1-|m|/n)^2 \approx \rho(m)$, we get

$$\begin{aligned}
\mathrm{AIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|}{n}} &&\approx \rho(m)e^{2\frac{|m|}{n}}(1-|m|/n)^2\\
\mathrm{AICc}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|-1}{n-|m|-2}} &&\approx \rho(m)e^{2\frac{|m|-1}{n-|m|-2}}(1-|m|/n)^2\\
\mathrm{FPE}(m) &= \frac{1}{n}\mathrm{RSS}(m)\frac{1+|m|/n}{1-|m|/n} &&\approx \rho(m)(1+|m|/n)(1-|m|/n)\\
\mathrm{BIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{\log(n)\frac{|m|}{n}} &&\approx \rho(m)n^{|m|/n}(1-|m|/n)^2.
\end{aligned}$$

Other model selectors

Consider AIC (Akaike, 1969), AICc (Hurvich & Tsai, 1989), FPE (Akaike, 1970), and BIC (Schwarz, 1978). Taking the exponential of the objective functions of AIC, AICc and BIC, and using the fact that $\mathrm{GCV}(m) = \frac{1}{n}\mathrm{RSS}(m)/(1-|m|/n)^2 \approx \rho(m)$, we get

$$\begin{aligned}
\mathrm{AIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|}{n}} &&\approx \rho(m)e^{2\frac{|m|}{n}}(1-|m|/n)^2\\
\mathrm{AICc}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{2\frac{|m|-1}{n-|m|-2}} &&\approx \rho(m)e^{2\frac{|m|-1}{n-|m|-2}}(1-|m|/n)^2\\
\mathrm{FPE}(m) &= \frac{1}{n}\mathrm{RSS}(m)\frac{1+|m|/n}{1-|m|/n} &&\approx \rho(m)(1+|m|/n)(1-|m|/n)\\
\mathrm{BIC}(m) &= \frac{1}{n}\mathrm{RSS}(m)e^{\log(n)\frac{|m|}{n}} &&\approx \rho(m)n^{|m|/n}(1-|m|/n)^2.
\end{aligned}$$

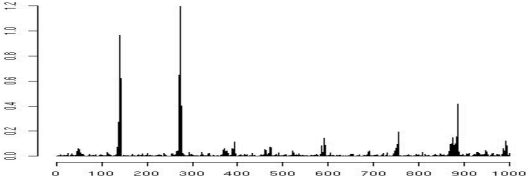
Simulation Scenario I

Consider **one** sample of size $n = 1300$ from (1) with $E[x_j] = 0$, $E[x_i x_j] = \delta_{i,j}$, and $E[u^2] = 1$.

Simulation Scenario I

Consider **one** sample of size $n = 1300$ from (1) with $E[x_j] = 0$, $E[x_i x_j] = \delta_{i,j}$, and $E[u^2] = 1$.

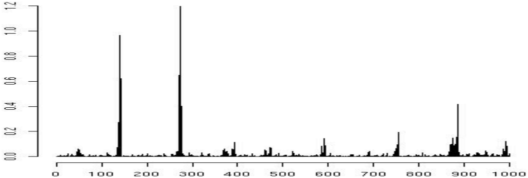
The first 1000 components of θ are shown (in absolute value) below, the remaining components are zero:



Simulation Scenario I

Consider **one** sample of size $n = 1300$ from (1) with $E[x_j] = 0$, $E[x_i x_j] = \delta_{i,j}$, and $E[u^2] = 1$.

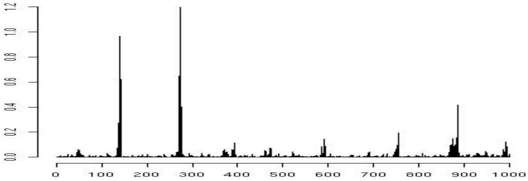
The first 1000 components of θ are shown (in absolute value) below, the remaining components are zero:



The non-zero coefficients of θ are 'sparse:' Most are small, but there are a few groups of adjacent large coefficients.

Simulation Scenario I

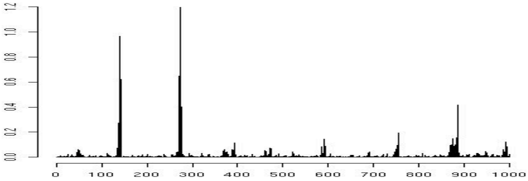
Consider **one** sample of size $n = 1300$ from (1) with $E[x_j] = 0$, $E[x_i x_j] = \delta_{i,j}$, and $E[u^2] = 1$.
The first 1000 components of θ are shown (in absolute value) below, the remaining components are zero:



Choose candidate models that can pick-out the few important groups: Divide the first 1000 coefficients of θ into 20 consecutive blocks of equal length and consider all candidate models that include or exclude one block at a time, resulting in 2^{20} candidate models.

Simulation Scenario I

Consider **one** sample of size $n = 1300$ from (1) with $E[x_j] = 0$, $E[x_i x_j] = \delta_{i,j}$, and $E[u^2] = 1$.
The first 1000 components of θ are shown (in absolute value) below, the remaining components are zero:

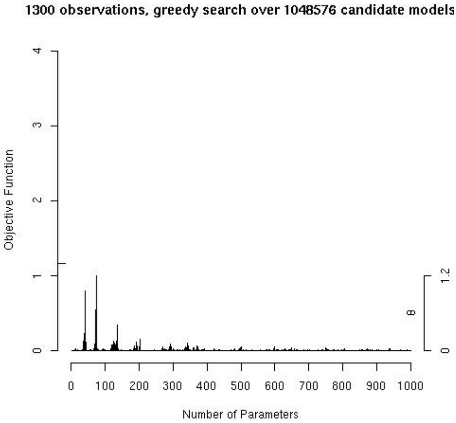


Model space is searched using a general-to-specific greedy strategy.

Simulation Scenario I

Results for X Gaussian, u Gaussian:

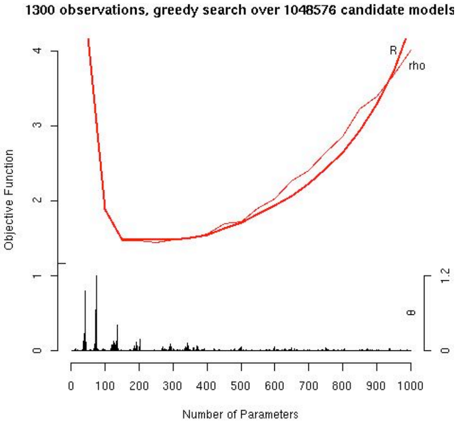
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

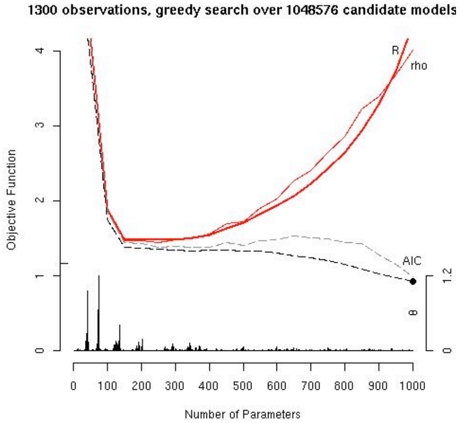
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

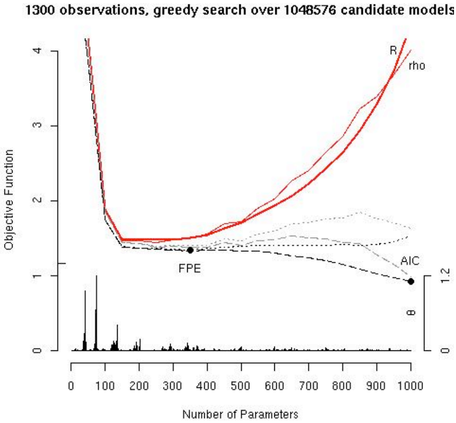
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

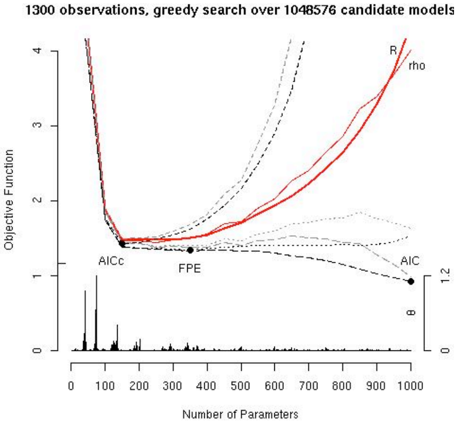
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

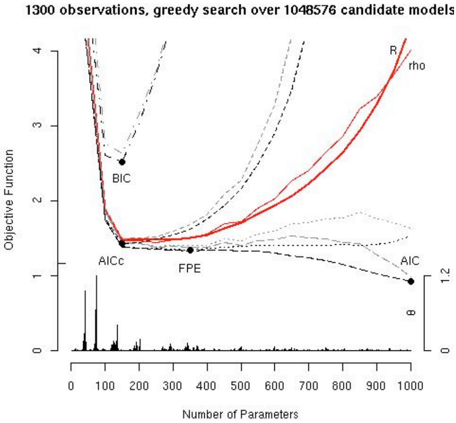
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

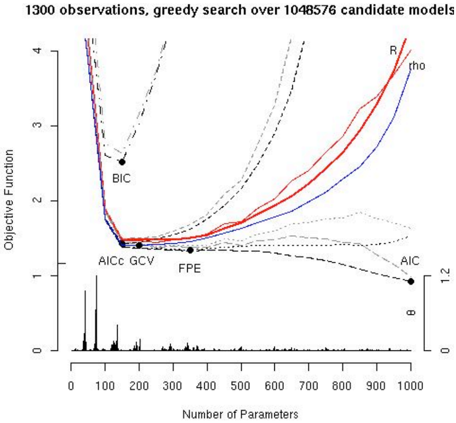
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

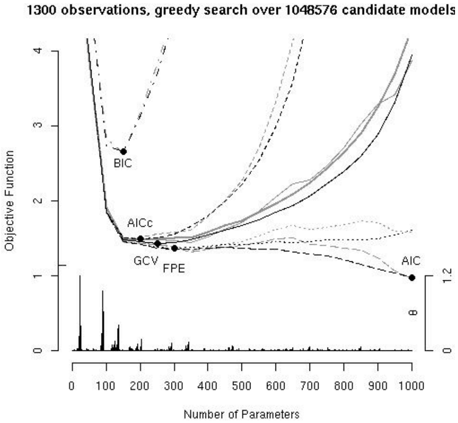
Run 1:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

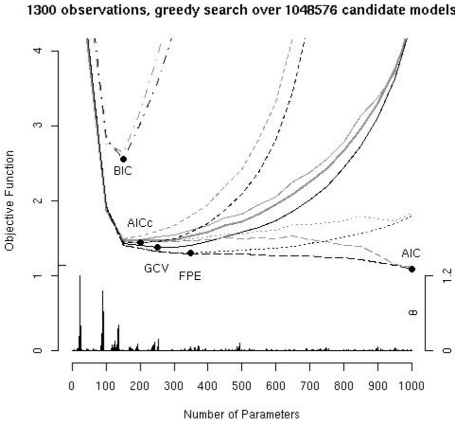
Run 2:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

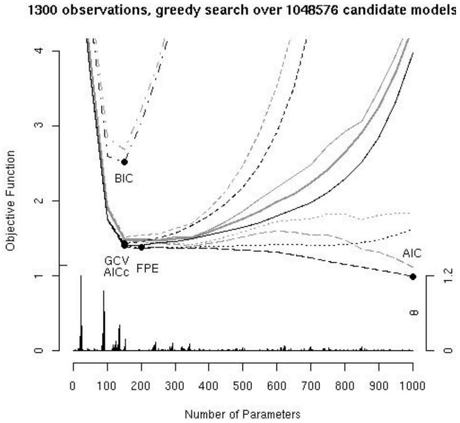
Run 3:



Simulation Scenario I

Results for X Gaussian, u Gaussian:

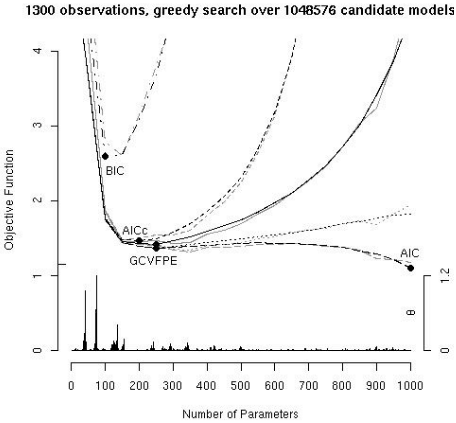
Run 4:



Simulation Scenario I

Results for X Exponential, u Bernoulli (scaled and centered).

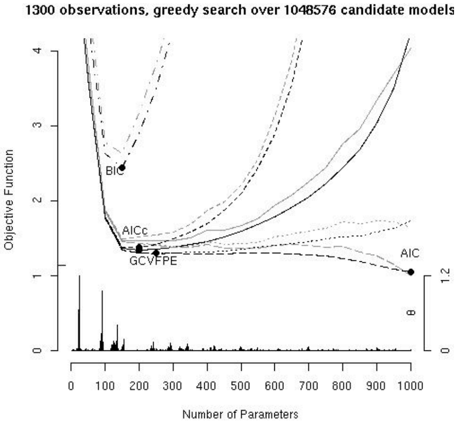
Run 1:



Simulation Scenario I

Results for X Bernoulli, u Exponential (scaled and centered).

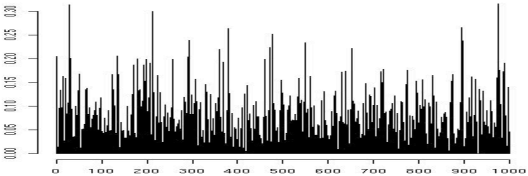
Run 1:



Simulation Scenario II

Consider the same setting as in Scenario I, but instead of a parameter θ that is 'sparse,' consider a case where none of the candidate models fits particularly well.

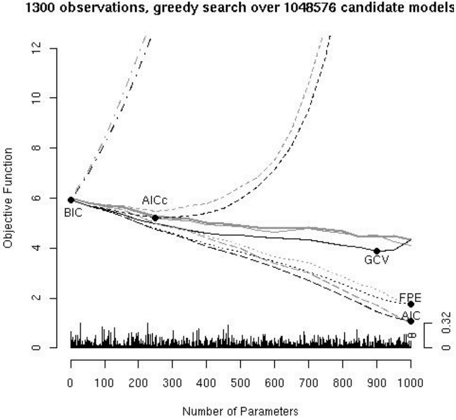
The first 1000 components of θ are shown (in absolute value) below, the remaining components are zero:



Simulation Scenario II

Results for X Gaussian, u Gaussian:

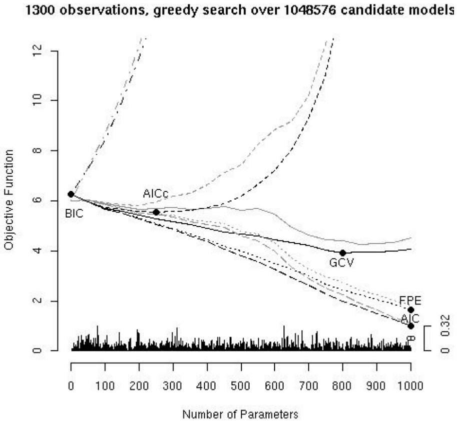
Run 1:



Simulation Scenario II

Results for X Exponential, u Bernoulli (scaled and centered).

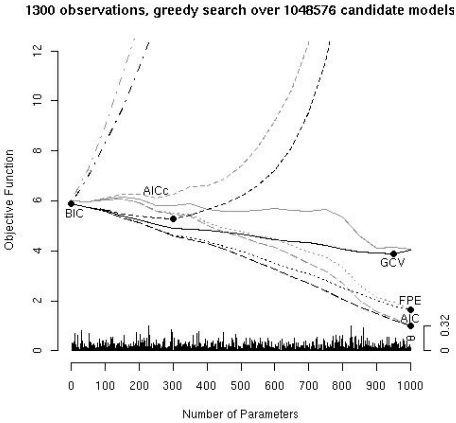
Run 1:



Simulation Scenario II

Results for X Bernoulli, u Exponential (scaled and centered).

Run 1:



Predictive Inference based on model m

Idea: Estimate the conditional distribution of the prediction error, i.e., $\mathbb{L}(m) \equiv N(\nu(m), \delta^2(m))$, by

$$\hat{\mathbb{L}}(m) \quad \equiv \quad N(0, \hat{\delta}^2(m)),$$

where $\hat{\delta}^2(m)$ is defined as $\hat{\rho}^2(m)$ before.

Theorem

For each fixed sample size n and uniformly over all data-generating processes as in (1), we have

$$\begin{aligned} & P\left(\left|\left| \hat{\mathbb{L}}(m) - \mathbb{L}(m) \right|\right|_{TV} > \frac{1}{\sqrt{n}} + \epsilon \right) \\ & \quad \leq \quad 7 \exp\left[-\frac{n-|m|}{2} \frac{\epsilon^2}{\epsilon+2} \right] \end{aligned}$$

for each $\epsilon \leq \log(2) \approx 0.69$.

Predictive Inference based on model m

Idea: Estimate the conditional distribution of the prediction error, i.e., $\mathbb{L}(m) \equiv N(\nu(m), \delta^2(m))$, by

$$\hat{\mathbb{L}}(m) \quad \equiv \quad N(0, \hat{\delta}^2(m)),$$

where $\hat{\delta}^2(m)$ is defined as $\hat{\rho}^2(m)$ before.

Theorem

For each fixed sample size n and uniformly over all data-generating processes as in (1), we have

$$\begin{aligned} P\left(\left|\left| \hat{\mathbb{L}}(m) - \mathbb{L}(m) \right|\right|_{TV} > \frac{1}{\sqrt{n}} + \epsilon \right) \\ \leq \quad 7 \exp\left[-\frac{n - |m|}{2} \, \frac{\epsilon^2}{\epsilon + 2} \right] \end{aligned}$$

for each $\epsilon \leq \log(2) \approx 0.69$.

Prediction intervals post model selection

Recall that $\hat{y}^{(f)}(m) - y^{(f)} \,||\, X, Y \sim N(\nu(m), \delta^2(m)) \equiv \mathbb{L}(m)$ for each $m \in \mathcal{M}$. Based on model m, the 'prediction interval'

$$\hat{y}^{(f)}(m) - \nu(m) \quad \pm \quad q_{\alpha/2}\delta(m)$$

has coverage probability $1-\alpha$ conditional on the training sample X, Y, but is infeasible.

In terms of width of this interval, the 'best' model is one that minimizes $\delta(m)$. Set

$$m_\circ \quad = \quad \operatorname{argmin}_{\mathcal{M}} \delta^2(m).$$

For fixed $m \in \mathcal{M}$, a feasible prediction interval is

$$\mathcal{I}(m): \qquad \hat{y}^{(f)}(m) \quad \pm \quad q_{\alpha/2}\hat{\delta}(m).$$

Prediction intervals post model selection

Recall that $\hat{y}^{(f)}(m) - y^{(f)} \,||\, X, Y \sim N(\nu(m), \delta^2(m)) \equiv \mathbb{L}(m)$ for each $m \in \mathcal{M}$. Based on model m, the 'prediction interval'

$$\hat{y}^{(f)}(m) - \nu(m) \quad \pm \quad q_{\alpha/2}\delta(m)$$

has coverage probability $1 - \alpha$ conditional on the training sample X, Y, but is infeasible.

In terms of width of this interval, the 'best' model is one that minimizes $\delta(m)$. Set

$$m_\circ \quad = \quad \operatorname{argmin}_{\mathcal{M}} \delta^2(m).$$

For fixed $m \in \mathcal{M}$, a feasible prediction interval is

$$\mathcal{I}(m): \qquad \hat{y}^{(f)}(m) \quad \pm \quad q_{\alpha/2}\hat{\delta}(m).$$

Prediction intervals post model selection

Recall that $\hat{y}^{(f)}(m) - y^{(f)} \mid\mid X, Y \sim N(\nu(m), \delta^2(m)) \equiv \mathbb{L}(m)$ for each $m \in \mathcal{M}$. Based on model m, the 'prediction interval'

$$\hat{y}^{(f)}(m) - \nu(m) \quad \pm \quad q_{\alpha/2}\delta(m)$$

has coverage probability $1 - \alpha$ conditional on the training sample X, Y, but is infeasible.

In terms of width of this interval, the 'best' model is one that minimizes $\delta(m)$. Set

$$m_\circ \quad = \quad \mathrm{argmin}_{\mathcal{M}} \delta^2(m).$$

For fixed $m \in \mathcal{M}$, a feasible prediction interval is

$$\mathcal{I}(m): \qquad \hat{y}^{(f)}(m) \quad \pm \quad q_{\alpha/2}\hat{\delta}(m).$$

Prediction interval is approx. valid & adaptive

Proposition

For each $\epsilon \leq \log 2$ and each fixed sample size n, we have

$$\left| \Big(1 - \alpha\Big) \;-\; P\Big(y^{(f)} \in \mathcal{I}(\hat{m}) \,\Big\|\, Y, X\Big) \right| \;\leq\; \frac{1}{\sqrt{n}} + \epsilon$$

and

$$\left| \log \frac{\hat{\delta}(\hat{m})}{\delta(m_\circ)} \right| \;\leq\; \epsilon,$$

except on an event whose probability is not larger than

$$11 \exp\left[\log \#\mathcal{M} - \frac{n - |\mathcal{M}|}{2} \frac{\epsilon^2}{\epsilon + 2} \right],$$

uniformly over all data-generating processes as in (1).

Conclusion

Caution:

The 'large p / small n' behavior of model selectors can be markedly different from their properties for 'small p / large n'.

Proof of concept: The two goals are achievable

In 'large p / small n' settings and under minimal assumptions, good models can be found, and the resulting prediction intervals post model selection are approximately valid and adaptive (in finite samples with high probability uniformly over all data-generating processes considered).

Conclusion

Caution:

The ‘large p / small n’ behavior of model selectors can be markedly different from their properties for ‘small p / large n’.

Proof of concept: The two goals are achievable

In ‘large p / small n’ settings and under minimal assumptions, good models can be found, and the resulting prediction intervals post model selection are approximately valid and adaptive (in finite samples with high probability uniformly over all data-generating processes considered).

Conclusion

Caution:

The 'large p / small n' behavior of model selectors can be markedly different from their properties for 'small p / large n'.

Proof of concept: The two goals are achievable

In 'large p / small n' settings and under minimal assumptions, good models can be found, and the resulting prediction intervals post model selection are approximately valid and adaptive (in finite samples with high probability uniformly over all data-generating processes considered).