Ein Überblick über Modellwahlverfahren

Universität Dortmund, Kolloquium

Hans R. Künsch

Seminar für Statistik

ETH Zürich

12. Juli 2005

Contents

1. Introduction
2. Akaike's Information Criterion
3. Rissanen's Minimum Description Length
4. Bayesian model selection
5. Selecting covariates in regression
6. Discussion

1. Introduction

1.1 Examples of model selection problems

A model selection problem arises whenever “the number of things that you don’t know is one of the things that you don’t know.” For instance

- Choice of explanatory variables in regression.
- Choice of orders p, q of an ARMA(p, q)-model.
- Choice of numbers of clusters in a cluster algorithm.
- ...

The choice of tuning constants in nonparametric curve estimation (e.g. bandwidth or smoothing parameters, number of knots in a CART procedure, choice of wavelet coefficients to truncate) can also be considered as model selection problem .

Finally we have the choice between two or more models that are not nested, e.g. the choice between a Fourier and a wavelet basis for expansion of a regression function.

1.2 Purposes of model selection

- Understanding and interpretation
- Prediction
- Inference about a particular parameter of interest common to all models.

1.3 Basic ideas for model selection

Ockham's razor: “Entities are not to be mulitplied beyond necessity”
(William of Ockham, 1285 – 1347/49).

The goal is a distinction between reproducible and non-reproducible structures in the data, or in other words a balance between systematic and random errors.

An **unbiased** (fair) assessment of goodness of fit of a model allows model selection (and is of independent interest).

The sum of squared errors or the maximum of log likelihood are biased towards complicated models. Possible corrections will be shown in this talk.

2. Akaike's Information Criterion (AIC)

2.1 Codes and entropy

For the following, we need some concepts from information theory.

Binary codes of a finite alphabet $\{1, 2, \ldots, m\}$ are equivalent to probability distributions on this set via

$$\text{code length of symbol } i \longleftrightarrow -\log p(i).$$

The **entropy** is the minimal expected code length:

$$H(P) = -\sum_i p(i) \log p(i) = \min_Q -\sum_i p(i) \log q(i).$$

In particular, the optimal code is based on the probabilities by which the symbols are produced.

The **Kullback-Leibler divergence** $K(P, Q)$ is the increase in expected code length when using q instead of p

$$-\sum_i p(i) \log q(i) + \sum_i p(i) \log p(i) = \sum_i p(i) \log \frac{p(i)}{q(i)}.$$

These concepts and quantities can be generalized to continuous symbols by discretizing the space. The expected code length contains then a term which diverges, but is independent of the chosen code.

2.2 The key idea of AIC

We use the following notation

Data: $\mathbf{y}^{obs} = (y_1^{obs}, \ldots, y_n^{obs})$.

Distribution that generated the data: G, with density $g(\mathbf{y})$.

Model distributions: F_θ, with densities $f(\mathbf{y}|\theta)$.

Parameter: θ in an open subset Θ of $\mathbb{R}^p$ (so p is the dimension of the parameter).

The lack of fit of a model distribution to the true distribution is measured by the Kullback-Leibler divergence $K(G, F_\theta) =$

$$\int \log \frac{g(\mathbf{y})}{f(\mathbf{y}|\theta)}\, g(\mathbf{y})d\mathbf{y} = \int \log g(\mathbf{y})\; g(\mathbf{y})d\mathbf{y} - \int \log f(\mathbf{y}|\theta)\; g(\mathbf{y})d\mathbf{y}.$$

If only differences in the lack of fit are of interest, we can drop the first term.

Hence the parameter of the model distribution that fits best to G ist

$$\overline{\theta} = \arg\min_{\theta} K(G, F_{\theta}) = \arg\min_{\theta} \int -\log(f(\mathbf{y}|\theta))\, g(\mathbf{y})d\mathbf{y}.$$

Moreover, $-\log f(\mathbf{y}^{obs}|\theta)$ is an unbiased estimate of $\int -\log f(\mathbf{y}|\theta)\, g(\mathbf{y})d\mathbf{y}$ for any fixed θ. Therefore the MLE

$$\widehat{\theta}(\mathbf{y}^{obs}) = \arg\min_{\theta}(-\log f(\mathbf{y}^{obs}|\theta))$$

maximizes the estimated goodness of fit.

However, $-\log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs}))$ is **not** an unbiased estimator of $\int -\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}^{obs}))\, g(\mathbf{y})d\mathbf{y}$ and thus cannot serve as a basis for model selection.

To see why, decompose

$$
\begin{aligned}
&-\log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) + \int \log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}^{obs}))\, g(\mathbf{y}) d\mathbf{y} \\
= \quad & -\log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) + \log f(\mathbf{y}^{obs}|\overline{\theta}) \\
& -\log f(\mathbf{y}^{obs}|\overline{\theta}) + \int \log f(\mathbf{y}|\overline{\theta})\, g(\mathbf{y}) d\mathbf{y} \\
& -\int \log f(\mathbf{y}|\overline{\theta})\, g(\mathbf{y}) d\mathbf{y} + \int \log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}^{obs}))\, g(\mathbf{y}) d\mathbf{y} \\
=: \quad & D_1 + D_2 + D_3.
\end{aligned}
$$

Then D_2 has expectation zero, but D_1 and D_3 are always negative by the definition of $\widehat{\theta}$ and $\overline{\theta}$.

2.2 Quantifying the bias

Use asymptotics for the bias of D_1 and D_3. For simplicity, consider i.i.d. observations:

$$g(\mathbf{y}) = \prod_i g_1(y_i), \quad f(\mathbf{y}|\theta) = \prod_i f_1(y_i|\theta).$$

Then in regular situations

$$\widehat{\theta}(\mathbf{y}^{obs}) \to \overline{\theta}, \quad \sqrt{n}(\widehat{\theta}(\mathbf{y}^{obs}) - \overline{\theta}) \xrightarrow{d} \mathcal{N}(0, J_0^{-1} I_0 J_0^{-1}),$$

where

$$\begin{aligned} J_0 &= -\int \frac{\partial^2}{\partial\theta\partial\theta^T} \log f_1(y_1|\overline{\theta})\, g_1(y_1) dy_1, \\ I_0 &= \int \frac{\partial}{\partial\theta} \log f_1(y_1|\overline{\theta}) \left(\frac{\partial}{\partial\theta} \log f_1(y_1|\overline{\theta}) \right)^T g_1(y_1) dy_1. \end{aligned}$$

Moreover,

$$\mathbf{E}[D_1] \approx \mathbf{E}[D_3] \approx -\frac{1}{2}\mathrm{trace}(J_0^{-1} I_0).$$

This gives (multiplying by 2 for convenience)

$$\text{AIC} = -2\ \log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) + 2\,\mathrm{trace}(J_0^{-1} I_0).$$

Estimation of $\mathrm{trace}(J_0^{-1} I_0)$:

- The original approach by Akaike uses $\mathrm{trace}(J_0^{-1} I_0) \approx p$. This is justified because $J_0 = I_0$ if the model holds.
- Takeuchi suggested to estimate J_0 and I_0 by averages over the data.
- Konishi and Kitagawa use the bootstrap to estimate the bias of D_1 and D_3. This allows to use other estimators than MLE.

2.3 Cross validation

This avoids the bias problem entirely by using

$$-\sum_{i=1}^{n} \log f_i(y_i^{obs}|\widehat{\theta}^{(-i)})$$

as an estimator of

$$\int -\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}^{obs}))\, g(\mathbf{y})d\mathbf{y} = -\sum_{i=1}^{n} \int \log f_i(y_i|\widehat{\theta}(\mathbf{y}^{obs}))\, g_i(y_i)dy_i.$$

Here $\widehat{\theta}^{(-i)}$ is the MLE based on $(y_j^{obs}, j \neq i)$. This is asymptotically equivalent to

$$-\sum_{i=1}^{n} \log f_i(y_i^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) + \mathrm{trace}(J_0^{-1} I_0).$$

2.4 The deviance information criterion

This is a Bayesian variant of AIC proposed by Spiegelhalter et al.:

$$\text{DIC} = -2\ \log f(\mathbf{y}^{obs}|\tilde{\theta}(\mathbf{y}^{obs})) + 2\, p_D$$

where $\tilde{\theta}(\mathbf{y}^{obs})$ is the posterior mean of θ and p_D is the Bayesian model complexity.

In AIC, the complexity of the model is

$$\mathrm{trace}(J_0^{-1} I_0) \approx 2\mathbf{E}[D_1] = \mathbf{E}[2 \log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) - 2 \log f(\mathbf{y}^{obs}|\overline{\theta})].$$

In analogy they define the Bayesian model complexity as

$$p_D = 2 \log f(\mathbf{y}^{obs}|\tilde{\theta}(\mathbf{y}^{obs})) - \mathbf{E}[2 \log f(\mathbf{y}^{obs}|\theta) \mid \mathbf{y}^{obs}].$$

This can easily be approximated by standard MCMC methods.

3. Rissanen's Minimum Description Length (MDL)

A precursor is Wallace and Boulton (1968).

3.1 The key idea of MDL

We have seen, that a distribution is equivalent to a code for the data. Moreover, this code achieves a good compression of the data iff the distribution is close to the distribution which generated the data.

For a whole class of distributions, can we find a code that comes close to the best compression achievable with these distributions ? In other words, is there a density $\overline{f}$ such that for any $\mathbf{y}$

$$-\log \overline{f}(\mathbf{y}) \approx \min_{\theta} \left(-\log f(\mathbf{y}|\theta)\right) ?$$

(Note that the right hand side is **not** minus the log of a density.)

Rissanen has shown that there are densities $\overline{f}$ such that

$$-\log \overline{f}(\mathbf{y}) = \min_{\theta} \left(-\log f(\mathbf{y}|\theta)\right) + \frac{p}{2} \log(n)$$

plus lower order terms, and this is the best what one can get (A rigorous statement needs more care).

Therefore

$$\min_{\theta} \big(-\log f(\mathbf{y}^{obs}|\theta)\big) + \frac{p}{2} \log(n) = -\log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) + \frac{p}{2} \log(n)$$

measures the compression of the data by the model and serves a model selection criterion.

3.2 Two part codes

How does the mysterious term $\frac{p}{2}\log(n)$ arise ?

$\min_\theta - \log f(\mathbf{y}|\theta)$ is not a code length because $\widehat{\theta}(\mathbf{y})$, the θ which achieves the minimum, depends on $\mathbf{y}$ and is thus unknown to the person who does the decoding. In order to get around this, encode $\widehat{\theta}(\mathbf{y})$ first and then encode $\mathbf{y}$ with the code corresponding to $f(.|\widehat{\theta}(\mathbf{y}))$.

In order to do this, one needs to choose a precision for encoding $\widehat{\theta}(\mathbf{y})$. The best choice turns out to be a precision $\delta = 1/\sqrt{n}$ for each component θ_i. Then the code length for encoding $\widehat{\theta}(\mathbf{y})$ is $\frac{p}{2}\log(n)$. This is obvious if Θ is the p-dimensional unit cube. The general situation is slightly more complicated.

3.3 Alternative codes

For a more specific criterion, take a density $\overline{f}$ which is asymptotically optimal.

Mixture MDL takes

$$\overline{f}(\mathbf{y}) = \int_{\Theta} f(\mathbf{y}|\theta)w(\theta)d\theta$$

with an arbitrary density w. A Laplace approximation gives again

$$-\log \overline{f}(\mathbf{y}^{obs}) \approx -\log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) + \frac{p}{2}\log(n).$$

Another possibility is **normalized MDL**:

$$\overline{f}(\mathbf{y}) = \frac{\max_{\theta} f(\mathbf{y}|\theta)}{\int \max_{\theta} f(\mathbf{y}'|\theta)d\mathbf{y}'}$$

(if the denominator is infinite, some modification is necessary).

4. Bayesian model selection

4.1 Bayesian theory

A Bayesian puts a **prior** π on the (finite) set $\mathcal{M}$ of models under consideration and a prior $\pi(\theta|M)$ on the parameters of each model and then computes $\pi(M|\mathbf{y}^{obs})$, the **posterior probabilities of the models given the data** .

By Bayes' formula

$$\pi(M|\mathbf{y}^{obs}) = \frac{\pi(M) \int_{\Theta(M)} \pi(\theta|M) f(\mathbf{y}^{obs}|\theta, M) d\theta}{\sum_{M'} \pi(M') \int_{\Theta(M')} \pi(\theta|M') f(\mathbf{y}^{obs}|\theta, M') d\theta}$$

The model which maximizes $\pi(M|\mathbf{y}^{obs})$ is being selected. For prediction, one can also **average over the different models** with weights $\pi(M|\mathbf{y}^{obs})$.

The **Bayes factors**

$$\frac{\pi(M|\mathbf{y}^{obs})}{\pi(M'|\mathbf{y}^{obs})} : \frac{\pi(M)}{\pi(M')} = \frac{\int_{\Theta(M)} \pi(\theta|M) f(\mathbf{y}^{obs}|\theta, M) d\theta}{\int_{\Theta(M')} \pi(\theta|M') f(\mathbf{y}^{obs}|\theta, M') d\theta}.$$

are independent of the priors and measure the relative strength of evidence in the data of model M over model M'.

The main problem is the evaluation of the integral

$$\int_{\Theta(M)} \pi(\theta|M) f(\mathbf{y}^{obs}|\theta, M) d\theta = f(\mathbf{y}^{obs}|M)$$

which is nothing else than the normalizing constant in Bayes formula. In most applications it is not available in closed form.

Improper uninformative priors create additional difficulties because then the integral does not exist.

4.2 Laplace approximation and BIC

We can approximate the log of the integrand in $f(\mathbf{y}^{obs}|M)$ by a quadratic around its maximum $\tilde{\theta}(\mathbf{y}^{obs})$, the mode of the posterior. This gives

$$f(\mathbf{y}^{obs}|M) \approx \mathrm{const.}\pi(\tilde{\theta}(\mathbf{y}^{obs})|M)\, f(\mathbf{y}^{obs}|\tilde{\theta}(\mathbf{y}^{obs}), M)\, |\det(H)|^{-1/2}$$

where H is minus the second derivative of the log posterior, evaluated at $\tilde{\theta}$.

Asymptotically, $\tilde{\theta}(\mathbf{y}^{obs})$ is equivalent to the MLE $\widehat{\theta}(\mathbf{y}^{obs})$ and H is approximately equal to the observed Fisher information which is proportional to the sample size n. Dropping all terms of constant order, we obtain

$$\log f(\mathbf{y}^{obs}|M) \approx \log f(\mathbf{y}^{obs}|\widehat{\theta}(\mathbf{y}^{obs})) - \frac{p(M)}{2}\log(n).$$

This gives the model selection criterion of Schwartz, usually called Bayesian Information Criterion (BIC), which has a substantially larger penalty than AIC.

4.3 Markov chain Monte Carlo methods

Today, the most widely discussed methods approximate $f(\mathbf{y}^{obs}|M)$ by **Markov chain Monte Carlo methods**.

If one runs separate Markov chains for each model, the normalizing constants are not obtained automatically. For this reason, DIC was proposed.

With reversible jump MCMC, we can estimate $\pi(M|\mathbf{y}^{obs})$ by counting how often the chain visit the model M.

5. Selecting covariates in regression

Data generating distribution: $Y_i = f(\mathbf{x}_i) + \varepsilon_i$
with ε_i independent, $\mathrm{E}[\varepsilon_i] = 0, \mathrm{E}[\varepsilon_i^2] = \tau^2$.

Model $M \subset \{1, 2, \ldots m\}$

$$Y_i = \sum_{j \in M} \beta_j x_i^{(j)} + \varepsilon_i, \quad \varepsilon_i \ i.i.d. \sim \mathcal{N}(0, \sigma^2).$$

The MLE for β in model M is the least squares estimator:

$$\widehat{\beta}_M = \arg \min_{\beta_i = 0, i \notin M} \|\mathbf{y} - X\beta\|_2^2 .$$

5.1 C_p, AIC, BIC, RIC, ...

If σ^2 is assumed to be known, we have

$$\text{AIC} = \frac{||\mathbf{y} - X\widehat{\beta}_M||_2^2}{\sigma^2} + 2|M|$$

This is up to a constant equal to Mallows' C_p and Akaike's final prediction error which provide unbiased estimates for the error in predicting new observations Y_i' at the same values $\mathbf{x}_i$.

If σ^2 is unknown, then C_p plugs in an estimate $\widehat{\sigma}^2$ common to all models, obtained either from the full model or from a nonparametric residual variance estimator.

With σ^2 unknown, we have

$$\text{AIC} = n \log \widehat{\sigma}^2_M + 2|M| = n \log \frac{||\mathbf{y} - X\widehat{\beta}_M||_2^2}{n} + 2|M|.$$

The differences are usually minor because for good models $\widehat{\sigma}^2_M \approx \widehat{\sigma}^2$ and thus

$$\log \widehat{\sigma}^2_M \approx \log \widehat{\sigma}^2 + \frac{||\mathbf{y} - X\widehat{\beta}_M||_2^2}{n\widehat{\sigma}^2} - 1.$$

There are criteria with other penalties. BIC has $\log n \cdot |M|$.

If we assume that the data were generated from the model, then we can compute the bias exactly. One obtains a corrected AIC with penalty

$$2\frac{n(|M|+1)}{n-|M|-2} \approx 2(|M| + |M|^2/n).$$

The risk inflation criterion (RIC) of Foster and George has penalty $2\log p \cdot |M|$. This is interesting because it depends on the number of models under consideration which is justified both intuitively and rigorously (Birgé and Massart).

5.2 Lasso, Lars and other algorithms

If p is large, it is difficult to compute the best model

$$\widehat{M} = \arg\min_M \left(\frac{||\mathbf{y} - X\widehat{\beta}_M||_2^2}{\widehat{\sigma}^2} + \lambda|M| \right)$$

(the value of λ depends on the criterion used).

The Lasso replaces the penalty $|M| = \|\beta_M\|_0$ by the L_1-norm, i.e. one minimizes

$$||\mathbf{y} - X\beta||_2^2 + \lambda||\beta||_1$$

over all β. For large values of λ, some coefficients β_j of the solution are exactly zero. Hence by selecting λ, we effectively perform model selection.

The criterion of the Lasso is convex, and the solution can be found by a quadratic programming algorithm. Least angle regression (LARS) is a very efficient algorithm to compute the Lasso solutions for all λ.

The relation between the models selected by Lasso and by minimizing AIC or BIC is however unclear.

A recent preprint of Candes and Tao consider the following similar method:

$$\widehat{\beta} = \arg\min\{||\beta||_1 \mid ||X^T(\mathbf{y} - X\beta)||_\infty \leq \lambda\sigma\}.$$

Again, this is convex and can be computed by linear programming. They give conditions on X and λ such that with high probability

$$||\widehat{\beta} - \beta||_2^2 \leq C\lambda \left(\sigma^2 + \sum_j \min(\beta_j^2, \sigma^2)\right).$$

6. Discussion

The number of proposals discussed here is already large, and there are more variants in the literature. The most important things to remember are AIC and BIC (minus log likelihood plus p and $\frac{p}{2}\log(n)$ respectively) and cross validation.

BIC has been shown to be consistent if the data are generated by one model with fixed dimension p, whereas AIC tends to overestimate the dimension in this case. On the other hand, AIC tends to do better for prediction.

An objective yardstick is helpful when a variety of models are tried on the same data. It is also necessary if one has to analyze many similar data sets in an automatic manner.

The development of model selection criteria has led to interesting ideas and deep theories, connecting statistics with information theory.

6.1 Some more critical points

Model selection supposes that the set of all models that will be considered is fixed in advance. It does not reflect the iterative nature of data analysis which develops new models from deficiencies of previous ones. It is not a complete substitute for assessing the generalizability of a selected model from independent test data.

In many applications, the idea of a single “best” model can be misleading.

The problem of quantifying uncertainty about the conclusions after model selection has been made is largely unsolved. Recent results by Pötscher indicate that this is very difficult. Presumably, one needs independent test data for this.