

Statistical Model Selection

Alex Potapov

January 19, 2009 Lab meeting

Motivation

Bythotrephes invasion in Ontario lakes. Spread with boaters.

Gravity model for boaters spread. Need a “portable” model, fitted for one region but applicable at another. **Not just predict**, but **predict for other lake systems**.

GM: average number of boaters travelling between lakes i and j

$$\lambda_{ij} = f(A_i, A_j, d_{ij}, \dots), \text{ e.g. } \lambda_{ij} = C \frac{A_i A_j}{d_{ij}^2}, \lambda_{ij} = C(\ln A_i)(\ln A_j) \exp(-\beta d_{ij})$$

Observed data assumed to be $n_{ij} \sim \text{Poisson}(\lambda_{ij})$

Standard approach: generalized linear model, power/exp terms

$$\lambda = \exp(a_0 + \sum a_i x_i); \quad \text{if } x_i = \ln A \Rightarrow \lambda \sim A^{a_i}$$

But: we are not sure, what variables and what terms form the best model

Should $\Delta\text{AIC}=14$ or 10 be ignored or not? Why? Basis?

A_i — lake i area, P_i — lake i perimeter, d_{ij} — distance between lakes i and j

#	Best Model	AIC	ΔAIC	RSS
0	$\lambda = C$	9563.9		3914.5
1	$\lambda = C (A_i A_j)^{0.3}$	7543.8	2020.1	3392.5
2	$\lambda = C (A_i A_j)^{0.3} d_{ij}^{-1.17}$	6120.7	1423.8	2466.2
3	$\lambda = C \frac{(A_i A_j)^{0.53}}{d_{ij}^{-1.19}} \exp\left(-2.00 \frac{A_i + A_j}{A_{\max}}\right)$	5894.0	226.7	2097.2
4	$\lambda = C \frac{(A_i A_j)^{0.56}}{d_{ij}^{-1.18}} \exp\left(-1.86 \frac{A_i + A_j}{A_{\max}} - 0.49 \frac{P_i + P_j}{P_{\max}}\right)$	5883.4	10.6	2059.6
5	$\lambda = C \frac{(A_i A_j)^{0.66}}{(\ln A_i \ln A_j)^{0.37} d_{ij}^{1.19}} \exp\left(-2.20 \frac{A_i + A_j}{A_{\max}} - 0.58 \frac{P_i + P_j}{P_{\max}}\right)$	5870.3	13.1	2114.1
6	$\lambda = C \frac{(A_i A_j)^{0.66} (P_i P_j)^{0.17}}{(\ln A_i \ln A_j)^{0.61} d_{ij}^{1.19}} \exp\left(-2.03 \frac{A_i + A_j}{A_{\max}} - 1.08 \frac{P_i + P_j}{P_{\max}}\right)$	5856.2	14.2	2100.6
7	$\lambda = C \frac{(A_i A_j)^{0.66} (P_i P_j)^{0.17}}{(\ln A_i \ln A_j)^{0.60} d_{ij}^{1.12}} \exp\left(-2.03 \frac{A_i + A_j}{A_{\max}} - 1.03 \frac{P_i + P_j}{P_{\max}} - 0.68 \frac{d_{ij}}{d_{\max}}\right)$	5854.0	2.2	2087.3

Statistical models

- **Data:** random values y_i and associated values of “regressor variables” \mathbf{x}_i , k -dimensional vectors, usually non-random.
- We want to be able to **make predictions** about y .
- **Statistical model:** probability distribution depending on \mathbf{x}_i and on some parameters:

$$y \sim \text{Prob. distr.} \left(\underbrace{f(x, \text{parameters}_1)}_{\text{structural part}}, \underbrace{\text{parameters}_2}_{\text{distributional parameters}} \right)$$

Examples of statistical models

- a) $y_i = f(\mathbf{x}_i, \theta) + \varepsilon_i$, or $y_i \sim N(f(\mathbf{x}_i, \theta), \sigma^2)$
 $f(\mathbf{x}_i, \theta)$ – deterministic component (structural part)
 σ^2 – parameter characterizing random component
- b) $y_i \sim \text{Poisson}(\lambda(\mathbf{x}_i, \theta))$ (e.g. intensity of boater flow)
- c) $y_i \sim \text{Bernoulli}(q(\mathbf{x}_i, \theta))$ (e.g. probability of lake invasion)
- d) $y_i \sim N(\mu, \sigma^2)$ – not interesting in this talk, no structural part

Problems with models

Nobody knows the true distribution for y , and even whether it is truly random!

George E. P. Box: “Essentially, all models are wrong, but some are useful”

Model selection task: how to choose a **most useful model** from a set of candidates? What does “useful” mean?

Model parsimony: if a simple and a complex model work similarly, one have to prefer the simple one.

Problems with models

Main problem in model selection: to decide whether the more complicated model performs sufficiently better to be accepted.

a) How to **measure** the model performance?

b) How to **draw conclusions** from these measurements.

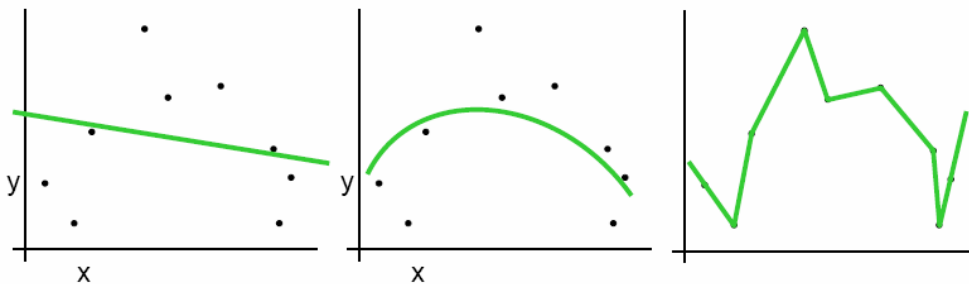
My impression:

for (a) – there are well elaborated standard techniques;

for (b) – there are always exceptions, often it is a **mixture of art and science**.

Will discuss in the end, why it is so.

So, model selection is a big problem...



Can one rely upon **only** quantitative criteria?

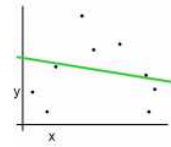


Minimum Residual Sum of Squares

$$\text{RSS} = \sum_{i=1}^n [y_i - f(x_i, \theta)]^2 = \min$$

Underfitting : too simple model,

$(f(x_i, \theta))$ does not approximate deterministic part of y



Overfitting : too complicated model, approximates part of noise



1) "Manual parsimony" – predefined model complexity

2) Regularization methods, e.g. $\min \left\{ \sum_{i=1}^n [y_i - f(x_i, \theta)]^2 + \lambda (\nabla f)^2 \right\}$,

no rules for automatic choice of λ (art)

Maximum Likelihood

Model for y : probability/density $p(y, \mathbf{x}, \theta)$

Likelihood : probability of the data for the given model

assuming data independent : $L(\theta) = \prod_i p(y_i, \mathbf{x}_i, \theta)$

To avoid too small values use log - likelihood : $l(\theta) = \sum_i \ln p(y_i, \mathbf{x}_i, \theta)$

Maximum likelihood estimate : $\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} l(\theta)$

Maximum Likelihood

Examples :

$$\text{Normal} \quad l(\theta | f) = \sum_i \left[-\frac{(y_i - f(x_i, \theta))^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right]$$

$$\text{Poisson} \quad l(\theta | f) = \sum_i \ln \left[\frac{\lambda(x_i, \theta)^{y_i}}{y_i!} e^{-\lambda(x_i, \theta)} \right]$$

ML is most popular method to find parameters for the given model.

Usually no account for parsimony in model comparison:
likelihood increases as the model becomes more complicated.

Exception: nested models and likelihood ratio test

Maximum Likelihood: nested models

Two models with the same distribution and f ,

one with $k_1 + k_2$ parameters : (θ_1, θ_2) , all to be determined

the other with k_1 parameters : $(\theta_1, \theta_2 = 0)$, only θ_1 to be determined

Example : Poisson models with $\lambda(\mathbf{x}, \theta) = \exp(\theta_0 + \theta_1 x + \theta_2 u + \theta_3 v)$

Model 1 : λ depends on $\{x, u, v\}$; $k_1 = 2, k_2 = 2$

Model 2 : λ depends only on $\{x\}$, $\theta_2 = \theta_3 = 0$

$l_1(\theta_0, \theta_1, \theta_2, \theta_3) > l_2(\theta_0, \theta_1, 0, 0)$, but how significant is the difference?

Hypothesis testing (H_0 vs H_1) or Likelihood Ratio Test ($l_1 - l_2 \sim \chi_{k_2}^2$).

Limitations : cannot distinguish e.g. between models with $\{x, u\}$ and $\{x, v\}$.

Entropy-based approaches: AIC

Let there be two probability distributions :

a) data according to unknown $q(y)$

b) model with assumed $p(y, \theta)$

$$\text{Entropy: } H[q, q] = -\int q \ln q \, dy = -\langle \ln q \rangle$$

What if we try to calculate

$$H[q, p] = -\int q \ln p \, dy = -\langle \ln p \rangle \approx -\frac{1}{n} \sum_{i=1}^n \ln p(y_i, \theta) = -\frac{1}{n} l(\theta)$$

Looks like negative likelihood, but interpretation is different!!!

Entropy-based approaches: AIC

Optimization problem : find p for which $H[q, p] = \min$

$$-\int q \ln p \, dy + \lambda \int p \, dy = \min \Rightarrow \int \left(\lambda - \frac{q}{p} \right) \delta p \, dy = 0,$$

$$\Rightarrow \lambda p = q \Rightarrow p = q; \quad \frac{d^2(-\ln p)}{dp^2} > 0, \Rightarrow \min (\sim \max L(\theta))$$

\Rightarrow may introduce measure of closeness of p to q ("distance") :

$$D_{KL} = H[q, p] - H[q, q] \geq 0.$$

Kullback – Leibler relative entropy.

Entropy - based selection criterion : take the model

with $\min D_{KL}$ or $\min H[q, p] = \min \int q(y) [-\ln p(y, \theta)] \, dy$

No need in nesting. But **how** to find $H[q, p]$?

Entropy-based approaches: AIC

$$H[q, p] = -\langle \ln p \rangle \approx -\frac{1}{n} \sum_{i=1}^n \ln p(y_i, \theta) = -\frac{1}{n} l(\theta)$$

$\min H[q, p]$ is approximately equivalent to $\max l(\theta)$.

Akaike (1974): $-\frac{1}{n} l(\theta)$ is a biased estimate of $H[q(y), p(y, \theta)]$

More accurate estimate is

$$H[q, p] \approx -\frac{1}{n} l(\theta) + \frac{k}{n}, \quad k = \text{the \# of parameters}$$

Akaike information criterion : take the model with **minimum** of

$$\text{AIC} = -2l(\theta) + 2k \approx 2nH[q, p]$$

Explicit parsimony term $+2k$ penalizes for using more parameters

AIC practice

Burnham & Anderson (2001): rule of thumb for model selection

Let there be 2 models:

model 1, simpler, AIC_1 greater;

model 2, more complicated, AIC_2 smaller, $< \text{AIC}_1$;

If $\text{AIC}_1 - \text{AIC}_2 = 0$ to 2 : models are practically equivalent

If $\text{AIC}_1 - \text{AIC}_2 = 4$ to 7 : model 2 is probably better

If $\text{AIC}_1 - \text{AIC}_2 > 8$: model 1 probably should be discarded

However...

AIC practice

...However, let there be two models and a data set of size n

$$l_1/n \rightarrow 0.20 \approx H_1, \quad k_1 = 2, \quad AIC_1 \approx 0.20n + 2k_1$$

$$l_2/n \rightarrow 0.18 \approx H_2, \quad k_2 = 4, \quad AIC_2 \approx 0.18n + 2k_2$$

Then

n	AIC_1	AIC_2	selection
10	6	9.8	1
1000	204	188	2

For small n AIC favors simple models,

For big n AIC does almost as ML and favors more complicated models.

Hypothesis: AIC is optimal for “soft science” with models chosen for good prediction rather than discovery of the truth

(Stone 1997, Ghosh & Samanta 2001)

In experiments AIC works better for complicated true models

(Burnham & Anderson 2004):

Cross-Validation

Possible source of selection problems: we estimate and test on the same data set.

For better results estimating and testing must be done on different data.

Cross-validation: data are split into calibration set and test set.

But: for small test set unsatisfactory testing, for big test set many data are lost for estimating.

Cross-Validation

Compromise: “**Leave One Out**”:

- estimate on $n-1$ points, validate on 1 points
- repeat n times validating on all n points
- average (or sum up) all validation results

For big n too many computations. Then use m points at a time.

If one uses likelihood for validating, then Leave One Out is asymptotically equivalent to AIC: for big n

$$(\text{Validated log-likelihood}) \approx (\text{log-likelihood}) - k \quad (\text{Stone, 1977})$$

Sometimes CV is very useful, however one may expect the same shortcomings as for AIC: oriented for prediction rather than “discovering the truth”.

Bayesian Model Selection

Methods work well. According to reported **experiments**, Bayesian selection provides the **highest rate of correct detection of true models**.

BUT. It uses “probability of parameter value” and “probability of model”.
What does it mean?

“**Frequentists**”: probability is the measure of frequency for a certain outcome of a repeated experiment. Parameter is fixed and there is no probability of it taking a certain value. True model is only one, and there is no probability for a model to be true. Hence, foundations for Bayesian selection are missing.

“**Bayesians**”: probability may reflect uncertainty of our knowledge as well. Since we are not sure about parameters and models, we may assign probabilities to them. It is just necessary to find the right way of such an assignment. Experiment reduces uncertainty and posterior probability is “more peaked” than the prior.

Interesting analogy with wave function in quantum mechanics. It cannot be measured, one can only observe the consequences of its existence. How should it be interpreted? Three versions: a) Copenhagen (almost Bayesian); b) Multi-world (“objective” but crazy), c) It is just an efficient numerical procedure.

Lot of room for philosophical discussions...

Bayesian Model Selection

Based upon the total probability formula :

$$P(\text{data, model}) = \underbrace{P(\text{model} | \text{data})}_{\text{selection criterion}} \underbrace{P(\text{data})}_{\substack{\text{does not} \\ \text{depend} \\ \text{on model}}} = \underbrace{P(\text{data} | \text{model})}_{\text{likelihood}} \underbrace{P(\text{model})}_{\text{prior probability}}$$

Or Bayes formula :

$$P(\text{model}_i | \text{data}) = \frac{P(\text{data} | \text{model}_i)P(\text{model}_i)}{\sum_k P(\text{data} | \text{model}_k)P(\text{model}_k)}$$

The result is a probability of parameter value or a probability of a certain model.

Model/parameter selection: choose one with the greatest probability.

Bayesian Model Selection

Selection tool: likelihood averaged over prior distribution of parameters

$$L_{A_i}(\text{data, model}_i, p_{A_i}) = \int \underbrace{L_i(\theta | \text{data, model}_i)}_{\text{usual likelihood}} \underbrace{p_{A_i}(\theta)}_{\text{prior}} d\theta$$

If prior probability of models π_i are equal, one selects the model with the greatest $L_{A_i}(\text{data, model}_i, p_{A_i})$.

Schwartz (1978): let $n \rightarrow \infty$, $\hat{\theta}$ is ML estimate, then

$$L_{A_i}(\text{data, model}_i, p_{A_i}) \approx L_i(\hat{\theta} | \text{data, model}_i) \times p_{A_i}(\hat{\theta}) \times \left(\sqrt{\frac{2\pi}{n}} \right)^{k_i},$$

$$\ln L_{A_i} \approx l(\hat{\theta}) - \frac{1}{2} k_i \ln n + \underbrace{\ln p_{A_i}(\hat{\theta}) + \frac{1}{2} \ln(2\pi)}_{\text{asymptotically negligible}} + o(1)$$

Bayesian Information Criterion

$$\text{BIC} = -2l_i(\hat{\theta}) + k_i \ln n$$

Prior distribution asymptotically disappears. For $n > 8$ there is much stronger penalizing of model complexity than in AIC case

Experiments: Burnham & Anderson (2004), Ghosh & Samanta (2001)

Conclusions:

G&S: AIC better selects models for predictions, BIC is better for “discovering the truth”

B&A: BIC better detects models with a few components (“effects”, variables), AIC better detects complicated models with many tapering effects. Model averaging gives the best prediction results.

Who's fault is it? What to do?

two major “Russian questions” :))

It appears that there is no single method that allows successful “blind model selection”, but there are many methods giving reasonable results in certain situations.

- Why?
- What a poor researcher should do?
- Is one allowed to use Bayesian methods of model selection without trusting in priors?

Hypothesis (my): *model selection is generically an ill-posed problem*

ILL-POSED problems: what are they?

Jaques Hadamard (1902): well-posed problem: solution **exists**, it is **unique**, and it **continuously** depends on input data. Belief: mathematics should study only well-posed problems.

Mid 20 century: ill-posed problems are very important for practice.

Typical scheme where they arise (inverse problems):

Source signal \Rightarrow **probed object** \Rightarrow **measured modified signal**

Problem: restore structure of the object from the measured signal.

Examples: computer tomography, probing structure of Earth etc.

Case of statistical data analysis: some resemblance

Source of randomness \Rightarrow **system structure** \Rightarrow **measured data**

Problem: restore the system structure from the data

ILL-POSED problems: how to deal with them?

Anrei Tikhonov (1943): **regularization** of an ill-posed problem.

Replace the unsolvable problem by a solvable one, which is “close” in a certain sense.

Example: minimizing RSS

$$\min \left\{ \underbrace{\sum_{i=1}^n [y_i - f(x_i, \Theta)]^2}_{\text{original problem}} + \underbrace{\lambda (\nabla f)^2}_{\text{regularization term}} \right\},$$

Typical regularization approaches:

- Replace an infinite-dimensional problem with a finite-dimensional (in our case: fix function and determine parameters) (continuity)
- Add special terms that penalize lack of smoothness (similar to penalizing model complexity). Magnitude of such terms should decrease as noise decreases. (existence & uniqueness)

There is no “universal” way of regularization for all ill-posed problems, only “universal ideas”. Hard to formalize “closeness”.

Bayesian selection = regularization?

Hypothesis: *Bayesian methods with special choices of priors may be considered just as a good regularization of the original problem.*

Such interpretation allows anyone to use them without going into philosophical debates about prior distributions.

Empirical evidence on model selection:

- If one needs just a good predictor for the same system, then optimal may be AIC + model averaging (Burnham & Anderson, 2001)
- If one needs to recover an interpretable model (“truth”), BIC is better for simple models with a few variables (Burnham & Anderson, 2004), otherwise elaborated Bayesian techniques may be much more efficient (Ghosh & Samanta, 2001).
- Other less formalized methods of regularization may work as well. For example, discard models with non-monotonous dependencies, oscillations etc.

Conclusions

1. If one needs to compare 2-3 models, usually simultaneous application of several conventional methods (AIC, LRT, CV) is enough . No problem!
2. Problems arise when one considers a series of models of increasing complexity and needs to decide when to stop. Then choice of a proper method and a certain regularization may be necessary.
 - a) Clearly define a problem formulation: what is needed.
 - b) Try to minimize the number of candidate models basing upon mechanistic arguments and general concepts about model structure.
 - c) Bayesian methods that have shown **good performance in tests** should be seriously considered: they may be considered just as a way of problem regularization. This allows one to avoid philosophical disputes. In particular, BIC may be useful.
3. Human intelligence is indispensable!!! If you have a good reason not to trust AIC/BIC/etc, rely upon less formal criteria!!!

Major References

- Akaike H., 1974. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 716-723.
- Burnham K.P., Anderson D.R., 2001. *Model selection and multimodel inference. A practical information-theoretic approach*. 2nd ed., Springer, N.Y.
- Burnham K.P., Anderson D.R., 2004. Multimodel inference. Understanding AIC and BIC in model selection. *Sociological methods & research* 33, 261-304.
- Forster M.R., 2000. Key concepts in model selection: performance and generalizability. *J. Math. Psychology* 44, 205-231.
- Ghosh J.K., Samanta T., 2001. Model selection – an overview. *Current Science* 80, No. 9-10, 1135-1144.
- Stone M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *J. Roy. Stat. Soc. B* 39, 44-47.
- Stone M., 1997. Comment to: J. Shao, An asymptotic theory for linear model selection. *Statistica Sinica* 7(1997), 221-264
- Schwarz G., 1978. Estimating the dimension of a model. *Ann. Stat.* 6, 461-464.
- Zucchini W., 2000. An introduction to Model selection. *J. Math. Psychology* 44, 41-61.