Model Selection

BEE 6940 Lecture 12

April 17, 2023

Table of Contents

Review: Model Assessment

Is Our Model Appropriate?

This means that are a large number of models under consideration.

In general, we are in an **$\mathcal{M}$-open setting**: no model is the "true" data-generating model, so we want to pick a model which performs well enough for the intended purpose.

The contrast to this is **$\mathcal{M}$-closed**, in which one of the models under consideration is the "true" data-generating model, and we would like to recover it.

Model Assessment and Hypothesis Testing

Evaluating the relative skill of different models can be thought of as a generalization of null hypothesis-testing.

- Can embed more nuanced and specific hypotheses;
- Compare proposed data-generating processes, instead of just comparing a "null" and an "alternative" under null assumptions.

Model Assessment Criteria

How do we assess models?

Generally, through **predictive performance**: how probable is some data (out-of-sample or the calibration dataset)?

Parsimony and Model Selection

What Is The Goal of Model Selection?

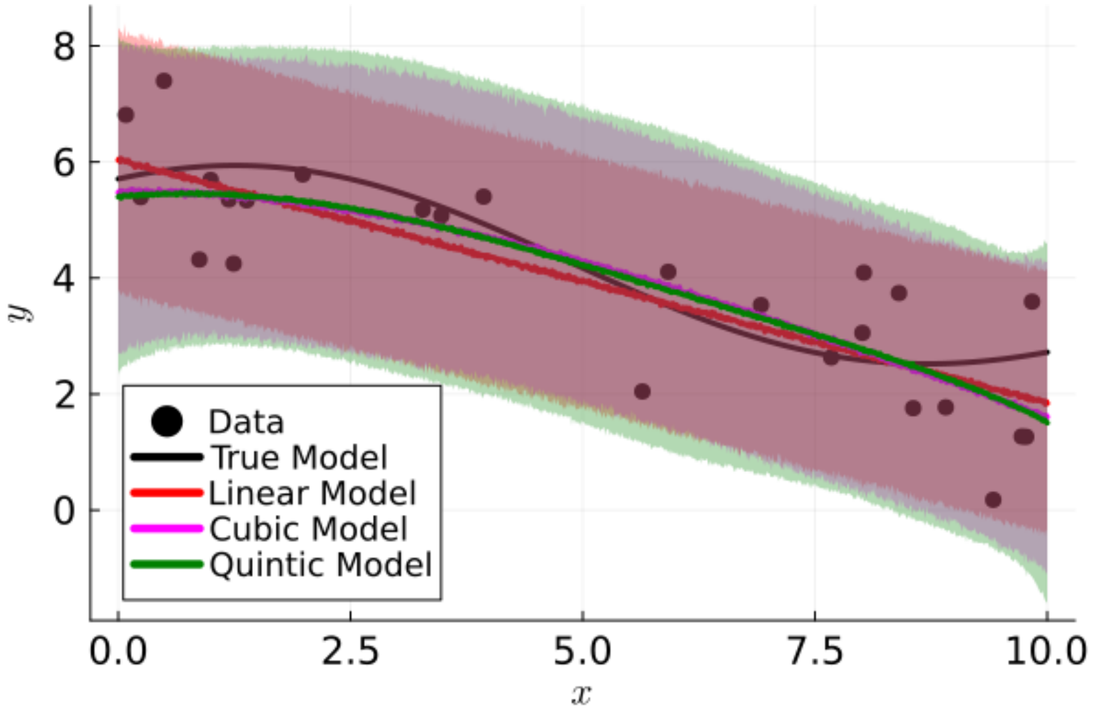
Key Idea: Model selection consists of navigating the bias-variance tradeoff.

Model error (*e.g.* RMSE) is a combination of *irreducible error*, *bias*, and *variance*.

- Bias can come from under-dispersion (too little complexity) or neglected processes;
- Variance can come from over-dispersion (too much complexity) or poor identifiability.

Bias-Variance Tradeoff

If either of these is high, the model's predictive ability will be poor!



Model "Complexity" and Bias/Variance

Model complexity is *not* necessarily the same as the number of parameters.

Sometimes processes in the model can compensate for each other, which can help improve the representation of the dynamics and reduce error/uncertainty even when additional parameters are included.

Occam's Razor

Occam's Razor:

> *Entities are not to be multiplied without necessity.*

- Credited to William of Ockham
- Appears much earlier in the works of Maimonides, Ptolemy, and Aristotle
- First formulated as such by John Punch (1639)

"Zebra" Principle

More colloquially:

> *When you hear hoofbeats, think of horses, not zebras.*
>
> *— Theodore Woodward*

Aside: What Error To Use?

There are many metrics we can use to assess error.

Common Framework: Specify a *loss function* which penalizes based on deviation between prediction (point or probabilistic) and the observation.

- Zero-One loss
- Logarithmic loss
- Quadratic loss

Aside: What Error To Use?

Can also think of inference in terms of loss-minimization relative to "true" parameter values, *e.g.*:

- the posterior mode minimizes the zero-one loss.
- the posterior median minimizes the linear loss
- the posterior mean minimizes the quadratic loss.

Point Estimate Error Functions

For point estimates, measures associated with loss functions are called "scoring functions" (overview: Gneiting (2011)):

$$\bar{S} = \frac{1}{n} \sum_{i=1}^{n} S(x_i, y_i)$$

For example, the squared scoring function results in the mean-squared error.

Probabilistic Forecasts

For probabilistic estimates, these are called "scoring rules" (overview: Gneiting & Raftery (2007)) and are based on the probability assigned to the observed event by the forecast:

- Quadratic score
- Logarithmic score
- Continuous Ranked Probability Score (CRPS)

Log-Likelihood as Predictive Fit Measure

The measure of predictive fit that we will use is the **log predictive density** or **log-likelihood** of a replicated data point/set, $p(y^{rep}|\theta)$.

For normally-distributed data, this is proportional to the mean-squared error.

Log-Likelihood as Predictive Fit Measure

Why use the log-likelihood density instead of the log-posterior?

- The likelihood captures the data-generating process;
- The posterior includes the prior, which is only relevant for parameter estimation.

Important: This means that the prior is still relevant in predictive model assessment, and should be thought of as part of the model structure!

Cross-Validation

The "gold standard" way to test for predictive performance is **cross-validation**:

1. Split data into training/testing sets;
2. Calibrate model to training set;
3. Check for predictive ability on testing set.

Cross-Validation

Leave-One-Out Cross-Validation: Drop one value, refit model on rest of data, check for prediction.

This is related to the posterior predictive distribution

$$p(y^{\text{rep}}|y) = \int_{\theta} p(y^{\text{rep}}|\theta)p(\theta|y)d\theta.$$

Cross-Validation

Leave-k-Out Cross-Validation: Drop k values, refit model on rest of data, check for prediction.

As $k \to n$, this reduces to the prior predictive distribution

$$p(y^{\text{rep}}) = \int_{\theta} p(y^{\text{rep}}|\theta)p(\theta)d\theta,$$

which is also the marginal likelihood of the model.

Cross-Validation

The problems:

- This can be very computationally expensive!
- We often don't have a lot of data for calibration, so holding some back can be a problem.
- How to divide data with spatial or temporal structure? This can be addressed by partitioning the data more cleverly (*e.g.* leaving out future observations), but makes the data problem worse.

Expected Out-Of-Sample Predictive Accuracy

Instead, we will try to compute the *expected out-of-sample predictive accuracy*.

The out-of-sample predictive fit of a new data point $\tilde{y}_i$ is

$$\begin{aligned}\log p_{\text{post}}(\tilde{y}_i) &= \log \mathbb{E}_{\text{post}}\left[p(\tilde{y}_i|\theta)\right] \\ &= \log \int p(\tilde{y}_i|\theta) p_{\text{post}}(\theta)\, d\theta.\end{aligned}$$

Expected Out-Of-Sample Predictive Accuracy

However, the out-of-sample data $\tilde{y}_i$ is itself unknown, so we need to compute the *expected out-of-sample log-predictive density*

$$\begin{aligned}
\text{elpd} &= \text{expected log-predictive density for } \tilde{y}_i \\
&= \mathbb{E}_P \left[\log p_{\text{post}}(\tilde{y}_i) \right] \\
&= \int \log \left(p_{\text{post}}(\tilde{y}_i) \right) P(\tilde{y}_i) \, d\tilde{y}.
\end{aligned}$$

Expected Out-Of-Sample Predictive Accuracy

But we don't know the "true" distribution of new data P!

We need some measure of the error induced by using an approximating distribution Q from some model.

Kullback-Leibler Divergence

Kullback-Leibler Divergence

Suppose a distribution P denotes "reality" and Q is an approximating distribution.

Kullback-Leibler divergence:

- Is a measure of the deviation between P and Q;
- Has a connection to information theory (average number of bits to re-encode samples from P using a Q-code);
- Can be interpreted as the "surprise" when using Q as an approximation to P.

Kullback-Leibler Divergence

Formula for K-L Divergence:

$$D_{KL}(P||Q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Note that D_{KL} is a *divergence*, not a *distance*, as it is not symmetric.

K-L Divergence and Scoring Functions

D_{KL} is the divergence resulting from the logarithmic scoring function.

In other words, it is the natural measure of model error when using the log-predictive density.

Computing K-L Divergence

However, computing K-L divergence is fraught for model selection: we don't actually know the "true" data generating model.

Information Criteria

Information Criteria Overview

"Information criteria" refers to a category of estimators of prediction error.

The idea: estimate predictive error without having access to the "true" model P.

Information Criteria Overview

There is a common framework for all of these:

If we compute the expected log-predictive density for the existing data $p(y|\theta)$, this will be too good of a fit and will overestimate the predictive skill for new data.

We can adjust for that bias by correcting for the *effective number of parameters*, which can be thought of as the expected degrees of freedom in a model contributing to overfitting.

Akaike Information Criterion

The "first" information criterion that most people see is the Akaike Information Criterion (AIC).

This uses a point estimate (the maximum-likelihood estimate $\hat{\theta}_{\mathrm{MLE}}$) to compute the log-predictive density for the data, corrected by the number of parameters k:

$$\widehat{\mathrm{elpd}}_{\mathrm{AIC}} = \log p(y|\hat{\theta}_{\mathrm{MLE}}) - k.$$

Akaike Information Criterion

The AIC is defined as $-2\widehat{\mathrm{elpd}}_{\mathrm{AIC}}$ (for "historical" reasons; this is called the deviance scale).

Due to this convention, lower AICs are better (they correspond to a higher predictive skill).

Akaike Information Criterion

In the case of a normal model with independent and identically-distributed data and uniform priors, k is the "correct" bias term so that $\widehat{\text{elpd}}_{\text{AIC}}$ converges to the K-L divergence (there are corrections when the sample size is sufficiently small).

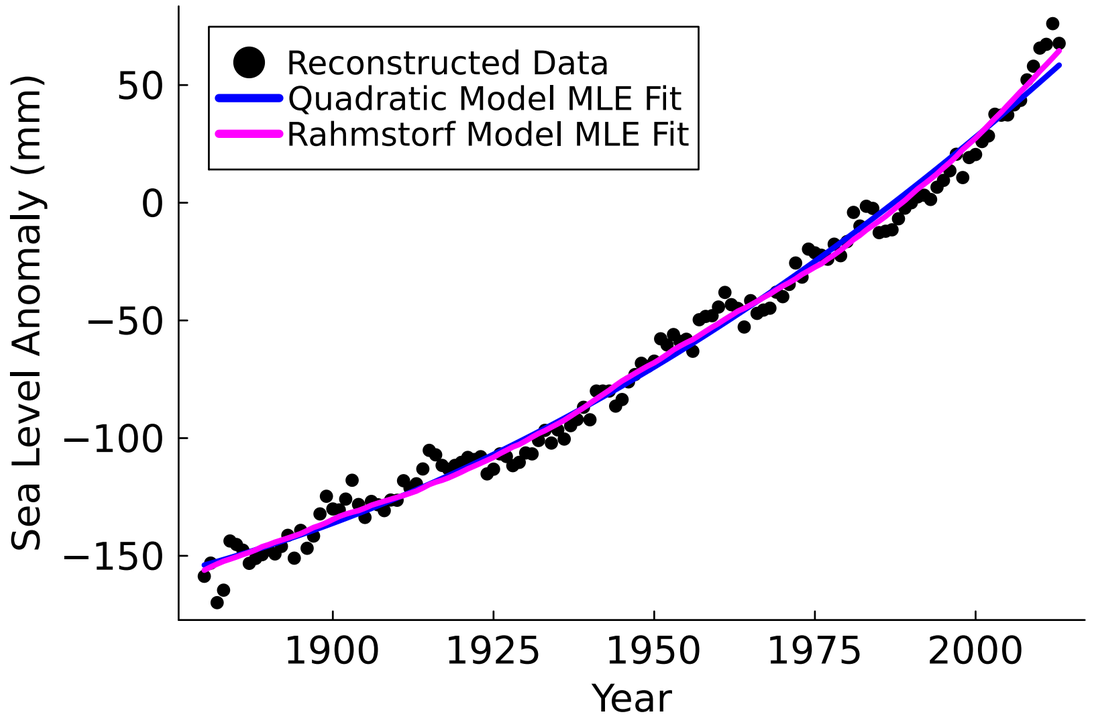
However, with more informative priors and/or hierarchical models, the bias correction k is no longer appropriate, as there is less "freedom" associated with each parameter.

AIC: Example with SLR data

Consider two SLR models:

- Quadratic
- Rahmstorf

We can think of these as alternative hypotheses about the influence of warming on SLR.



AIC: Example with SLR data

These both have four parameters (including the error variance), so $k = 4$, and the difference is in the log-likelihood of the MLE estimate.

- Quadratic AIC: 895
- Rahmstorf AIC: 864

AIC: Example with SLR data

The actual values here don't matter; what's important when comparing models within a set $\mathcal{M}$ are the differences

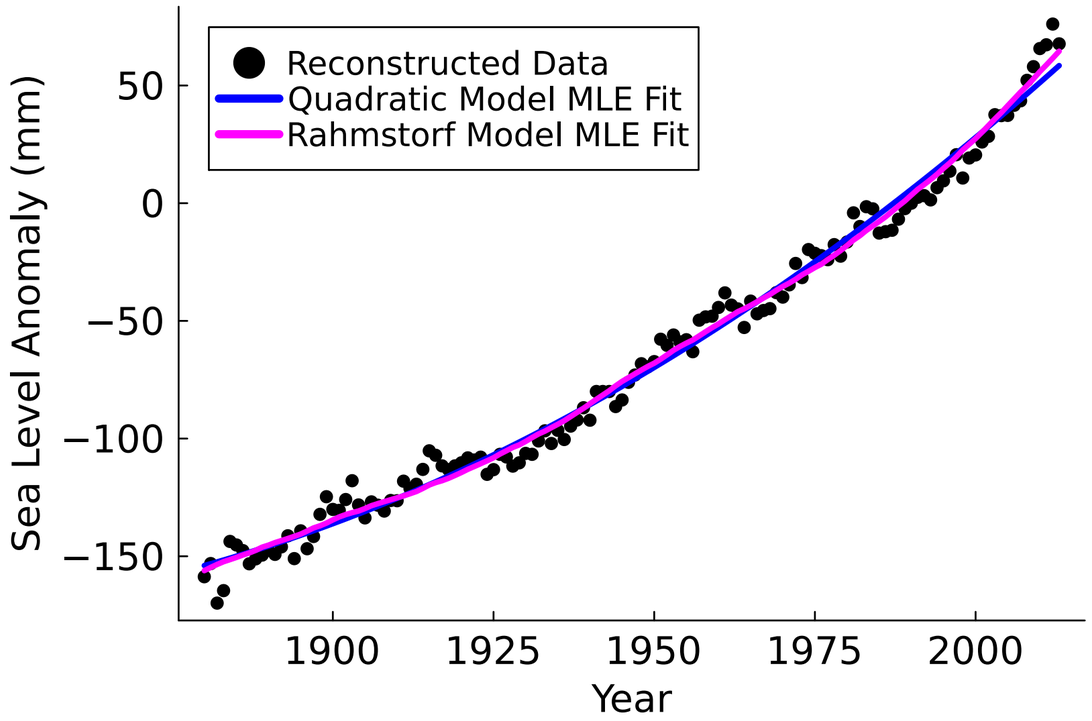
$$\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}.$$

Some basic rules of thumb (from Burnham & Anderson (2004)):

- $\Delta_i < 2$ means the model has "strong" support across $\mathcal{M}$;
- $4 < \Delta_i < 7$ suggests "less" support;
- $\Delta_i > 10$ suggests "weak" or "no" support.

AIC: Example with SLR data

So in this case, the quadratic model has weak support relative to the Rahmstorf model, which might be interpreted as supporting the hypothesis that SLR increases are related to temperature increases.



AIC and Model Averaging

$\exp(-\Delta_i/2)$ can be thought of as a measure of the likelihood of the model given the data y. The ratio

$$\exp(-\Delta_i/2)/\exp(-\Delta_j/2)$$

can approximate the relative evidence for M_i versus M_j.

AIC and Model Averaging

This gives rise to the idea of *Akaike weights*:

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{m=1}^{M} \exp(-\Delta_m/2)}.$$

Model projections can then be weighted based on w_i, which can be interpreted as the probability that M_i is the K-L minimizing model in $\mathcal{M}$.

Aside: Model Averaging vs. Selection

Model averaging can sometimes be beneficial vs. model selection, as model selection can introduce bias from the selection process (this is particularly acute for stepwise selection due to path-dependence).

Deviance Information Criterion

The Deviance Information Criterion (DIC) is a more Bayesian generalization of AIC which uses the posterior mean

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}\left[\theta|y\right]$$

and a bias correction derived from the data.

Deviance Information Criterion

$$\widehat{\mathrm{elpd}}_{\mathrm{DIC}} = \log p(y|\hat{\theta}_{\mathrm{Bayes}}) - p_{\mathrm{DIC}},$$

where

$$p_{\mathrm{DIC}} = 2\left(\log p(y|\hat{\theta}_{\mathrm{Bayes}}) - \mathbb{E}_{\mathrm{post}}\left[\log p(y|\theta)\right]\right).$$

Then, as with AIC,

$$\mathrm{DIC} = -2\widehat{\mathrm{elpd}}_{\mathrm{DIC}}.$$

DIC: Effective Number of Parameters

What is the meaning of p_{DIC}?

- The difference between the average log-likelihood (across parameters) and the log-likelihood at a parameter average measures "degrees of freedom".
- Note that p_{DIC} can be negative if the posterior mean is far from the posterior mode: this suggests that the model is poorly constrained and may do better on new data than the existing data.
- The DIC adjustment assumes independence of residuals for fixed θ.

AIC vs. DIC

AIC and DIC often give similar results, but don't have to. For example, for the SLR example, with relatively uninformative priors,

$$\Delta_{\text{DIC}} \approx 4,$$

which based on the AIC scale suggests limited (but stronger) evidence for the quadratic model.

The key difference is the impact of priors on parameter estimation and model degrees of freedom.

Convergence of AIC and DIC

Both AIC and DIC converge to the K-L divergence.

Also valuable: they both also converge to expected leave-one-out cross-validation.

Watanabe-Akaike Information Criterion (WAIC)

WAIC is a *fully* Bayesian generalization of AIC.

$$\widehat{\mathrm{elpd}}_{\mathrm{WAIC}} = \sum_{i=1}^{n} \log \int p(y_i|\theta) p_{\mathrm{post}}(\theta)\, d\theta - p_{\mathrm{WAIC}},$$

where

$$p_{\mathrm{WAIC}} = \sum_{i=1}^{n} \mathrm{Var}_{\mathrm{post}} \left(\log p(y_i|\theta)\right).$$

Watanabe-Akaike Information Criterion (WAIC)

p_{WAIC} is an estimate of the number of "unconstrained" parameters in the model.

- A parameter counts as 1 if its estimate is "independent" of the prior;
- A parameter counts as 0 if it is fully constrained by the prior.
- A parameter gives a partial value if both the data and prior are informative.

Watanabe-Akaike Information Criterion (WAIC)

- WAIC can be viewed as an approximation to leave-one-out CV, and averages over the entire posterior, vs. AIC and DIC which use point estimates.
- But it doesn't work well with highly structured data; no real alternative to more clever uses of Bayesian cross-validation.

Bayesian LOO-CV

By default, Bayesian LOO-CV is extremely expensive:

$$\text{loo-cv} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}(y_i),$$

which requires refitting the model without y_i for every data point.

Can think of the "number of effective parameters" as the difference between the log-predictive density of the data set minus the LOO-CV estimate.

Bayesian LOO-CV: More Advanced Methods

There are approximations to Leave-One-Out Cross-Validation which use *importance sampling* to avoid this, and these can be extended to time series.

See

- Vehtari et al (2015) on "Pareto-smoothed "
- Bürkner et al (2020) on time-series "leave-future-out CV".

"Bayesian" "Information" Criterion

$$\text{BIC} = -2 \log p(y|\hat{\theta}_{\text{MLE}}) + k \log n.$$

BIC:

- Is not Bayesian (it relies on the MLE);
- Has no relationship to information theory (unlike AIC/DIC);
- Assumes $\mathcal{M}$-closed (e.g. that the true model is under consideration);

"Bayesian" "Information" Criterion

BIC approximates the *prior* log-predictive likelihood and leave-k-out cross-validation (hence the extra penalization for additional parameters).

Differences between BIC values can therefore be interpreted as Bayes factors, which are ratios of marginalized likelihoods (see Kass & Raftery (1995) for more on Bayes factors).

This is why it's odd when model selection consists of examining both AIC and BIC: these are different quantities with different purposes!

Key Takeaways

Key Takeaways

- Model selection is a balance between bias (underfitting) and variance (overfitting).
- For predictive assessment, leave-one-out cross-validation is an ideal, but hard to implement in practice (particularly for time series).
- AIC and DIC can be used to approximate K-L divergence and leave-one-out cross-validation.
- BIC is an entirely different measure, approximating the prior predictive distribution.

Key Consideration

Model selection can result in significant biases when separated from hypothesis-driven model development.

- There's no free lunch: better off thinking about the scientific or engineering problem you want to solve and use domain knowledge/checks rather than throwing a large number of possible models into the machinery.
- Regularizing priors reduce potential for overfitting.
- Model averaging (Hoeting et al (1999)) and stacking (Yao et al (2018)) can combine multiple models as an alternative to selection.

Upcoming Schedule

Upcoming Schedule

Wednesday: Discussion of Höge et al (2019).

Next Monday: Emulation of Complex Models