More on information criteria and LOO-CV

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 14, 2020

Outline

Last time:

- Intro information theory
- Definitions of AIC, WAIC

Today:

- More on AIC/WAIC
- Leave-one-out cross-validation and approximate calculations

Information theory

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

$$D_{KL}(p, q) = \sum p_i (\log_2 p_i - \log_2 q_i)$$

Interpretation:

- p is the true outcome distribution
- q is the model predictive distribution
- KL divergence measures incorrectness, in some way

KL divergence for model comparison

We think of $D_{KL}(p, q)$ as measuring the distance from our model, q, to the truth, p.

Problem: we don't know p and never will!

But this isn't an obstacle for comparing models, because if we have two models q and r, then

$$D_{KL}(p, q) - D_{KL}(p, r) = \sum p_i(\log_2(r_i) - \log_2(q_i))$$

i.e. the $H(p)$ term drops out. We still don't know p_i, but we can estimate this from a sample of observations (because the observations are drawn from p_i)

Log score and deviance

Scoring models using log probabilities:

$$\log \text{score} \qquad S(q) = \sum \log(q_i)$$

Deviance:

$$D = -2 * S(q) = -2 * \sum \log(q_i)$$

In Bayesian world, the posterior isn't one model, it's a distribution of models – so we should average:

$$\text{lppd}(y, \theta) = \sum_i \log \left(\frac{1}{S} \sum_s p(y_i|\theta_s) \right)$$

(log pointwise predictive density)

Out-of-sample prediction error

The problem is to estimate prediction error (evaluated by log score) out of sample.

- Adding parameters generally improves fit within the sample
- Eventually, adding parameters reduces accuracy out of the sample (overfitting)
- How can we predict out-of-sample prediction accuracy?
  - Cross-validation
  - Information criteria

Goals

Since we use lppd to estimate the fit of our model, our goal with all of these tools is to estimate what our lppd will be on new data.

In other words, all of the following are estimates of some form of

$$\text{elpd} = -2\mathbb{E}(\log p(\tilde{y}|y))$$

the expected log predictive density of a new data point. In some cases (e.g. with AIC) we'll calculate the expected deviance of a new data set of the same size (can work either way).

Overfitting in action

To demonstrate overfitting, we'll consider a few models fit to fake data.

True data-generating process:

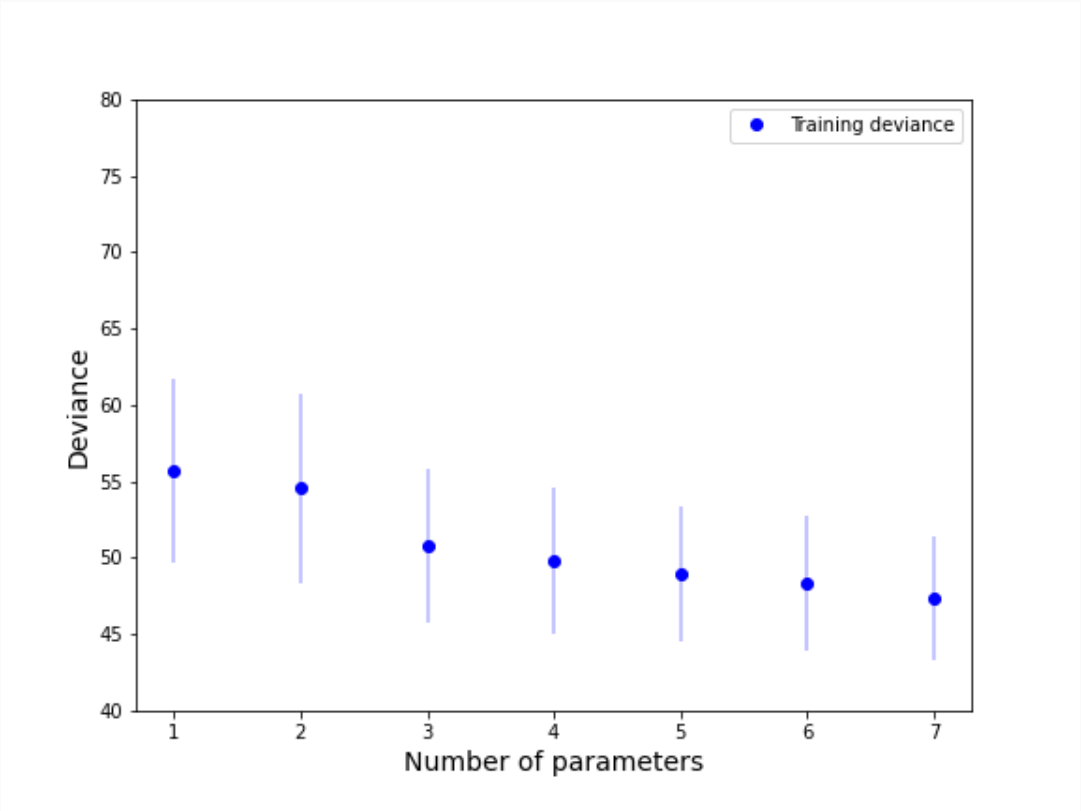
$$y_i \sim \text{Normal}(\mu_i, 1)$$
$$mu_i = 0.15x_1 - 0.40x_2$$

We'll fit models with the same likelihood and

$$\mu_i = \alpha$$
$$\mu_i = \alpha + \beta_1 x_1$$
$$\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2$$
$$\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$\dots$$

Overfitting in action

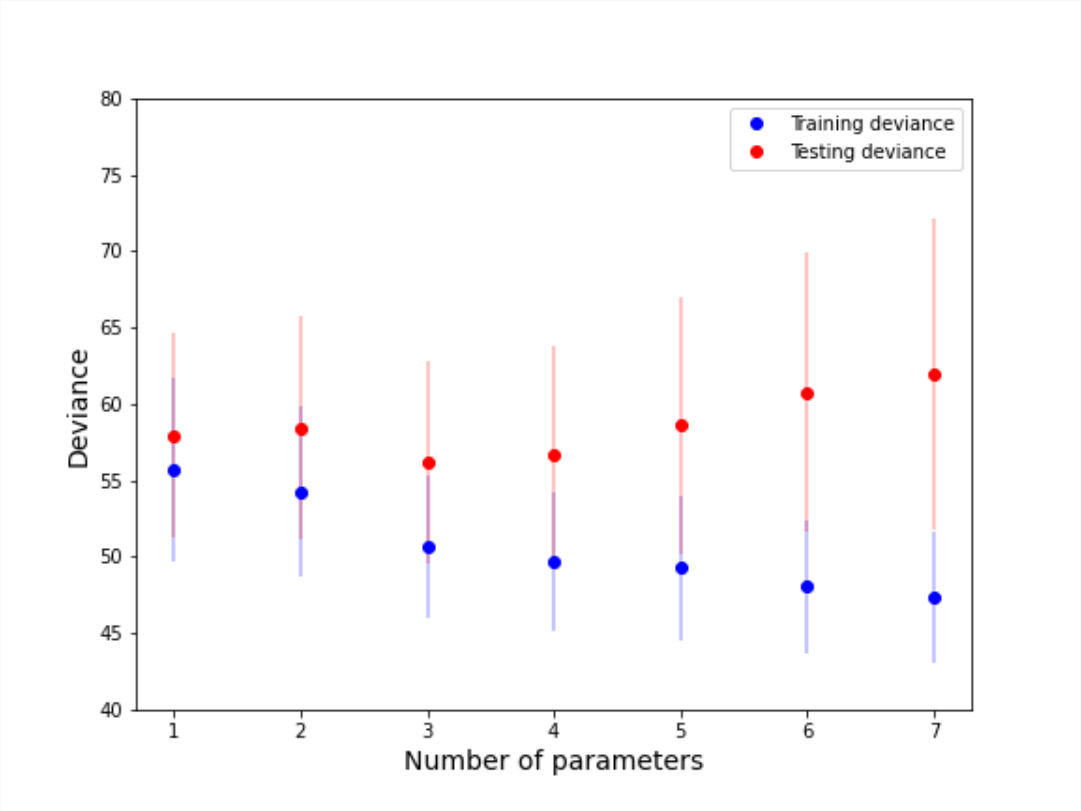
On the training set:



Remember: past 3 parameters, the predictors have no relationship to y in the true data generating process

Overfitting in action

Add in the testing set:



As expected, the additional parameters just make matters worse.

Akaike information criterion (AIC)

AIC: named for Akaike (but he called it "an information criterion"). Attempts to estimate the out-of-sample deviance.

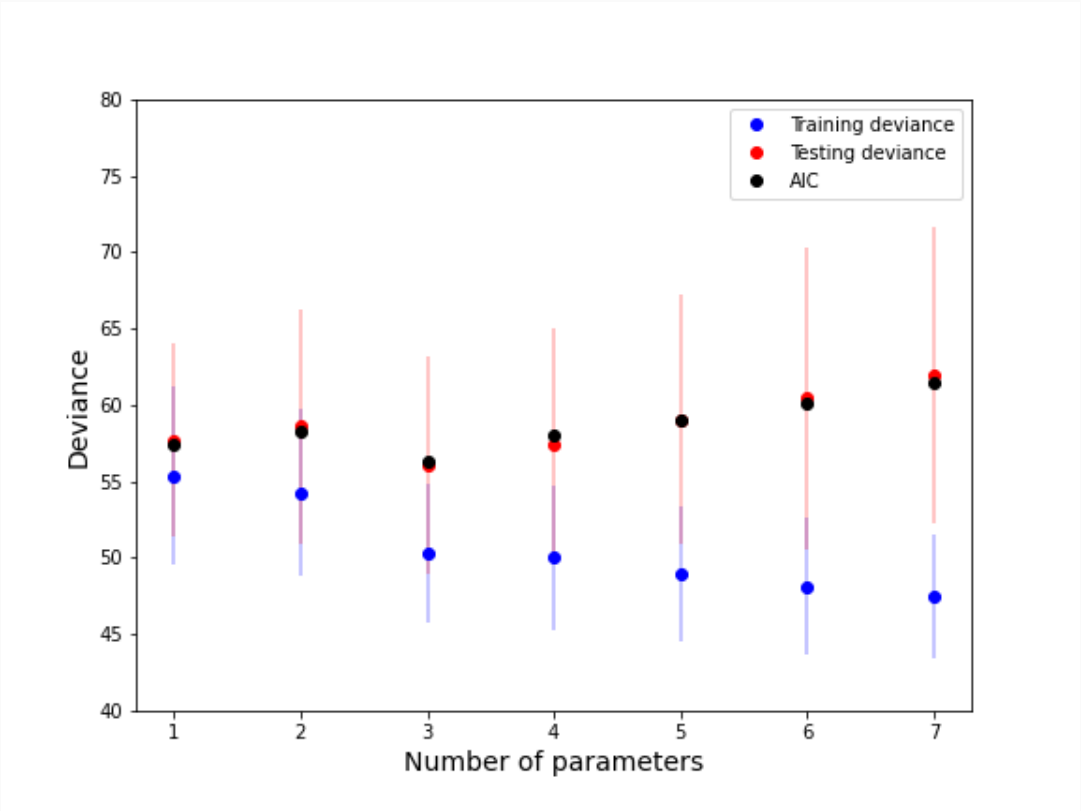
Assuming a point estimate $\hat{\theta}$ for model parameters, calculate the log score and apply a penalty to correct for overfitting:

$$AIC = D_{\text{train}} + 2k$$

k is the number of parameters. Assumes Gaussian posterior.
Where it comes from: Taylor expansion around the posterior mode.

Overfitting in action

Adding the AIC:



We see that for this model, the AIC is a good estimate of the out-of-sample deviance.

Widely applicable information criterion

Introduced by Watanabe (2010); a more Bayesian generalization of the AIC.

$$WAIC = -2(\text{lppd}(y, \theta) - \sum_i \text{var} \log(p(y_i|\theta)))$$

lppd replaces the training deviance; pointwise variance in log posterior density generalizes the parameter count (effective number of parameters).

Reduces to AIC in the special case where AIC is exact: models with flat priors and Gaussian posterior.

Leave-one-out cross-validation

LOO cross-validation

Idea behind cross-validation:

- Hold out some of your data for evaluation
- Fit the model on the remaining data
- Evaluate the model by estimating lppd on the held-out data
- Repeat, with different partitionings

Importance sampling

Instead of refitting and resampling the model for each held-out point, we can use the technique of *importance sampling.*

Idea: want to calculate $\mathbb{E}_p(h(\theta))$, but we can't generate samples from $p(\theta)$; however, we can generate random values from an approximation $g(\theta)$. Then,

$$E(h(\theta)) = \frac{\int h(\theta)p(\theta)d\theta}{\int p(\theta)d\theta}$$

Multiply and divide by $g(\theta)$ to get

$$E(h(\theta)) = \frac{\int [h(\theta)p(\theta)/g(\theta)]g(\theta)d\theta}{\int [p(\theta)/g(\theta)]d\theta}$$

Importance sampling

Then, we can estimate using draws θ_s from $g(\theta)$:

$$E(h(\theta)) \approx \frac{\sum_s h(\theta_s) w(\theta_s)}{\sum_s w(\theta_s)}$$

where

$$w(\theta_s) = \frac{p(\theta_s)}{g(\theta_s)}$$

are the *importance weights*.

IS-LOO

In LOO-CV we are trying to estimate $\log p(y_i|y_{-i})$, where y_{-i} denotes the set of observed y values without y_i.

Using importance sampling, we estimate the posterior predictive distribution given y_{-i} by

$$p(\tilde{y}|y_{-i}) = \frac{\sum_s w_{i,s} p(\tilde{y}|\theta_s)}{\sum w_{i_s}}$$

where the weights are $w_{i,s} = \frac{1}{p(y_i|\theta_s)}$

Evaluating this at $\tilde{y} = y_i$, we get

$$p(y_i|y_{-i}) \approx \frac{1}{\frac{1}{S}\sum_s \frac{1}{p(y_i|\theta_s)}}$$

So we can calculate this from a single posterior sample $\{\theta_s\}_{s \in S}$.

Pareto smoothing

In practice, importance sampling behaves poorly if the distribution of importance weights have high variance/long tails.

A fix: Pareto smoothing. Take the top 20% of importance weights, fit a generalized Pareto distribution to them; then replace those top weights with appropriate quantiles from the Pareto distribution.

Pareto smoothing

In practice, importance sampling behaves poorly if the distribution of importance weights have high variance/long tails.

A fix: Pareto smoothing. Take the top 20% of importance weights, fit a generalized Pareto distribution to them; then replace those top weights with appropriate quantiles from the Pareto distribution.

- `pm.compare` can be used to calculate WAIC or PSIS-LOO from a sample trace.

Summary

Today:

- Model comparison using WAIC / LOO.
- Note these are trying to estimate the same thing, so if they disagree substantially then something is funny

Further reading:

- *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC*
  `https://arxiv.org/pdf/1507.04544.pdf`