Computational Cognitive Science

Lecture 8: Model comparison and selection 2

Chris Lucas

School of Informatics

University of Edinburgh

October 10, 2019

Readings

- Chapter 10.6 of F&L
- Chapter 11 of F&L

Recommended:

- "Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method" (2010) by Wagenmakers et al. (Link)

Model comparison

- We discussed some methods for comparing models using likelihoods under MLEs, and predictive acccuracy

Identifiability

If likelihood doesn't depend strongly on parameters:

- Effectively a simpler model
- Parameters may not be identifiable

Pros:

- Less likely to have unwanted flexibility
- Sometime we want parameters *not* to matter
  - "Nuisance parameters"
  - Ideally we integrate them out and forget about them

Identifiability

Cons:

- If parameters have important psychological interpretations, they should be identifiable
- Some flexibility is important – a model should be as simple as possible, but no simpler

Identifiability

1. Sometimes non-identifiability is inherent in a model; function from parameters to data isn't invertible
   - E.g., using mean response times to infer parameters of Weibull distribution
2. Sometimes identification is impossible in practice, because data are too sparse or noisy

Identifiability

If identifiability is important, we can perform an identifiability "sanity check" before collecting data.

1. Math, e.g., Jacobian rank (See F&L 10.6.1)
2. Simulate and recover – how accurate are inferences when there's a known truth, and data matched to experiment
   - Optimistic, by still useful
   - Can serve purposes similar to a power analysis

Identifiability

The same principles apply to model comparison – it's good to check that a given experiment can distinguish between models

Bayesian model comparison

Last time we discussed approaches to model comparison that don't involve marginal likelihoods.

Today we'll talk about using and approximating marginal likelihoods.

Bayesian model comparison

Even if we have marginal likelihoods under several models, we can't compute $P(\mathcal{M}|\mathbf{y})$:

We would need to sum over all possible models and data distributions for all of them. Instead:

- Given any two models $\mathcal{M}_1$ and $\mathcal{M}_1$, we can compute the ratio of their posterior probabilities:

$$\frac{P(\mathcal{M}_1|\mathbf{y})}{P(\mathcal{M}_2|\mathbf{y})} = \frac{P(\mathbf{y}|\mathcal{M}_1)P(\mathcal{M}_1)}{P(\mathbf{y}|\mathcal{M}_2)P(\mathcal{M}_2)}$$

(The impossible sums cancel out)

Bayesian model comparison

There is unlikely to be consensus about $P(\mathcal{M})$; in practice people use *Bayes factors*:

- Posterior ratio given equal prior probability
- How strongly you'd have to prefer a model *a priori* in order to (still) favor it *a posteriori*

Lots of opinions about what constitutes a "convincing" Bayes factor

Estimating marginal likelihoods

If we have no closed-form solution for a marginal likelihood, what can we do?

We have several options, including:

1. Numerical integration
2. Importance sampling
3. ~~Harmonic mean estimation~~
4. Transdimensional MCMC
5. Savage-Dickey density ratio
6. BIC

Numerical integration

Use a general-purpose algorithm to integrate a function within a hypercube.

- Easy!
- Requires bounds on the high-density parts of the space
- Falls apart in high-dimensional spaces
- Risks missing narrow peaks

Numerical integration: Example

Recall that the normal density function is $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\mu-x)^2}{2\sigma^2}}$.

```
library(cubature)
sd=10;mu=.5
unnGauss <- function(x) {exp(-(mu-x)^2/(2*sd^2))}
adaptIntegrate(unnGauss,c(-1E3),c(1E3))
```

Result:

```
$integral
[1] 25.06628
```

```
> sqrt(2*pi*sd^2)
[1] 25.06628
```

(Also see F&L listing 11.1)

Importance sampling

Suppose we have:

- $p(\boldsymbol{\theta}|\mathcal{M})$
- $p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$

and we want:

- $p(\mathbf{y}|\mathcal{M}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$
- $E[\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}]$

Importance sampling

We can estimate what we want, using *importance sampling*:

(omitting $\mathcal{M}$)

1. Draw J samples from a **normalized** *proposal distribution* $g(\boldsymbol{\theta})$
2. Weight each sample: $w_j = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}$
3. The expectation of $\boldsymbol{\theta}$ is approximately $\frac{\sum_j w_j \boldsymbol{\theta}^{(j)}}{\sum_j w_j}$
4. The marginal likelihood is approximately $\frac{1}{J}\sum_j w_j$

If the variance of the weights is low, these are *probably* trustworthy estimates.

Simple Monte Carlo integration

A special case of importance sampling:

1. Sample from the prior (usually easy)
2. Weight by likelihood: $w_j = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} = p(\mathbf{y}|\boldsymbol{\theta})$

Easy and acceptable if you think the samples will cover high-density areas of the posterior.

Better: Find a (normalized) proposal function that resembles your posterior

A demonstration

Our problem boils down to estimating the integral of a function.

We can use standard probability densities to see this works, since we know their integral (over **data**, not parameters) is 1.

Gaussian:

$$\int_x e^{-\frac{(\mu - x)^2}{2\sigma^2}}\, dx = \sqrt{2\pi\sigma^2}$$

Let's integrate over x (analogous to our parameters in model selection).

Importance sampling

```r
impSamp <- function(targD,ef) {
  nSamps = 40000 # The more the better
  # Using a student's t distribution, df=1
  proposals <- rt(nSamps,1)
  pDens <- dt(proposals,1)
  unnP <- targD(proposals)
  w <- unnP/pDens
  print(paste("Expected value of target function:",
    sprintf("%2.3f",sum(w*ef(proposals))/sum(w))))
  print(paste("Average importance weight:",
    sprintf("%2.3f",sum(w)/nSamps)))
}
```

Gaussian example

```
sd=.05;mu=5;
unnGauss <- function(x) {exp(-(mu-x)^2/(2*sd^2))}
# Real
print(sprintf("Real: %2.3f",sqrt(2*pi*sd^2)))
# Numerical
adaptIntegrate(unnGauss,c(-1E3),c(1E3))
# Importance
impSamp(unnGauss,function(x) x)
```

Gaussian example

```
sd=.05;mu=5

Real normalization constant Z: 0.125

Cubature estimate of Z: 0 (oops)

Importance sampling:
  Estimated mean: 5.000
  Avg. importance weight (Z): 0.132 (close)
```

Importance sampling

If we were interested in the marginal likelihood, we would propose μ and σ rather than x (and would need priors for both)

Importance sampling

- General-purpose Monte Carlo method for approximating parameter distributions
- Can exploit knowledge about high-density regions of posterior
- Can compute expectations of functions of params
- Requires good proposals
  - Increasingly so as dimensionality goes up
- In these cases, additional tricks may be necessary, e.g.,
  - **annealed** importance sampling (link)
  - Inference trees (link)

Harmonic mean estimation

See "The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever" by Radford Neal. (link)

Excerpts:

- "abysmal performance in most real problem[s]"
- "the total unsuitability of the harmonic mean estimator should have been apparent within an hour of its discovery"

Don't use it.

Transdimensional MCMC

- Use Markov Chain Monte Carlo, combining multiple models into a single overarching one
- Nice in principle
- Often difficult / fiddly in practice
- Out of scope for this course

Savage-Dickey density ratio

- Efficiently compare nested probabilistic models
- Effectively a better and Bayesian alternative to the likelihood-ratio test
    - See recommended reading to learn more

BIC

$$BIC = 2 \cdot NLL + K \log(N)$$

where N is the number of data points and K is the number of parameters.

- Motivated by model comparison per se, not prediction
- Can be understood as a “minimum description length” approach
- Like AIC, a model-comparison method that boils down to MLE likelihoods and counting parameters
- Like AIC, rests on assumptions; guarantees are asymptotic
- Easy!
- Safer than AIC if arguing for a more complex model

Summary

- If you care about parameter identifiability, check!
- Many methods for approximating the marginal likelihood
- Easy cases:
  - low-dimensional models (numerical integration, importance sampling)
  - nested models (Savage-Dickey)
  - conjugate priors (didn't discuss)
- Hard case: High-dimensional, non-conjugate, non-nested
  - Transdimensional MCMC (out of scope)
  - Clever/lucky proposals
  - Annealed importance sampling (out of scope)