Bayesian Model Selection: Demonstration with POPAN and Hunt's Data

Robert W Rankin and Krista E. Nicholson

Georgetown University, Murdoch University)

October 28, 2017

Model Selection

The model selection world may be divided into two camps...

Hypothesis-falsification/confirmation

- goal: find the "true model"
- properties: consistency
- iconic techniques: Bayes Factors [Kass and Raftery, 1995], the BIC [Schwarz, 1978], consistent selectors (e.g., Adaptive-Lasso)[Zou, 2006], APE[Wagenmakers et al., 2006].

Prediction / estimation

- goal: minimize a predictive loss
- properties: efficiency
- iconic techniques: cross-validation, AIC, WAIC [Watanabe, 2010], Lasso [Efron et al., 2004], boosting, model-averaging, random-effects

Two Paradigms

For more about the philosophical distinctions and practical differences between the predictive approaches and hypothesis-testing/confirmation approaches, read:

- predictive approaches: **?**
- Bayes Factor approaches: [Kass and Raftery, 1995]
- all-together: Shiffrin et al. [2008]
- philosophical distrinctions in CMR: Rankin 2017, PhD Thesis, Murdoch University

Other

There are many other Bayesians techniques that have a flavour of model-selection or multi-model inference, with properties that are less-well-understood (or even arbitrary).

- stochastic variable search [George and McCulloch, 1993]
- spike-and-slab priors [Miller 2006]
- hierarchical-models [Rankin et al 2016]

The above are easily implimented in JAGS/BUGS
For more information about these, see the many reviews like: Shiffrin et al. [2008], Hooten and Hobbs [2015], O'Hara and Sillanpää [2009]

Model Selection

For our purposes, we will focus on two Bayesian approximations to:

- posterior predictive loss and LOO-cross-validation using the Watanabe-Akaike Information Criteria (WAIC)
- Bayes Factors using the Harmonic Mean estimate of marginal likelihoods

WAic

The WAIC

- asymptotically selects the same models as AIC and LOOCV (under certain conditions)

-goal: minimize a posterior predictive loss

$$WAIC = -2 \sum_{i=1}^{n} \log \left(\frac{1}{J} \sum_{j=1}^{J} \mathcal{L}(\mathbf{y}_i | \boldsymbol{\theta}^{(j)}) \right) + 2 p_{\text{WAIC}}$$

where ..

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \mathbb{V}[\log \mathcal{L}(\mathbf{y}_i | \boldsymbol{\theta}^{(j)})]$$

- recommended for estimation problems (e.g., want a time-series of N, want to quantify the difference in survival between males and females, etc.). I.e., in situations where you want to get close to the truth
- has a similar intuition as the AIC: first term a goodness-of-fit quantity, 2nd a complexity penalty

WAIC

WAIC notes:

- the WAIC approximation assumes independence among *observations* or “data-points" (i.e., the i in the summation terms) will
- sequential captures are correlated
- we use consider each individual capture-history $y_i = \{000010101\}$ to be a unit of data. Captures-histories are assumed independent.

WAIC in JAGS

In JAGS, the WAIC is a two-step process.

1. Step one, calculate the logdensity per “data-point"

```
for(i in 1:N.obs){ # loop through individuals
  for(t in 1:T){ # loop through time
     llit[i,t] <- logdensity.cat(y[i,t], em[1:E, z[i,t],t])
  } #
  lli[i] <- sum(llit[i,1:T])
} # individuals N.obs
}
```

- `N.obs`: number of observed individuals (not pseudo-individuals),
- `E` : number of capture events (for POPAN, E=2)
- `lli[i]`: is $\log \mathcal{L}(\mathbf{y}_i|\boldsymbol{\theta}^{(j)})$ in term two of the WAIC. I.e., the log-density of the data.

WAIC in JAGS

2. Step Two: Use the samples of `lli[i]` for all `i:N.obs` in the following function:

```
waic1.jags <- function(post,n.obs,loglike.grep.txt = "lli"){
    logsumexp <- function(x){
       max.x <- max(x);
       max.x - log(length(x)) +log(sum(exp(x-max.x)))}
    ll.col <-grep(loglike.grep.txt,colnames(post[[1]]))
    if(length(post)>1){
      lli.samp <- as.matrix(do.call("rbind",post)[,ll.col])
     } else {
       lli.samp <- as.matrix(post[[1]][,ll.col])}
    logElike <- apply(lli.samp,2,logsumexp)
    Eloglike <- apply(lli.samp,2,mean)
    p_waic1 <- 2*sum(logElike-Eloglike)
    logsumf.i <- apply(lli.samp,2,logsumexp)
    lppd <- sum(logsumf.i)
    waic1 <- -2*(lppd-p_waic1)
    return(waic1)}
```

Yikes! That was crazy!

The previous just uses a common "log-sum-exp" trick to compute the WAIC. Don't worry about it.

WAIC Approximation

- "We approximated the WAIC using the Complete Data log-Likelihood (logCDL) of each capture history given their latent state sequences, per MCMC iteration. In expectation (over the posterior density), the Expected CDL is should be the same as the expectation of the exact log-likelihood"

The WAIC and model-averaging

- The best WAIC model has some important optimality properties for estimation.
- it is better (for estimation) to do model averaging with "pseudo-posterior model probabilities" based on the WAIC

$$p^*(M_j|\mathbf{Y}) = \frac{e^{(-0.5\Delta\text{WAIC}_j)}}{\sum_{M_k \in \mathcal{M}} e^{(-0.5\Delta\text{WAIC}_k)}} \quad (1)$$

. . . a.k.a. WAIC-weights.
You can then do model averaging. . .

$$p*(N_t|\mathbf{Y}) = p^*(M_1|\mathbf{Y})p(N_t|\mathbf{Y}, M_1) + \cdots + p^*(M_k|\mathbf{Y})p(N_t|\mathbf{Y}, M_k)$$

The WAIC and model-averaging

Model-averaging of posterior distributions (like N_t or recruits) with JAGS output:
Assuming you have the SAME number of MCMC samples per model...

1. loop through each model M_j with WAIC-weight $p^*(M_j|\mathbf{Y})$
2. sub-sample (without replacement) each j MCMC sample of $\theta^{(j)}$ with probability $p^*(M_j|\mathbf{Y})$ (i.e., the WAIC-weight)
3. `rbind` all the $\theta^{(j|k)}$ sub-samples together.

For this to be a good approximation, it is important to have A LOT of MCMC samples per model (>5000).

EXERCISE:

EXERCISE : run 4 POPAN models and calculate the WAIC, using Hunt et 2017 Sousa data

1. $\psi(t)\phi(t)p(t)$ fully-time-varying
2. $\psi(t)\phi(\cdot)p(t)$ time-constant `phi`
3. $\psi(t)\phi(t)p(\cdot)$ time-constant `p`
4. $\psi(t)\phi(\cdot)p(\cdot)$ time-constant `p` and `phi`

In each case, calculate the WAIC for comparison

References

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053604000000067.

Mevin Hooten and N Thompson Hobbs. A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28, February 2015. ISSN 0012-9615. doi: 10.1890/14-0661.1.

Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430): 773–795, June 1995. ISSN 0162-1459. doi: 10.1080/01621459.1995.10476572.

R. B. O'Hara and M. J. Sillanpää. A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–117, March 2009. ISSN 1936-0975. doi: 10.1214/09-BA403.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136.

Richard Shiffrin, Michael Lee, Woojae Kim, and Eric-Jan Wagenmakers. A survey of model evaluation approaches with a tutorial on Hierarchical Bayesian methods. *Cognitive Science: A Multidisciplinary Journal*, 32(8): 1248–1284, December 2008. ISSN 0364-0213. doi: 10.1080/03640210802414826.

Eric-Jan Wagenmakers, Peter Grünwald, and Mark Steyvers. Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50(2):149–166, April 2006. ISSN 0022-2496. doi: 10.1016/j.jmp.2006.01.004.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v11/watanabe10a.html.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476): 1418–1429, 2006. ISSN 0162-1459. doi: 10.1198/016214506000000735.