Predictive information criteria in hierarchical Bayesian models for clustered data

Sophia Rabe-Hesketh & Daniel Furr

Education & Biostatistics
University of California, Berkeley
sophiarh@berkeley.edu

Joint work with Ed Merkle

Psychological Sciences, University of Missouri

StanCon 2018, Asilomar, Pacific Grove

Outline of Talk

- **Predictive information criteria**
  DIC and WAIC with connections to leave-one-out (LOO) cross-validation
- **Hierarchical Bayesian models for clustered data**
  Mixed/multilevel models (MLM), structural equation models (SEM), item response theory (IRT) models
- **Marginal versus conditional versions of DIC, WAIC, and LOO**
- **Application to IRT**

Targets of predictive information criteria (non-hierarchical Bayesian model)

- Model likelihood: $f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta})$ Model prior: $p(\boldsymbol{\theta})$
- Assess model by how well it predicts future, out-of-sample data $\boldsymbol{y}^{\mathsf{r}}$
- Measure of prediction error (scoring function) is deviance:

$$-2\log f(\boldsymbol{y}^{\mathsf{r}}|\boldsymbol{\theta}) = -2\sum_{i=1}^{N} \log f(y_i^{\mathsf{r}}|\boldsymbol{\theta})$$

- What to do about unknowable $\boldsymbol{\theta}$?
  - **DIC**: plug in posterior mean $\widetilde{\boldsymbol{\theta}}$

$$\text{plug-in deviance} = -2\log f(\boldsymbol{y}^{\mathsf{r}}|\widetilde{\boldsymbol{\theta}})$$

  - **WAIC**: integrate over $p(\boldsymbol{\theta}|\boldsymbol{y})$, but use **pointwise** predictive densities

$$-2\,\text{log pointwise predictive density} = -2\sum_{i=1}^{N} \log \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} f(y_i^{\mathsf{r}}|\boldsymbol{\theta})$$

- Targets are **expectations** of the above over out-of-sample data $\boldsymbol{y}^{\mathsf{r}}$

[Gelman, Hwang & Vehtari, 2014]

DIC (non-hierarchical Bayesian model)

- Expectation of plug-in deviance (pid) over distribution of $\boldsymbol{y}^{\mathsf{r}}$:

$$\text{expected pid} = -2\mathsf{E}_{\boldsymbol{y}^{\mathsf{r}}} \log f(\boldsymbol{y}^{\mathsf{r}}|\widetilde{\boldsymbol{\theta}})$$

- Data-generating distribution of $\boldsymbol{y}^{\mathsf{r}}$ unknown & validation data $\boldsymbol{y}^{\mathsf{r}}$ not available
- Use within-sample pid and penalize for using data twice

$$\text{DIC} = -2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + 2p_{\text{D}}$$

where p_{D} in penalty term is effective number of parameters

$$p_{\text{D}} = \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[-2\log f(\boldsymbol{y}|\boldsymbol{\theta})] - [-2\log f(\boldsymbol{y}|\tilde{\boldsymbol{\theta}})]$$

- Posterior means estimated as averages over MCMC draws

[Spiegelhalter, Best, Carlin & van der Linde, 2002]

WAIC (non-hierarchical Bayesian model)

- -2 expected log pointwise predictive density (elppd) over distribution of $\boldsymbol{y}^{\mathrm{r}}$:

$$-2\,\text{elppd} = -2\sum_{i=1}^{N} \mathsf{E}_{\boldsymbol{y}^{\mathrm{r}}} \log \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} f(y_i^{\mathrm{r}}|\boldsymbol{\theta})$$

- Data-generating distribution of $\boldsymbol{y}^{\mathrm{r}}$ unknown & validation data $\boldsymbol{y}^{\mathrm{r}}$ not available
- Use within-sample lppd and penalize for using data twice

$$\text{WAIC} \;=\; -2\sum_{i=1}^{N} \log \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} f(y_i|\boldsymbol{\theta}) + 2p_{\mathsf{W}}$$

where p_{W} in penalty term is effective number of parameters

$$p_{\mathsf{W}} \;=\; \sum_{i=1}^{N} \mathsf{Var}_{\boldsymbol{\theta}|\boldsymbol{y}} \log f(y_i|\boldsymbol{\theta})$$

- Posterior means and variances estimated from MCMC draws
- Asymptotically equivalent to LOO cross-validation (LOO-CV)

LOO-CV and PSIS-LOO (non-hierarchical Bayesian model)

- Same target as WAIC

$$-2\,\text{elppd} = -2\sum_{i=1}^{N} \mathsf{E}_{\boldsymbol{y}^{\mathrm{r}}} \log \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} f(y_i^{\mathrm{r}}|\boldsymbol{\theta})$$

- Estimate using LOO-CV

$$-2\,\text{LOO-CV} \;=\; -2\sum_{i=1}^{N} \log \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}_{-i}} f(y_i|\boldsymbol{\theta})$$

  - Where $\boldsymbol{y}_{-i}$ is the "training" data without unit i
  - Requires running MCMC on each of N training datasets
- Approximate by Pareto-smoothed importance sampling (PSIS)
  - Idea of importance sampling (IS)

$$-2\,\text{IS-LOO} \;=\; -2\sum_{i=1}^{N} \log \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} \underbrace{\left[\frac{p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})}{p(\boldsymbol{\theta}|\boldsymbol{y})}\right]}_{\text{importance ratio}} f(y_i|\boldsymbol{\theta})$$

  - Importance ratios $\propto \frac{1}{f(y_i|\boldsymbol{\theta})}$; Unstable, hence Pareto smoothing

[Vehtari, Gelman & Gabry, 2017]

Hierarchical Bayesian models for clustered data

Stage		MLM Example	Densities, general notation	
3	Responses	$y_{ij} \sim N(\alpha+\zeta_j, \sigma^2)$	$f_c(y_{ij}\vert\boldsymbol{\omega},\boldsymbol{\zeta}_j)$	$\boldsymbol{\omega} \equiv (\alpha,\sigma^2)'$
		unit $i=1,\ldots,n_j$		$\boldsymbol{\zeta}_j \equiv \zeta_j$
2	Direct param.	$\zeta_j \sim N(0,\psi)$	$g(\boldsymbol{\zeta}_j\vert\boldsymbol{\psi})$	$\boldsymbol{\psi} \equiv \psi$
		cluster $j=1,\ldots,J$		
		Fully Bayesian $\downarrow$		
		prior for α,σ^2	$p(\boldsymbol{\omega})$	
1	Hyperparameters	hyperprior for ψ	$p(\boldsymbol{\psi})$	

- ζ_j are direct parameters, varying intercepts/coefficients (MLM), latent variables (SEM/IRT), missing data
- In Bayesian setting, ambiguous whether ζ_j are parameters or (latent) variables

Revisit DIC
Two versions of the likelihood (or deviance)

- **Conditional likelihood**: $\prod_j f_c(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\zeta}_j)$, where

$$f_c(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\zeta}_j) = \prod_{i=1}^{n_j} f_c(y_{ij}|\boldsymbol{\omega},\boldsymbol{\zeta}_j)$$

  - Natural definition in `Stan` (or `BUGS`/`JAGS`) code
  - Condition on $\boldsymbol{\omega}$ and $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1{}',\ldots,\boldsymbol{\zeta}_J{}')'$
- **Marginal likelihood**: $\prod_j f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\psi})$, where

$$f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\psi}) = \int f_c(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\zeta}_j) g(\boldsymbol{\zeta}_j|\boldsymbol{\psi}) d\boldsymbol{\zeta}_j$$

  - Natural in maximum likelihood (ML) estimation (e.g., `lmer` in R)
  - Condition on $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$, the only parameters in ML setting
  - In MLM example, $f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\psi})$ is MVN with means α, variances $\psi+\sigma^2$, and covariances ψ

Conditional and marginal DIC

- **Conditional DIC**
  ζ (and ω) "in focus" [Spiegelhalter, Best, Carlin & van der Linde, 2002]

$$\mathrm{DIC_c} = -2\log f_c(\boldsymbol{y}|\tilde{\boldsymbol{\omega}},\tilde{\boldsymbol{\zeta}}) + 2p_{\mathrm{Dc}}$$

$$p_{\mathrm{Dc}} = \mathrm{E}_{\boldsymbol{\omega},\boldsymbol{\zeta}|\boldsymbol{y}}[-2\log f_c(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\zeta})] + 2\log f_c(\boldsymbol{y}|\tilde{\boldsymbol{\omega}},\tilde{\boldsymbol{\zeta}})$$

  - Used in almost all application, easy with `Stan`, `BUGS`, `JAGS`
- **Marginal DIC**
  $\boldsymbol{\psi}$ (and $\boldsymbol{\omega}$) "in focus"

$$\mathrm{DIC_m} = -2\log f_m(\boldsymbol{y}|\tilde{\boldsymbol{\omega}},\tilde{\boldsymbol{\psi}}) + 2p_{\mathrm{Dm}}$$

$$p_{\mathrm{Dm}} = \mathrm{E}_{\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y}}[-2\log f_m(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\psi})] + 2\log f_m(\boldsymbol{y}|\tilde{\boldsymbol{\omega}},\tilde{\boldsymbol{\psi}})$$

  - Provided by R package `blavaan` [Merkle & Rosseel, 2018] for SEM (which evaluates $f_m(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\psi})$ using `lavaan`) and by `Mplus`
  - Efficient adaptive quadrature to evaluate intractable integrals [Furr, 2017; Rabe-Hesketh, Skrondal & Pickles, 2005]

Revisit WAIC
Two versions of predictive distributions

- **Posterior predictive distribution** for new unit in *existing* cluster

$$\mathrm{E}_{\boldsymbol{\omega},\zeta_j|\boldsymbol{y}} f_c(y^r_{ij}|\boldsymbol{\omega},\zeta_j) = \int f_c(y^r_{ij}|\boldsymbol{\omega},\zeta_j)\underbrace{\left[\int p(\boldsymbol{\zeta}_j|\boldsymbol{y}_j,\boldsymbol{\omega},\boldsymbol{\psi})p(\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}\right]}_{p(\boldsymbol{\omega},\zeta_j|\boldsymbol{y})} d\boldsymbol{\omega}d\boldsymbol{\zeta}_j$$

  Uses **posterior** for $\zeta_j \Rightarrow$ directly influenced by $\boldsymbol{y}_j$
  $\Rightarrow$ treats ζ_j and therefore cluster as within-sample
- **Mixed predictive distribution** for new units in *new* cluster:

$$\mathrm{E}_{\boldsymbol{\omega},\psi|\boldsymbol{y}} f_m(\boldsymbol{y}^r_j|\boldsymbol{\omega},\boldsymbol{\psi}) = \int \underbrace{\left[\int f_c(\boldsymbol{y}^r_j|\boldsymbol{\omega},\zeta_j)g(\boldsymbol{\zeta}_j|\boldsymbol{\psi})d\boldsymbol{\zeta}_j\right]}_{f_m(\boldsymbol{y}^r_j|\boldsymbol{\omega},\boldsymbol{\psi})} p(\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\omega}d\boldsymbol{\psi}$$

  Uses **prior** for ζ_j
  $\Rightarrow$ treats ζ_j and therefore cluster as out-of-sample

[Gelman, Meng & Stern, 1996]

Conditional WAIC and LOuO-CV

$$\mathrm{WAIC_c} = -2\sum_{j=1}^{J}\sum_{i=1}^{n_j}\log\left[\mathrm{E}_{\boldsymbol{\omega},\zeta_j|\boldsymbol{y}} f_c(y_{ij}|\boldsymbol{\omega},\boldsymbol{\zeta}_j)\right] + 2p_{\mathrm{Wc}}$$

$$p_{\mathrm{Wc}} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathrm{Var}_{\boldsymbol{\omega},\zeta_j|\boldsymbol{y}}\left[\log f_c(y_{ij}|\boldsymbol{\omega},\boldsymbol{\zeta}_j)\right]$$

- Same target as leave-one-unit out (LOuO) CV

$$-2\,\mathrm{LOuO\text{-}CV} = -2\sum_{j=1}^{J}\sum_{i=1}^{n_j}\log \mathrm{E}_{\boldsymbol{\omega},\zeta_j|\boldsymbol{y}_{-ij}} f_c(y_{ij}|\boldsymbol{\omega},\zeta_j)$$

- $\mathrm{WAIC_c}$ and PSIS-LOuO provided by combination of `Stan` and R package `loo` [Vehtari, Gelman & Gabry, 2016]

Marginal WAIC and LOcO-CV

$$\mathrm{WAIC_m} = -2\sum_{j=1}^{J}\log\left[\mathrm{E}_{\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y}} f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\psi})\right] + 2p_{\mathrm{Wm}}$$

$$p_{\mathrm{Wm}} = \sum_{j=1}^{J}\mathrm{Var}_{\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y}}\left[\log f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\boldsymbol{\psi})\right]$$

- Same target as leave-one-cluster out (LOcO) CV

$$-2\,\mathrm{LOcO\text{-}CV} = -2\sum_{j=1}^{J}\log \mathrm{E}_{\boldsymbol{\omega},\psi|\boldsymbol{y}_{-j}} f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\psi)$$

- Can compute PSIS-LOcO using `loo` package with posterior samples of $f_m(\boldsymbol{y}_j|\boldsymbol{\omega},\psi)$ as input; automated in `blavaan` for SEM!
- Ever used??
  - Hinted at [e.g., Gelman, Hwang & Vehtari, 2014 – Section 2.5]
  - Used for unclustered data with latent variables (e.g., overdispersed Poisson, meta-analysis) [Li, Qui & Feng, 2016; Millar, 2018]

WAIC and LOO-CV for unclustered data

- In unclustered data with **univariate** y_j (instead of $\boldsymbol{y}_j$), posterior predictive density collapses to mixed predictive density

$$\mathsf{E}_{\boldsymbol{\omega},\zeta_j|\boldsymbol{y}} f_{\mathsf{c}}(y_j^{\mathsf{r}}|\boldsymbol{\omega},\zeta_j) = \int f_{\mathsf{c}}(y_j^{\mathsf{r}}|\boldsymbol{\omega},\zeta_j) \underbrace{\left[\int p(\zeta_j|\cancel{y_j},\cancel{\boldsymbol{\omega}},\boldsymbol{\psi}) p(\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi}\right]}_{p(\boldsymbol{\omega},\zeta_j|\boldsymbol{y})} d\boldsymbol{\omega} d\zeta_j$$

$$= \int \underbrace{\left[\int f_{\mathsf{c}}(y_j^{\mathsf{r}}|\boldsymbol{\omega},\zeta_j) g(\zeta_j|\boldsymbol{\psi}) d\zeta_j\right]}_{f\mathsf{m}(y_j^{\mathsf{r}}|\boldsymbol{\omega},\boldsymbol{\omega})} p(\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\omega} d\boldsymbol{\psi} = \mathsf{E}_{\boldsymbol{\omega},\boldsymbol{\psi}|\boldsymbol{y}} f_{\mathsf{m}}(\boldsymbol{y}_j^{\mathsf{r}}|\boldsymbol{\omega},\boldsymbol{\psi})$$

  No data for unit $j \Rightarrow$ **posterior** for ζ_j equals **prior** for ζ_j
- Therefore conditional PSIS-LOO makes no sense and not clear what WAIC$_{\mathsf{c}}$ represents!

[Millar, 2018]

8 schools example
WAIC and LOO-CV for unclustered data

- Meta-analysis of SAT prep. programs in 8 schools ($j = 1, \ldots, 8$)
- Effect size estimates y_j with standard error estimates σ_j
- Hierarchical model

$$y_j|\zeta_j,\sigma_j^2 \sim N(\zeta_j,\sigma_j^2), \quad \zeta_j|\mu,\tau^2 \sim N(\mu,\tau^2), \quad p(\alpha,\tau) \propto 1$$

$$f_{\mathsf{c}}(y_j|\zeta_j) = N(y_j|\zeta_j,\sigma_j^2) \qquad f_{\mathsf{m}}(y_j|\tau^2) = N(y_j|\mu,\tau^2+\sigma_j^2)$$

- Scale data $y_j^* = S \times y_j$, σ_j unchanged [Vehtari, Gelman & Gabry, 2017]

Scale factor S	WAIC$_{\mathsf{c}}$	LOO-CV	WAIC$_{\mathsf{m}}$
1	61.8	62.6	62.6
4	68.7	86.0	85.5

- WAIC$_{\mathsf{c}}$ terrible approximation to LOO-CV when $S = 4$
  [Vehtari, Gelman & Gabry, 2017 – Figure 1a (did not consider WAIC$_{\mathsf{m}}$)]
- WAIC$_{\mathsf{m}}$ much better approximation to LOO-CV
  Also found in other applications [Li, Qui & Feng, 2016; Millar, 2018]

Dan Furr: Application to IRT

Discussion

- Make informed choice between conditional and marginal ICs
- Marginal ICs generally more justified than conditional ICs
  - Want to assess specification of prior $g(\zeta_j|\boldsymbol{\psi})$
  - And/or want to generalize to other clusters
- Theoretical problems with conditional ICs
  - WAIC$_{\mathsf{c}}$ and PSIS-LOuO make no sense for unclustered data
  - WAIC$_{\mathsf{c}}$ does not meet regularity conditions: (a) $y_{ij}|\boldsymbol{\omega},\zeta_j$ not iid (b) number of parameters increases with sample size [Millar, 2018]
  - Penalty term for DIC$_{\mathsf{c}}$ problematic because number of parameters increases with sample size [Plummer, 2008]
- Empirical problem with conditional ICs
  - Both WAIC$_{\mathsf{c}}$ and DIC$_{\mathsf{c}}$ can have huge Monte Carlo errors
  - WAIC$_{\mathsf{c}}$ can be poor approximation to PSIS-LOuO

References to other authors

- Li, Qui & Feng (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing* 26, 881-897.
- Gelman, Meng & Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6, 733-807.
- Gelman, Hwang & Vehtari (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 24, 997-1016.
- Millar (2018). Conditional vs. marginal estimation of predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*. In press.
- Plummer (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* 9, 523-539.
- Spiegelhalter, Best, Carlin & van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* 64, 583-639.
- Vehtari, Gelman & Gabry (2016). `loo`: Efficient leave-one-out crossvalidation and WAIC for Bayesian models. https://github.com/stan-dev/loo
- Vehtari, Gelman & Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27, 1413-1432.
- Watanabe (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11, 3571-3594.

References to our work

- Furr (2017). Bayesian and frequentist cross-validation methods for explanatory item response models. PhD Thesis. UC Berkeley
- Merkle & Rosseel (2018). `blavaan`: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*. In Press.
- Merkle, Furr & Rabe-Hesketh (2018). Bayesian model assessment: Use of conditional vs marginal likelihoods. *arXiv:1802.04452*. http://arxiv.org/abs/1802.04452
- Rabe-Hesketh, Skrondal & Pickles (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* 128, 301-323.

- Web page on Education Research using Stan: `https://education-stan.github.io` (contributions welcome)
  - Tutorial and case-studies on IRT
  - Papers that use Stan in education research, broadly construed