Evaluating predictive loss for models with observation-level latent variables

Russell Millar
University of Auckland

Dec 2015

Notation

- $\boldsymbol{y} = (y_1, ..., y_n)$, observations with density $p(\boldsymbol{y})$
- $\boldsymbol{\theta} \in \mathbb{R}^d$, parameter vector
- $p(\boldsymbol{y}|\boldsymbol{\theta})$, the model
- $p(\boldsymbol{\theta})$, prior
- $\boldsymbol{z}$, future realizations from true distribution of $\boldsymbol{y}$.
- $D(\boldsymbol{\theta}) = -2 \log p(\boldsymbol{y}|\boldsymbol{\theta})$, deviance function

DIC, the Dirty Information Criterion

Widely used: Spiegelhalter et al. (2002) $> 6\,500$ cites.

DIC can be written as

$$\mathrm{DIC} = \overline{D(\boldsymbol{\theta})} + p \ ,$$

where p is a penalty term to correct for using the data twice.

A Taylor series expansion of $D(\boldsymbol{\theta})$ around $\overline{\boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}]$ "suggests" that p can be estimated as the posterior expected value of $D(\boldsymbol{\theta}) - D(\overline{\boldsymbol{\theta}})$, giving

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}}) \ .$$

- Not invariant to re-parameterization due to use of $\overline{\boldsymbol{\theta}}$. ☹☹☹
- p_D can be negative if deviance is not concave. ☹☹☹
- Never explicitly stated what DIC is trying to estimate!!!

WAIC, Widely Applicable Information Criteria

Sumio Watanabe (2009) developed a singular learning theory derived using algebraic geometry results developed by Heisuke Hironaka (who earned a Fields medal in 1970 for his work).

It is assumed that $p(y_i|\boldsymbol{\theta})$ are independent.

WAIC, Widely Applicable Information Criteria

Sumio Watanabe (2009) developed a singular learning theory derived using algebraic geometry results developed by Heisuke Hironaka (who earned a Fields medal in 1970 for his work).

It is assumed that $p(y_i|\boldsymbol{\theta})$ are independent.

Watanabe defines several WAIC variants. One particular variant has gained popularity due to:

- It's asymptotic equivalence with Bayesian leave-one-out cross-validation (LOO-CV), Watanabe (2010).
- It's high degree of approximation to its **target loss**

WAIC, Widely Applicable Information Criteria

$$
\begin{aligned}
\mathrm{WAIC} &= -2\sum_{i=1}^{n} \log p(y_i|\boldsymbol{y}) + 2V \\
&= = -2\sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} + 2V \ ,
\end{aligned}
$$

where

$$V = \sum_{i=1}^{n} \mathrm{Var}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(y_i|\boldsymbol{\theta})) \ .$$

Watanabe showed that $E_Y[\mathrm{WAIC}]$ is an asymptotically unbiased estimator of $E_Y(\mathrm{B})$ where

$$\mathrm{B} = -2\sum_{i=1}^{n} E_{Z_i}\left[\log p_i(z_i|\boldsymbol{y})\right] = -2\sum_{i=1}^{n} E_{Z_i}\left[\log \int p(z_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}\right] \ .$$

This holds under very general conditions, including for non-identifiable, singular and unrealizable models.

LOO-CVL, Leave-one-out Cross-validation

Letting $\boldsymbol{y}_{-i}$ denote the observations with y_i removed, a natural approximation for B is the LOO-CVL estimator

$$\mathrm{CVL} = \sum_{i=1}^{n} \mathrm{CVL}_i \ ,$$

where

$$\begin{aligned} \mathrm{CVL}_i &= -2 \log p(y_i | \boldsymbol{y}_{-i}) \\ &= -2 \log \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta} \ . \end{aligned} \tag{1}$$

CVL has asymptotic bias of $O(1/n)$ as an estimator of B.

LOO-CVL, Leave-one-out Cross-validation

Letting $\boldsymbol{y}_{-i}$ denote the observations with y_i removed, a natural approximation for B is the LOO-CVL estimator

$$\mathrm{CVL} = \sum_{i=1}^{n} \mathrm{CVL}_i \ ,$$

where

$$\begin{aligned} \mathrm{CVL}_i &= -2 \log p(y_i|\boldsymbol{y}_{-i}) \\ &= -2 \log \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta} \ . \end{aligned} \quad (1)$$

CVL has asymptotic bias of $O(1/n)$ as an estimator of B.

But, direct estimation of CVL can be **very** computationally intensive since it requires samples from n posteriors $p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}), i = 1, ..., n$. This direct estimator will be denoted $\widehat{\mathrm{CVL}}$.

Importance sampling approximation to LOO-CVL

$p(y_i|\boldsymbol{y}_{-i})$ can be expressed as the harmonic mean of $p(y_i|\boldsymbol{\theta})$ with respect to the full posterior,

$$p(y_i|\boldsymbol{y}_{-i}) = \left(\int \frac{1}{p(y_i|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \right)^{-1} ,$$

and so $p(y_i|\boldsymbol{y}_{-i})$ can be estimated as

$$\widehat{p}(y_i|\boldsymbol{y}_{-i}) = \frac{S}{\sum_{s=1}^{S} \frac{1}{p(y_i|\boldsymbol{\theta}^{(s)})}} , \tag{2}$$

where $\boldsymbol{\theta}^{(s)}, s = 1, ..., S$, is a sample from $p(\boldsymbol{\theta}|\boldsymbol{y})$. Thus, each $\text{CVL}_i, i = 1, ..., n$ and hence $\mathrm{CVL} = \sum_{i=1}^{n} \mathrm{CVL}_i$ can be estimated from a single posterior sample.

The importance-sampling estimator of CVL will be denoted $\widehat{\mathrm{IS}}\mathrm{CVL}$.

Importance sampling approximation to LOO-CVL

$p(y_i|\boldsymbol{y}_{-i})$ can be expressed as the harmonic mean of $p(y_i|\boldsymbol{\theta})$ with respect to the full posterior,

$$p(y_i|\boldsymbol{y}_{-i}) = \left(\int \frac{1}{p(y_i|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \right)^{-1} ,$$

and so $p(y_i|\boldsymbol{y}_{-i})$ can be estimated as

$$\widehat{p}(y_i|\boldsymbol{y}_{-i}) = \frac{S}{\sum_{s=1}^{S} \frac{1}{p(y_i|\boldsymbol{\theta}^{(s)})}} , \qquad (2)$$

where $\boldsymbol{\theta}^{(s)}, s = 1, ..., S$, is a sample from $p(\boldsymbol{\theta}|\boldsymbol{y})$. Thus, each $\text{CVL}_i, i = 1, ..., n$ and hence $\mathrm{CVL} = \sum_{i=1}^{n} \mathrm{CVL}_i$ can be estimated from a single posterior sample.

The importance-sampling estimator of CVL will be denoted $\widehat{\mathrm{IS}}\mathrm{CVL}$.

Note that (2) can be highly unstable when $\boldsymbol{\theta}^{(s)}$ is in the tails of $p(y_i|\boldsymbol{\theta}^{(s)})$.

Importance sampling approximation to LOO-CVL

It is very useful to quantify the reliability of importance sampling using the notion of effective sample size. The effective sample size is with respect to a sample from $p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})$ for evaluating CVL_i using (1).

For observation i, ESS_i can be calculated as

$$\text{ESS}_i = \frac{n\overline{w_i}^2}{\overline{w_i^2}} \ ,$$

where $w_{si} = p(y_i|\boldsymbol{\theta}^{(s)})^{-1}$ and $\overline{w_i}$ is the mean of the weights $w_{si}, s = 1, ..., S$, and $\overline{w_i^2}$ is the mean of the squared weights $w_{si}^2, s = 1, ..., S$.

Evaluation of predictive loss

Recent work has examined the relative performance of WAIC, CVL and IS-CVL in the context of normal models.

I have been examining their performance with regard to:

- Model focus (i.e., level of hierarchy at which likelihood is specified).
- Use with non-normal data.

Evaluation of predictive loss

Recent work has examined the relative performance of WAIC, CVL and IS-CVL in the context of normal models.

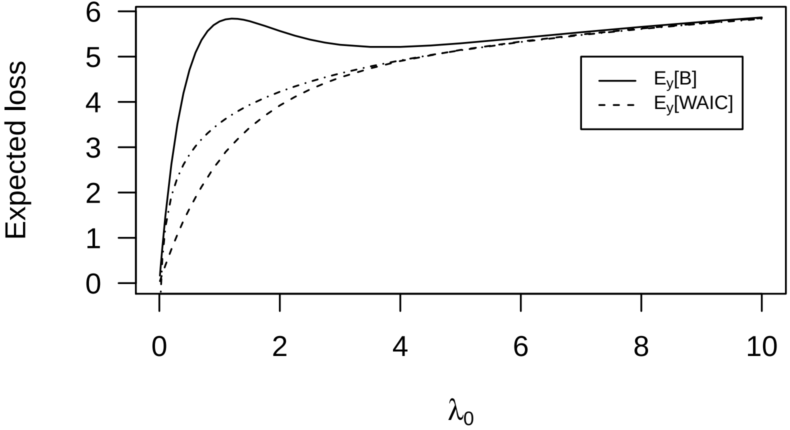
I have been examining their performance with regard to:

- Model focus (i.e., level of hierarchy at which likelihood is specified).
- Use with non-normal data.

Models for over-dispersed count data incorporate both of these issues.

E.g., the negative binomial density can be expressed directly (marginal focus), or as a Poisson density conditional on an underlying gamma latent variable (conditional focus).

Evaluation of predictive loss, $y \sim \mathrm{Pois}(\lambda)$



WAIC approximation not so good until normal approximation (to Poisson) kicks in at around $\lambda_0 = 5$.

Evaluation of predictive loss, $y \sim \mathrm{Pois}(\lambda)$

FYI, the underlying R code to numerically evaluate B for $y \sim \mathrm{Pois}(\lambda_0)$.

```
BayesLoss=function(y,lambda0,alpha=0.001,beta=0.001) {
  yrep_limits=qpois(c(1e-15,1-1e-15),lambda0)
  yrep_grid=seq(yrep_limits[1],yrep_limits[2]) #Grid of values for reps
  grid_probs=dpois(yrep_grid,lambda0) #Probabilities over the grid
  grid_pd=dnbinom(yrep_grid,size=y+alpha,mu=(y+alpha)/(beta+1)) #Pred densi
  BLoss=-2*sum(grid_probs*log(grid_pd)) #Predictive loss, B, for a given y
  return(BLoss) }
```

Simulation study with over-dispersed count data

How well can the predictive criteria distinguish the following three models?

- Poisson: $y_i|\mu \sim \mathrm{Pois}(\mu)$
- PGA: $y_i|\lambda_i \sim \mathrm{Pois}(\lambda_i)$ where $\lambda_i \sim \Gamma(\alpha, \alpha/\mu)$
- PLN: $y_i|\lambda_i \sim \mathrm{Pois}(\lambda_i)$ where $\lambda_i \sim \mathrm{LN}(\log(\mu) - 0.5\tau^2, \tau^2)$

These are conditional-level specifications.

Simulation study with over-dispersed count data

How well can the predictive criteria distinguish the following three models?

- Poisson: $y_i|\mu \sim \text{Pois}(\mu)$
- PGA: $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$ where $\lambda_i \sim \Gamma(\alpha, \alpha/\mu)$
- PLN: $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$ where $\lambda_i \sim \text{LN}(\log(\mu) - 0.5\tau^2, \tau^2)$

These are conditional-level specifications.

For the PLN the marginal-level likelihood is

$$p(y_i|\mu,\tau) = \int \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right) \left(\frac{e^{-(\log \lambda_i - \nu)^2/2\tau^2}}{\sqrt{2\pi}\tau\lambda_i} \right) d\lambda_i \ ,$$

where $\nu = \log(\mu) - 0.5\tau^2$.

Simulation study with over-dispersed count data

How well can the predictive criteria distinguish the following three models?

- Poisson: $y_i|\mu \sim \text{Pois}(\mu)$
- PGA: $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$ where $\lambda_i \sim \Gamma(\alpha, \alpha/\mu)$
- PLN: $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$ where $\lambda_i \sim \text{LN}(\log(\mu) - 0.5\tau^2, \tau^2)$

These are conditional-level specifications.

For the PLN the marginal-level likelihood is

$$p(y_i|\mu,\tau) = \int \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right) \left(\frac{e^{-(\log \lambda_i - \nu)^2/2\tau^2}}{\sqrt{2\pi}\tau\lambda_i} \right) d\lambda_i \ ,$$

where $\nu = \log(\mu) - 0.5\tau^2$.

...or just `dpoilog(y[i],nu,tau)` in R.

Simulation study with over-dispersed count data

The simulation generated $y_i, i = 1, ..., 160$ from each of the three models (using $\mu = 1$ and $\tau = 1.5$), and fitted each of the three models to these data.

$\widehat{\mathrm{WAIC}}_c$ and $\widehat{\mathrm{ISCVL}}_c$ denote the predicted losses estimated using conditional-level likelihood.

Denoted $\widehat{\mathrm{WAIC}}_m$ and $\widehat{\mathrm{ISCVL}}_m$ at marginal level.

It can be shown that:

- CVL_c and CVL_m are identical, and are valid approximations to B_m.
- WAIC_m is a valid approximation to B_m.
- WAIC_c may, or may not, be a valid approximation to B_c.

Simulation study: Conditional-level comparison

True model	Criterion	Fitted model P	PGA	PLN	Propn minimum P	PGA	PLN
P	$\widehat{\text{ISCVL}}_c$	419.1	419.6	419.5	**0.83**	0.10	0.07
	$\widehat{\text{WAIC}}_c$	419.1	419.0	419.1	**0.60**	0.28	0.12
	min ESS	4612	207	1359			
PGA	$\widehat{\text{ISCVL}}_c$	731.0	272.8	291.2	0.00	**0.99**	0.01
	$\widehat{\text{WAIC}}_c$	730.9	219.4	240.1	0.00	**1.00**	0.00
	min ESS	188	2	2			
PLN	$\widehat{\text{ISCVL}}_c$	643.5	374.5	377.4	0.00	0.66	**0.34**
	$\widehat{\text{WAIC}}_c$	644.2	319.0	333.5	0.00	1.00	**0.00**
	min ESS	23	2	2			

Table : Mean values (over 100 simulations) of $\widehat{\text{ISCVL}}$ and $\widehat{\text{WAIC}}$, and hierarchical means of minimum ESS, from fitting Poisson (P), Poisson-gamma (PGA) and Poisson-lognormal (PLN) models to simulated data. The posterior sample size was 5 000.

Simulation study: Marginal-level comparison

True model	Criterion	Fitted model P	PGA	PLN	Propn minimum P	PGA	PLN
P	$\widehat{\mathrm{ISCVL}}_m$	419.1	419.6	419.6	**0.87**	0.06	0.07
	$\widehat{\mathrm{WAIC}}_m$	419.1	419.6	419.6	**0.87**	0.06	0.07
	min ESS	4612	4439	4424			
PGA	$\widehat{\mathrm{ISCVL}}_m$	731.0	345.9	351.2	0.00	**0.94**	0.06
	$\widehat{\mathrm{WAIC}}_m$	730.9	345.9	351.2	0.00	**0.94**	0.06
	min ESS	188	1070	4166			
PLN	$\widehat{\mathrm{ISCVL}}_m$	643.5	412.8	406.6	0.00	0.20	**0.80**
	$\widehat{\mathrm{WAIC}}_m$	644.2	412.6	406.5	0.00	0.20	**0.80**
	min ESS	23	40	952			

Table : Mean values (over 100 simulations) of $\widehat{\mathrm{ISCVL}}$ and $\widehat{\mathrm{WAIC}}$, and hierarchical means of minimum ESS, from fitting Poisson (P), Poisson-gamma (PGA) and Poisson-lognormal (PLN) models to simulated data. The posterior sample size was 5 000.

Application to counts of goatfish

Application to counts of goatfish

	Fitted model			
Criterion	P	PGA	PLN	Δ
Conditional				
$\widehat{\mathrm{CVL}}_c$	482.1	349.7	355.1	5.4
$\widehat{\mathrm{ISCVL}}_c$	479.8	319.9	328.7	8.8
$\widehat{\mathrm{WAIC}}_c$	477.5	273.9	286.0	12.1
min ESS	14.3	4.3	1.5	
Marginal				
$\widehat{\mathrm{CVL}}_m$	482.1	349.7	355.1	5.4
$\widehat{\mathrm{ISCVL}}_m$	479.8	349.6	355.1	5.5
$\widehat{\mathrm{WAIC}}_m$	477.5	348.2	354.5	6.3
min ESS	14.3	189.7	2108.6	

Table : $\widehat{\mathrm{CVL}}$, $\widehat{\mathrm{ISCVL}}$, $\widehat{\mathrm{WAIC}}$ and minimum effective sample size from fitting Poisson (P), Poisson-gamma (PGA) and Poisson-lognormal (PLN) models to goatfish count data. Δ gives the difference between the PGA and PLN losses. The posterior sample size was 10 000.

Summary: Take home advice

- Use marginal-level likelihood where possible (it has fatter tails than conditional-level likelihood).
- Here, $\widehat{\mathrm{CVL}}_c$ was reliable at conditional level.
- Be sure to check effective sample size if using $\widehat{\mathrm{IS}}\mathrm{CVL}$ (an ESS in the 100's appeared to be enough).
- Regularized forms of $\widehat{\mathrm{IS}}\mathrm{CVL}$ were examined, but did not provide any improvement.
- It is a good idea to evaluate both $\widehat{\mathrm{IS}}\mathrm{CVL}$ and $\widehat{\mathrm{WAIC}}$ - and hope that they are little different (since they are different approximations to the same thing).
- WAIC can be unreliable if $\mathrm{Var}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(y_i|\boldsymbol{\theta})) > 1$ for any i (this corresponds to a high influence point and the underlying WAIC approximation to B is liable to be inaccurate).