Beyond DIC:

New developments in Bayesian model comparison

Russell Millar
University of Auckland

Nov 2014

Notation

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

Expected predictive loss

The predictive deviance for future observations, $-2 \log p(\mathbf{z}|\boldsymbol{\theta})$, is a commonly used loss function.

We don't know $\boldsymbol{\theta}$ or $\mathbf{z}$, so use the expected (with respect to future observations) posterior mean of this predictive deviance

$$\mathrm{G}(\mathbf{y}) = -2 E_Z E_{\boldsymbol{\theta}|\mathbf{y}} [\log p(\mathbf{z}|\boldsymbol{\theta})] = -2 E_Z \left[\int \log p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right] .$$

DIC is motivated by the idea that $\mathrm{G}(\mathbf{y})$ can be estimated using the within-sample version:

$$\overline{D(\boldsymbol{\theta})} = -2 E_{\boldsymbol{\theta}|\mathbf{y}} [\log p(\mathbf{y}|\boldsymbol{\theta})] = -2 \int \log p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} .$$

Note that $\overline{D(\boldsymbol{\theta})}$ uses the data twice, and hence underestimates $\mathrm{G}(\mathbf{y})$.

The Dirty information criterion, DIC

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 $\overline{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}|\mathbf{y}}[\boldsymbol{\theta}]$ suggests that p can be estimated as the posterior expected value of $D(\boldsymbol{\theta}) - D(\overline{\boldsymbol{\theta}})$, giving

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

Yikes! Not invariant to re-parameterization due to use of $\overline{\boldsymbol{\theta}}$. ☹☹☹

Also, p_D can be negative if deviance is not concave. ☹☹☹

The Dirty information criterion, DIC

If $D(\boldsymbol{\theta}) - D(\bar{\boldsymbol{\theta}})$ has an approximate chi-square distribution then its posterior variance is approximately twice its posterior mean, leading to the alternative estimate

$$\begin{aligned} p_V &= 0.5\,\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(D(\boldsymbol{\theta})) \\ &= 2\,\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log p(\mathbf{y}|\boldsymbol{\theta}))\ . \end{aligned}$$

This gives re-parameterization invariance, but is more reliant on the deviance being approximately quadratic in shape, and p_V can be numerically unstable in MCMC simulations.

The Dirty information criterion, DIC

If $D(\boldsymbol{\theta}) - D(\overline{\boldsymbol{\theta}})$ has an approximate chi-square distribution then its posterior variance is approximately twice its posterior mean, leading to the alternative estimate

$$\begin{aligned} p_V &= 0.5\,\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(D(\boldsymbol{\theta})) \\ &= 2\,\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log p(\mathbf{y}|\boldsymbol{\theta}))\ . \end{aligned}$$

This gives re-parameterization invariance, but is more reliant on the deviance being approximately quadratic in shape, and p_V can be numerically unstable in MCMC simulations.

These justifications of DIC assume the model is regular. That is, identifiable with non-singular Fisher information (i.e., Hessian) matrix at $\overline{\boldsymbol{\theta}}$. Then $p \to d$ as $n \to \infty$.

Problems with DIC

Mixture models are known to be problematic for DIC. E.g.,

$$p(y|\mu_1, \mu_2, \sigma_1, \sigma_2, b) = bN(\mu_1, \sigma_1^2) + (1-b)N(\mu_2, \sigma_2^2) \ .$$

- Mixture models are not identifiable due to label switching, i.e.,

  $$p(y|\mu_1, \mu_2, \sigma_1, \sigma_2, b) = p(y|\mu_2, \mu_1, \sigma_2, \sigma_1, 1-b)$$

  - although this can be addressed by imposing parameter constraints
- The likelihood is not concave (i.e., deviance is not convex) and hence p_D may be negative and p_V may be erroneous.

Problems with DIC

Several works have argued that DIC under-penalizes model complexity (van-der Linde 2005; Ando, 2007, 2011; Plummer 2008) and have argued the use of

$$\begin{aligned} \mathrm{DIC}^* &= \mathrm{DIC} + p \\ &= \overline{D(\boldsymbol{\theta})} + 2p \ . \end{aligned}$$

DIC^* can be justified on the basis that it is the unbiased estimator of the *unconditional* expected predictive loss

$$\mathcal{G}(n) = E_Y[\mathrm{G}(\mathbf{y}))] = -2E_Y E_Z E_{\boldsymbol{\theta}|\mathbf{y}} [\log p(\mathbf{z}|\boldsymbol{\theta})] \ .$$

Note the additional expectation with respect to the data $\mathbf{y}$.

DIC is a negatively biased estimator of $\mathcal{G}(n)$.

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.

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 calls $\overline{D(\boldsymbol{\theta})}$ Gibbs training loss, and denotes it G_T. He defined

$$\mathrm{WAIC}_G = \mathrm{G}_T + 2V = \overline{D(\boldsymbol{\theta})} + 2V$$

where

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

Watanabe showed that $E_Y[\mathrm{WAIC}_G]$ is an asymptotically unbiased estimator of $\mathcal{G}(n)$ under very general conditions, including for singular and unrealizable models.

For regular realizable models, $V \to d$.

Widely Applicable Information Criteria

Watanabe also considered the unconditional expected predictive loss

$$\mathcal{B}(n) = E_Y(\mathrm{B}(\mathbf{y})) \ ,$$

where

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

Define $\mathrm{WAIC}_B = B_T + 2V$, where

$$\begin{aligned} \mathrm{B}_T &= -2\sum_{i=1}^{n} \log p(y_i|\mathbf{y}) \\ &= -2\sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ . \end{aligned}$$

Watanabe showed that $E_Y[\mathrm{WAIC}_B]$ is asymptotically unbiased for $\mathcal{B}(n)$.

WAIC_B and Bayesian leave-one-out cross validation

Proofs in Watanabe (2009) are very inaccessible.

However, Watanabe (2010) showed that WAIC_B is asymptotically equivalent to Bayesian leave-one-out cross-validation loss.

WAIC_B and Bayesian leave-one-out cross validation

Proofs in Watanabe (2009) are very inaccessible.

However, Watanabe (2010) showed that WAIC_B is asymptotically equivalent to Bayesian leave-one-out cross-validation loss.

Define
$$\mathcal{F}_i(\alpha) = -\log \int p_i(y_i|\boldsymbol{\theta})^\alpha \prod_{j \neq i} p_j(y_j|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \ .$$

Then,

$$-2 \log p(y_i|\mathbf{y}_{-i}) = -2 \log \frac{\int \prod_{i=1}^n p_i(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \prod_{j \neq i} p_j(y_j|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = 2(\mathcal{F}_i(1) - \mathcal{F}_i(0))$$

$$-2 \log p(y_i|\mathbf{y}) = -2 \log \frac{\int p_i(y_i|\boldsymbol{\theta})^2 \prod_{j \neq i} p_j(y_j|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \prod_{i=1}^n p_i(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = 2(\mathcal{F}_i(2) - \mathcal{F}_i(1))$$

WAIC_B and Bayesian leave-one-out cross validation

Proofs in Watanabe (2009) are very inaccessible.

However, Watanabe (2010) showed that WAIC_B is asymptotically equivalent to Bayesian leave-one-out cross-validation loss.

Define
$$\mathcal{F}_i(\alpha) = -\log \int p_i(y_i|\boldsymbol{\theta})^{\alpha} \prod_{j \neq i} p_j(y_j|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \ .$$

Then,

$$-2 \log p(y_i|\mathbf{y}_{-i}) = -2 \log \frac{\int \prod_{i=1}^{n} p_i(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \prod_{j \neq i} p_j(y_j|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = 2(\mathcal{F}_i(1) - \mathcal{F}_i(0))$$

$$-2 \log p(y_i|\mathbf{y}) = -2 \log \frac{\int p_i(y_i|\boldsymbol{\theta})^2 \prod_{j \neq i} p_j(y_j|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \prod_{i=1}^{n} p_i(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = 2(\mathcal{F}_i(2) - \mathcal{F}_i(1))$$

The equivalence is deduced from Taylor series expansions around $\alpha = 1$. The second order difference between $-2 \log p(y_i|\mathbf{y})$ and $-2 \log p(y_i|\mathbf{y}_{-i})$ is $-2\mathcal{F}_i^{''}(1) = 2\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}\left(l_i(\boldsymbol{\theta})\right)$.

Closing remarks: Where to from here?

- Current work is looking at extending WAIC to models where y_i are not conditionally independent. E.g., times series, spatial networks.

Closing remarks: Where to from here?

- Current work is looking at extending WAIC to models where y_i are not conditionally independent. E.g., times series, spatial networks.
- For hierarchical models, $\mathbf{y}$ can be partitioned into conditionally independent groups (one level below that of the model focus).

Closing remarks: Where to from here?

- Current work is looking at extending WAIC to models where y_i are not conditionally independent. E.g., times series, spatial networks.
- For hierarchical models, $\mathbf{y}$ can be partitioned into conditionally independent groups (one level below that of the model focus).
- In worst case, where $\mathbf{y}$ considered as 1 independent group, V reduces to p_V. That is, WAIC_G reduces to DIC^* in worst case.

Closing remarks: Where to from here?

- Current work is looking at extending WAIC to models where y_i are not conditionally independent. E.g., times series, spatial networks.
- For hierarchical models, $\mathbf{y}$ can be partitioned into conditionally independent groups (one level below that of the model focus).
- In worst case, where $\mathbf{y}$ considered as 1 independent group, V reduces to p_V. That is, WAIC_G reduces to DIC^* in worst case.
- WAIC_B has been used in a handful of published works and appears to be the more popular of the two WAICs - likely due to its equivalence with Bayesian LOO-CV.

Closing remarks: Where to from here?

- Current work is looking at extending WAIC to models where y_i are not conditionally independent. E.g., times series, spatial networks.
- For hierarchical models, $\mathbf{y}$ can be partitioned into conditionally independent groups (one level below that of the model focus).
- In worst case, where $\mathbf{y}$ considered as 1 independent group, V reduces to p_V. That is, WAIC_G reduces to DIC^* in worst case.
- WAIC_B has been used in a handful of published works and appears to be the more popular of the two WAICs - likely due to its equivalence with Bayesian LOO-CV.
- However, WAIC_B has been shown to be asymptotically equivalent to DIC for regular realizable models, and DIC is known to overfit. So, there may be some justification for preferring WAIC_G (i.e., it may be better to target $\mathcal{G}(n)$ rather than $\mathcal{B}(n)$).

More widely applicable information criterion?

$$\mathrm{MWAIC}_B = -2\sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\theta}, \mathbf{y}_{-i}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} + 2\sum_{i=1}^{n} \mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log p(y_i|\boldsymbol{\theta}, \mathbf{y}_{-i})) \ . \tag{1}$$