

# Posterior Predictive Criteria for Model Comparison

ANGELIKA VAN DER LINDE

University of Bremen, Germany

October 2005

1. Introduction
2. Posterior predictive model comparison
3. Posterior predictive entropy  
and coefficients of determination
4. Model fit and model complexity
5. Discussion

# 1. Introduction

## 1.1 Preliminaries

$Y$  dependent variable

$X = (X_1 \dots X_p)^T$  independent random vector

$p(x, y|\theta)$  joint density

prior:  $p(\theta)$ , random variable:  $\vartheta_{prior}$

model:  $p(x, y, \theta) = p(x, y|\theta)p(\theta)$

specification depends on sampling scheme

under *conditional* sampling

$$p(x, y, \theta) = p(y|x, \theta)p(x)p(\theta)$$

$p(x)$  empirical, not estimable

observations according to *design*  $d$

reduction under conditional sampling:

$$p(y_d, x_d|\theta) = p(y_d|x_d, \theta), \quad p(x_d) = 1$$

posterior:  $p(\theta|data)$ , random variable  $\vartheta_{post}$

## 1.2 Prediction

$X$  should “explain”  $Y$

interest in prediction of  $Y$  based on  $X$  and  $\vartheta$

prior prediction ( $\rightarrow$  marginal density)

current observations (data) and  $\vartheta_{prior}$

posterior prediction ( $\rightarrow$  predictive density )

future observations (replicated experiment) and  $\vartheta_{post}$

notation:

future observations resulting from the same experiment

$\tilde{X}, \tilde{Y}$  joint sampling

$\tilde{Y}_d$  conditional sampling (given  $x_d$ )

## 2. Posterior predictive model comparison

### 2.1 General approach

models  $M_i \in \mathfrak{M}$  of rv  $Z$  for comparison

with densities  $p_i(z|\theta_i)$  and predicting densities  $\tilde{p}_i$

options to specify  $\tilde{p}_i$

- *posterior predictive* :  $\tilde{p}_i(\tilde{z}) = \bar{p}_i(\tilde{z}) = E_{\vartheta_{post}^i} p_i(\tilde{z}|\theta_i)$
- with *estimate*  $\hat{\theta}_i$  :
  - e.g.  $\hat{\theta}_i = \bar{\theta}_i =$  posterior mean,  $\tilde{p}_i(\tilde{z}) = p_i(\tilde{z}|\bar{\theta}_i)$
  - e.g.  $\hat{\theta}_i = \theta_i^{mod} =$  posterior mode,  $\tilde{p}_i(\tilde{z}) = p_i(\tilde{z}|\theta_i^{mod})$

*utility* of model  $M_i$  if  $\tilde{z}$  occurs:

$$u(M_i, \tilde{z}) = \log \tilde{p}_i(\tilde{z})$$

*expected utility*  $\bar{u}_q(M_i)$  of model  $M_i$

under *actual belief*  $q$  :

$$\bar{u}_q(M_i) = \int q(\tilde{z}) \log \tilde{p}_i(\tilde{z}) d\tilde{z}$$

options to specify actual belief  $q$  :

- *model average*, requires probabilities on models  
e.g.  $q(\tilde{z}) = \sum_j \tilde{p}_j(\tilde{z})P(M_j|data)$   
e.g. parameterizing models by hyperparameters
- more *complex model*  
e.g. encompassing model  
e.g. semi-parametric model
- *avoidance*  
e.g. cross-validation  
e.g. current belief:  $q = \tilde{p}_i$

any combination of  
predicting density and current belief  
yields criterion for model comparison

application under *conditional sampling*:

$$\tilde{Z} \simeq \tilde{Y}_d \text{ (given } x_d)$$

examples

<i>criterion</i>		<i>actual</i>	<i>predicting</i>
$H_i(\tilde{Y}_d)$	$\doteq$	$q = \bar{p}_i$	$\tilde{p}_i(\tilde{y}_d) = \bar{p}_i(\tilde{y}_d)$
$DIC_i$	$\doteq$	$q = \bar{p}_i$	$\tilde{p}_i(\tilde{y}_d) = p_i(\tilde{y}_d \bar{\theta}_i)$

application under *joint sampling*:

$$\tilde{Z} \simeq \tilde{Y}|\tilde{X} \text{ and average wrt } \tilde{X}$$

example

<i>critierion</i>	<i>actual</i>	<i>predicting</i>
$H_i(\tilde{Y} \tilde{X})$	$q(\tilde{x}, \tilde{y}) = \bar{p}_i(\tilde{x}, \tilde{y})$	$\tilde{p}_i(\tilde{y}) = \bar{p}_i(\tilde{y} \tilde{x})$

explicitly:

Deviance Information Criterion

$$DIC_i = -2E_{\tilde{Y}_d|y_d} \log p_i(\tilde{y}_d|\bar{\theta}) = -2E_{\vartheta_{post}}[E_{\tilde{Y}_d|\theta} \log p_i(\tilde{y}_d|\bar{\theta})]$$

posterior predictive entropy (conditional sampling)

$$2H_i(\tilde{Y}_d) = -2E_{\tilde{Y}_d|y_d} \log \bar{p}_i(\tilde{y}_d) = -2E_{\vartheta_{post}}[E_{\tilde{Y}_d|\theta} \log \bar{p}_i(\tilde{y}_d)]$$

posterior predictive entropy (joint sampling)

$$2H(\tilde{Y}|\tilde{X}) = -2E_{\tilde{X}|y_d} E_{\tilde{Y}|\tilde{x},y_d} \log \bar{p}_i(\tilde{y}|\tilde{x})$$

interpretation: using  $\vartheta_{post}$

minimize uncertainty about  $\tilde{Y}_d$  ( $\tilde{Y}$ ) given  $x_d$  ( $\tilde{x}$ )

### 3. Posterior predictive entropies and coefficients of determination

Kullback-Leibler discrepancy between densities  $f$  and  $g$

$$D_{KL}(f, g) = E_f \log\left(\frac{f}{g}\right)$$

interpret  $D_{KL}(\tilde{p}_i, \tilde{g})$  as general (posterior) coefficient  
of determination (of  $Y$  by  $X$  in model  $M_i$ )

for a choice of  $g$  representing independence of  $X$  and  $Y$

under joint sampling:  
 if models differ in specification of dependence, i.e.  
 with  $p_i(x, y|\theta_i) = p_i(y|x, \theta_i)p(x)$ ,  $g(x, y) = g(y)p(x)$

$$D_{KL}(\tilde{p}_i, \tilde{g}) = -E_{\tilde{p}_i} \log \tilde{g}_y(\tilde{y}) - H_i(\tilde{Y}|\tilde{X})$$

under conditional sampling with  $g(y_d)$   
 reduction to

$$D_{KL}(\tilde{p}_i, \tilde{g}) = -E_{\tilde{p}_i} \log \tilde{g}(\tilde{y}_d) - H_i(\tilde{Y}_d)$$

if  $H_i(\tilde{Y}|\tilde{X})$  or  $H_i(\tilde{Y}_d)$  small,  
 coefficient of determination large  
 though optimizations (“absolute” vs “relative”)  
 not equivalent

## Question

(1) model comparison  
based on posterior predictive entropy  
comes close to model comparison  
based on coefficients of determination  
emphasizing “model fit”  
rather than “model complexity”

(2) model comparison  
based on posterior predictive entropy  
comes close to model comparison  
based on DIC (special case: AIC)  
trading off “model fit” and “model complexity”

how to strike the balance ??? subtle discussion!

## 4. Model fit and model complexity

basic *decomposition*

$$H(Z) = H(Z|\vartheta) + I(Z, \vartheta)$$

where  $I(Z, \vartheta) = D_{KL}(p(z, \theta), p(z)p(\theta)) =$  mutual info  
between  $Z$  and  $\vartheta$

application:

$$\begin{aligned} H_i(\tilde{Y}_d) &= H_i(\tilde{Y}_d|\vartheta_{post}) + I_i(\tilde{Y}_d, \vartheta_{post}) \\ &= -E_{\vartheta_{post}}[E_{\tilde{Y}_d|\theta} \log p_i(\tilde{y}_d|\theta)] + I_i(\tilde{Y}_d, \vartheta_{post}) \\ &\approx -E_{\vartheta_{post}}[\log p_i(y_d|\theta)] + I_i(\tilde{Y}_d, \vartheta_{post}) \\ &= -\text{“adequacy”} \quad + \quad \text{penalty for complexity} \end{aligned}$$

define

$$D_i(\theta) = -2 \log p_i(y_d|\theta), \quad \bar{D}_i = E_{\vartheta_{post}} D_i(\theta), \\ p_{D_i} = \bar{D}_i - D_i(\bar{\theta})$$

then

$$2H_i(\tilde{Y}_d) \approx \bar{D}_i + 2I_i(\tilde{Y}_d, \vartheta_{post}) \\ = D_i(\bar{\theta}) + p_{D_i} + 2I_i(\tilde{Y}_d, \vartheta_{post})$$

note:

$$\bar{D}_i = D_i(\bar{\theta}) + p_{D_i} \\ \text{“adequacy”} = \text{“fit”} + \text{“complexity”}$$

#### 4.1 DIC (conditional sampling)

decomposition for  $p(\tilde{z}|\bar{\theta})$  :

$$-2E_{\tilde{Z}|data} \log p(\tilde{z}|\bar{\theta}) = 2H(\tilde{Z}|\vartheta_{post}) + 2E_{\vartheta_{post}^i} D_{KL}(p_i(\tilde{z}|\theta)||p_i(\tilde{z}|\bar{\theta}))$$

where

$$\begin{aligned} & 2E_{\vartheta_{post}^i} D_{KL}(p_i(\tilde{z}|\theta)||p_i(\tilde{z}|\bar{\theta})) \\ & \underset{Taylor2}{\approx} E_{\vartheta_{post}^i} D_S(p_i(\tilde{z}|\theta)||p_i(\tilde{z}|\bar{\theta})) \\ & : = E_{\vartheta_{post}^i} D_{KL}(p_i(\tilde{z}|\theta)||p_i(\tilde{z}|\bar{\theta})) + E_{\vartheta_{post}^i} D_{KL}(p_i(\tilde{z}|\bar{\theta})||p_i(\tilde{z}|\theta)) \end{aligned}$$

and in expofams

$$E_{\vartheta_{post}^i} D_S(p_i(\tilde{z}|\theta)||p_i(\tilde{z}|\bar{\theta})) = J(\tilde{Z}, \vartheta_{post}) := E_{\vartheta_{post}^i} D_S(p_i(\tilde{z}|\theta)||\bar{p}_i(\tilde{z}))$$

application:  $\tilde{Z} = \tilde{Y}_d$

$$\begin{aligned} -2E_{\tilde{Y}_d|y_d} \log p_i(\tilde{y}_d|\bar{\theta}) &= 2H_i(\tilde{Y}_d|\vartheta_{post}) + 2E_{\vartheta_{post}^i} D_{KL}(p_i(\tilde{y}_d|\theta)||p_i(\tilde{y}_d|\bar{\theta})) \\ \underset{DIC_i}{=} &= \bar{D}_i + p_{D_i} \\ &= D_i(\bar{\theta}) + 2p_{D_i} \end{aligned}$$

## 4.2 Comparison and extension

(subscript i omitted)

in expofams

$p_D \geq 0$

$p_D$  estimates  $J(\tilde{Y}_d, \vartheta_{post}) \underset{Taylor2}{\approx} 2E_{\vartheta_{post}^i} D_{KL}(p_i(\tilde{z}|\theta) || p_i(\tilde{z}|\bar{\theta})) \geq 2I(\tilde{Y}_d, \vartheta_{post})$

in general

rhs  $2H(\tilde{Y}_d|\vartheta_{post}) + J(\tilde{Y}_d, \vartheta_{post})$

corresponds to

odd lhs  $-E_{\vartheta_{post}} [E_{\tilde{Y}_d|\theta} \log p(\tilde{y}_d|\theta) + E_{\tilde{Y}_d|y_d} \log p(\tilde{y}_d|\theta)]$

in DIC-target

posterior variance of  $\theta$  not fully acknowledged

extension DIC+:

$$-2E_{\vartheta_{post}} E_{\tilde{Y}_d|y_d} \log p(\tilde{y}_d|\theta) = 2H(\tilde{Y}_d|\vartheta_{post}) + \boxed{2}J(\tilde{Y}_d, \vartheta_{post})$$

comparison DIC -  $H(\tilde{Y}_d)$

$$\begin{aligned}
2H(\tilde{Y}_d) : & \quad -2E_{\tilde{Y}_d|y_d}[\log p(\tilde{y}_d|y_d)] & = & \quad 2H(\tilde{Y}_d|\vartheta_{post}) & + & \quad 2I(\tilde{Y}_d, \vartheta_{post}) \\
DIC : & \quad -2E_{\tilde{Y}_d|y_d}[\log p(\tilde{y}_d|\bar{\theta})] & & & & \\
expofams: & & \approx & \quad 2H(\tilde{Y}_d|\vartheta_{post}) & + & \quad J(\tilde{Y}_d, \vartheta_{post}) \\
& & & \quad \text{'adequacy'} & + & \quad \text{'complexity'} \\
& & \approx & \quad \bar{D} & + & \quad p_D \\
& & (\approx & \quad D(\bar{\theta}) & + & \quad 2p_D) \\
\nabla \text{ expofams: } & \quad H(\tilde{Y}_d|\vartheta_{post}) - E_{\vartheta_{post}}E_{\tilde{Y}_d|y_d}[\log(\tilde{y}_d|\theta)] & = & \quad 2H(\tilde{Y}_d|\vartheta_{post}) & + & \quad J(\tilde{Y}_d, \vartheta_{post}) \\
DIC+ : & \quad -2E_{\vartheta_{post}}E_{\tilde{Y}_d|y_d}[\log(\tilde{y}_d|\theta)] & = & \quad 2H(\tilde{Y}_d|\vartheta_{post}) & + & \quad 2J(\tilde{Y}_d, \vartheta_{post}) \\
expofams: & & \approx & \quad \bar{D} & + & \quad 2p_D \\
& & (\approx & \quad D(\bar{\theta}) & + & \quad 3p_D) \\
AIC : & \quad -2E_{Y_d}E_{\tilde{Y}_d}[\log p(\tilde{y}_d|\hat{\theta}_{ML}(y_d))] & & & & \\
& & \approx & \quad -2\log p(y_d|\hat{\theta}_{ML}(y_d)) & + & \quad 2p
\end{aligned}$$

## 5. Discussion

“true” vs current belief:

maximization of expected utility

$\Leftrightarrow$  minimization of KL-discrepancy  $D_{KL}(q, \tilde{p}_i)$

original motivation of AIC

use of current belief: “good model assumption”

## 5.1 Targets in model comparison

$$\begin{array}{l} \text{penalties for model complexity} \\ H(\tilde{Y}_d) : I(\tilde{Y}_d, \vartheta_{post}) \\ DIC : J(\tilde{Y}_d, \vartheta_{post}) \\ DIC+ : 2J(\tilde{Y}_d, \vartheta_{post}) \end{array}$$

where  $I \stackrel{\text{expofams, Taylor2}}{\leq} J \leq 2J$

experience of AIC (limiting DIC):  
 $J$  not always sufficient

## 5.2 Estimates

model adequacy  $H(\tilde{Y}_d|\vartheta_{post})$  estimated by  $\bar{D}$

model complexity

- $p_D$  valid in *expofams* (log concave densities)
- $p_D$  not valid for *finite mixtures* (neg values)  
possibly promising (Richardson, 2002)

$$p'_D = \bar{D} + 2 \log(E_{\vartheta_{post}} p(y_d|\theta))$$

- no *generally* valid simple estimate yet of

$$E_{\vartheta_{post}^i} D(p(\tilde{y}_d|\theta)||p(\tilde{y}_d|\bar{\theta})) \quad \text{or} \quad J(\tilde{Y}_d, \vartheta_{post})$$

Beirlant et al (1997); Paninski, Neural Computing (2003)

naive MC estimates often with bias and large variance

### 5.3 Example (clustering)

multinomial response

$$Y \in \{1, \dots, K\}, \quad Y \sim M(1, \pi), \quad Y|x \sim M(1, \pi(x))$$

joint sampling (discrimination):

$$p(x, y|\theta) = p(x|y, \theta)p(y|\theta)$$

$$p(x|j, \theta) \doteq N(\mu_j, \Sigma_j), \quad \theta = (\mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K, \pi)$$

marginal sampling of  $X$  (clustering):

$$p(x, y|\theta) = p(y|x, \theta)p(x|\theta)$$

$$\begin{aligned} p(x|\theta) &= \sum_{j=1}^K p(x|j, \theta)p(j|\theta) \\ &= \pi_j p_{N(\mu_j, \Sigma_j)}(x) + \left(1 - \sum_{j=1}^{K-1} \pi_j\right) p_{N(\mu_K, \Sigma_K)}(x) \end{aligned}$$

$$\theta = (\mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K, \pi)$$

$$p(y|x, \theta) = p(x|y, \theta)p(y|\theta)/p(x|\theta)$$

interpretation of entropies

finite mixture

$$p(x) = \sum_{j=0}^1 p(x|j)P(Y = j)$$

measure of dependence

$$I(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

minimize  $H(Y|X)$  = “measure of overlap”

minimize  $H(X|Y)$   $\Leftrightarrow$  maximize “homogeneity of clusters”

criteria for model choice

number of clusters  $K$

uncertainty: which prediction ??

entropy based

(i) focus on  $X$

not targeted to clustering

(ii) focus on  $Y$

with ML-estimate

$$\text{“}NEC\text{”} \triangleq \frac{H(Y|X, \hat{\theta})}{H(X, Y|\hat{\theta})}$$

(Celeux and Soromenho; 1996), experience:

NEC does not optimize “shape of the model”

i.e. within cluster distributions

(iii) focus on  $X$  and  $Y$   
prior predictive (BIC)

$$\text{"ICL"} \doteq H(X, Y|\theta) = \underbrace{H(X|\theta)}_{\substack{\text{adequacy} \\ X \text{ observed}}} + \underbrace{H(Y|X, \theta)}_{\substack{\text{penalty} \\ \text{for overlap}}}$$

(Biernacki, Celeux, Govaert; 2000), experience:  
complexity neglected  
does not always detect true  $K$   
→ *true versus predictive* model

predictive: DIC

(i) focus on  $X$

$p(x)$  *observed* likelihood (finite mixture)

$p(x|y, \theta)$  *conditional* likelihood (often expofam)

(ii) focus on  $Y$

$p(y|x, \theta)$  not yet considered

(iii) focus on  $X$  and  $Y$

$p(x, y|\theta)$  *completed* likelihood (non expofam)

current research (Celeux, Robert, Titterington et al)