

DIC in variable selection

Angelika van der Linde

*Institute of Statistics, University of Bremen,
PO Box 330440, 28334 Bremen, Germany
Email: avdl@math.uni-bremen.de*

Abstract

Model comparison is discussed from an information theoretic point of view. In particular the posterior predictive entropy is related to the target yielding DIC and modifications thereof. The adequacy of criteria for posterior predictive model comparison is also investigated depending on the comparison to be made. In particular variable selection as a special problem of model choice is formalized in different ways according to whether the comparison is a comparison across models or within an encompassing model and whether a joint or conditional sampling scheme is applied. DIC has been devised for comparisons across models. Its use in variable selection and that of other criteria is illustrated for a simulated data set.

Key words: posterior predictive entropy, mutual information, model comparison, hypothesis testing

Running head: DIC

1 Introduction

The Deviance Information Criterion (DIC) was introduced by SPIEGELHALTER et al. (2002) as an easily computable and rather universally applicable Bayesian criterion for posterior predictive model comparison. Like many other criteria it compromises between data fit and model complexity, and it generalizes AIC which appears as a special case under a vague prior. Although a ‘decision theoretic’ derivation of DIC was given and the term of model complexity further analyzed in (VAN DER LINDE, 2004), DIC is still not yet fully understood. For example, experiences indicating that DIC works well if the sampling distribution belongs to an exponential family but less so if it is a mixture are hard to explain, and hence modifications are hard to justify. In this note DIC is investigated as special estimate of an information theoretic target.

A natural information theoretic approach in regression analysis is to consider the mutual information between (future) observations and the covariates (a measure of dependence) as a major quantity of interest. This quantity is different from an expected utility, and in general information based criteria provide an alternative to decision theoretically derived criteria. Several commonly used criteria for model comparison may be interpreted from this point of view, and DIC is put into this perspective.

Which criterion can be used also depends on how the models to be compared are set up. Under the heading of variable selection as a special problem of model choice at least three different problems are addressed:

- within an encompassing model the problem, whether a submodel predicts future observations as well as the full model:

(i) under conditional sampling this problem is posed as a problem of testing the hypothesis that some of the regression coefficients are equal to zero, and information theory then provides useful test statistics;

(ii) under joint sampling this problem amounts to the problem of identifying a parsimonious submodel based on the mutual information between the response variable and subsets of covariates defining the submodels;

(iii) - across models the problem, which of the competing models based on different subsets of covariates is to be preferred.

DIC is a criterion applicable to version (iii) of the problem of variable selection.

In this paper the cross-classification of criteria for variable selection related to the idea of mutual information with set-ups for model comparison ((i)-(iii)) is described in general and, given this background, DIC in variable selection is investigated as a particular instance. The paper is organized as follows. Basic definitions are introduced in section 2. In section 3 problems of variable selection are discussed where the notion of encompassing models is interpreted in terms of coherent marginal distributions including those of the parameters rather than just nested models. In section 4 DIC is re-examined within an information theoretic framework. A brief discussion in section 5 concludes the paper.

2 Dependence between covariates and response

2.1 Preliminaries

Let X denote a vector of covariates, Y a (for convenience) univariate response variable and ϑ a random parameter taking values θ . Let further (X_d, Y_d) denote the matrix of all variables observed in an experiment according to an experimental design d . The joint probability density $p(x, y, \theta)$ may be specified in one of two ways,

$$p(x, y, \theta) = p(x, y|\theta)p(\theta), \quad (1)$$

or by the factorization

$$p(x, y, \theta) = p(y|x, \theta)p(x)p(\theta) \quad (2)$$

corresponding to joint or conditional sampling respectively. In both settings X is regarded to be random although under conditional sampling which is frequent in

regression analysis, the distribution of X may not be estimable and may have to be substituted by the empirical distribution according to the experimental design.

In order to distinguish between prior and posterior distribution and the corresponding random variables a subscript will be used. Hence ϑ_{prior} denotes the random variable induced by the prior distribution and ϑ_{post} is the random variable corresponding to the posterior distribution. Let (\tilde{X}, \tilde{Y}) denote a vector with the same conditional (given θ) distribution as (X, Y) yielding a single future observation. Similarly let $(\tilde{X}_d, \tilde{Y}_d)$ be the matrix yielding future observations if the whole experiment is repeated. In posterior predictive model assessment we are interested in the joint density of $(\tilde{X}, \tilde{Y}, \vartheta_{post})$ or $(\tilde{X}_d, \tilde{Y}_d, \vartheta_{post})$. Referring to this joint density in particular the posterior predictive density for the response is simply written as $p(\tilde{y})$ or $p(\tilde{y}_d)$.

Dependence between the response variable, the covariates and the parameters is described by the mutual information. For variables U, V, W and all densities denoted by p , we have the following definitions (KULLBACK, 1968; COVER and THOMAS, 1991): The mutual information between U and V is given by

$$\begin{aligned} I(U, V) & : = \int p(u, v) \log \frac{p(u, v)}{p(u)p(v)} d(u, v) \\ & = H(U) - H(U|V), \end{aligned}$$

where $H(U) = -E_U(\log p(u))$ is the entropy of U and $H(U|V) = E_V(H(U|v)) = -E_V E_{U|v}(\log p(u|v))$ is the conditional entropy of U given V . $I(U, V)$ is the *directed divergence* $D(p(u, v)||p(u)p(v))$ between $p(u, v)$ and $p(u)p(v)$. A symmetrized version is the *divergence*

$$\begin{aligned} J(U, V) & = D(p(u, v)||p(u)p(v)) + D(p(u)p(v)||p(u, v)) \\ & = \int (p(u, v) - p(u)p(v)) \log \frac{p(u, v)}{p(u)p(v)} d(u, v). \end{aligned}$$

The conditional mutual information $I(U, V|W)$ is defined as

$$\begin{aligned} I(U, V|W) & = H(U|W) - H(U|V, W) \\ & = \int p(w)p(u, v|w) \log \frac{p(u, v|w)}{p(u|w)p(v|w)} d(u, v, w). \end{aligned}$$

2.2 The chain rule

A key equation in posterior predictive model assessments is the chain rule (COVER AND THOMAS, 1991, p.22)

$$\begin{aligned} I(\tilde{Y}, (\tilde{X}, \vartheta_{post})) & = I(\tilde{Y}, \tilde{X}) + I(\tilde{Y}, \vartheta_{post}|\tilde{X}) \\ & = I(\tilde{Y}, \vartheta_{post}) + I(\tilde{Y}, \tilde{X}|\vartheta_{post}). \end{aligned} \tag{3}$$

All quantities describe mutual information between future observations of \tilde{Y} and model components (covariates and parameters), and thus all quantify in some way

how the response \tilde{Y} is ‘explained’ by a model specified by (1) or (2). Intuitively the various ways of conditioning are hard to distinguish and careful interpretations are needed. Roughly, the measure of dependence between \tilde{Y} and \tilde{X} has been considered as a quantity of interest in model comparison whereas the measure of dependence between \tilde{Y} and ϑ_{post} quantifies model complexity. Here are some readings.

(1) *Coefficient of determination*

$$I(\tilde{Y}, \tilde{X} | \vartheta_{post}) = E_{\vartheta_{post}}(I(\tilde{Y}, \tilde{X} | \theta)),$$

and $I(\tilde{X}, \tilde{Y} | \theta)$ is related to a coefficient of determination in regression by the transformation $R_Y^2(\theta) = 1 - \exp(-2I(\tilde{X}, \tilde{Y} | \theta))$. For example, if (\tilde{X}, \tilde{Y}) are jointly Gaussian, $I(\tilde{X}, \tilde{Y} | \theta) = \frac{1}{2} \log(\sigma_{\tilde{Y}|\theta}^2) - \frac{1}{2} \log(E_{\tilde{X}}(\sigma_{\tilde{Y}|\tilde{x},\theta}^2))$. The conventional R^2 is obtained plugging in the maximum likelihood variance estimates, that is, without adjustment for the estimation of means in the degrees of freedom. For a comprehensive analysis of coefficients of determination see (VAN DER LINDE and TUTZ, 2004). \square

(2) *Conditional entropy and expected utility*

$$I(\tilde{Y}, \tilde{X}) = H(\tilde{Y}) - H(\tilde{Y} | \tilde{X}).$$

The conditional density $p(\tilde{y} | \tilde{x})$ already specifies the association between \tilde{Y} and \tilde{X} and posterior predictive model comparison might be reduced to the comparison of the entropies $H(\tilde{Y} | \tilde{X})$ (to be minimized). We have

$$-H(\tilde{Y} | \tilde{X}) = E_{\tilde{X}} E_{\tilde{Y} | \tilde{x}}(\log p(\tilde{y} | \tilde{x})) = E_{\tilde{X}} E_{\tilde{Y} | \tilde{x}}(\log E_{\vartheta_{post}}(p(\tilde{y} | \tilde{x}, \theta))). \quad (4)$$

Under conditional sampling (2) with n observations $H(\tilde{Y} | \tilde{X}) = E_X(H(\tilde{Y} | \tilde{x})) = \sum_{\tilde{x}} H(\tilde{Y} | \tilde{x})/n$. Note that the conditional entropy cannot be interpreted as an expected utility w.r.t. an ‘actual belief’ because the posterior predictive distribution yielding the expectation is model specific (see BERNARDO and BERMUDEZ, 1985). Estimating (4) by $\sum_{i=1}^n \log E_{\vartheta_{post}}(p(y_i | x_i, \theta))/n$ does use the data twice but comes close to the cross-validatory estimate $\sum_{i=1}^n \log E_{\vartheta_{post}^{-i}}(p(y_i | x_i, \theta))/n$ of a posterior expected utility if n is not too small. \square

$I(\tilde{Y}, \tilde{X})$ and $I(\tilde{Y}, \tilde{X} | \vartheta_{post})$ account for the variability of ϑ_{post} in different ways. $I(\tilde{Y}, \tilde{X} | \vartheta_{post})$ is merely an average measure of conditional dependence while $I(\tilde{Y}, \tilde{X})$ measures dependence more directly based on the marginal distribution having integrated out θ .

3 Problems of variable selection

We now assume that there are p covariates, $X = (X_1, \dots, X_p)$. Posing a problem of variable selection we ask which covariates are needed to explain or predict the

response. The explanatory value of each subvector $X_{(s)} = (X_{i_1}, \dots, X_{i_k})$ can only be assessed given a model, and three cases are usually distinguished.

An encompassing model including all covariates is fixed, and any comparison is made for submodels,

- (i) under joint sampling,
- (ii) under conditional sampling.
- (iii) Comparisons are to be made across models.

By a model we again mean a joint distribution of (X, Y, ϑ) , not only a conditional distribution of Y given (X, ϑ) . The notion of ‘nested models’, where $p_1(y|x, \theta_1) = p_2(y|x, \theta)$, $\theta = (\theta_1, \theta_2)$, applies in all cases. In particular in an encompassing model there is only one prior distribution, whereas across models different priors are involved.

3.1 Analysis within an encompassing model: joint sampling

As the encompassing model is fixed as a reference, a measure of dimension reduction or simplification may be applied when assessing a subset of covariates. If the encompassing model is given by (1), the mutual information between \tilde{Y} and $\tilde{X}_{(s)}$, respectively the posterior predictive entropy $H(\tilde{Y}|\tilde{X}_{(s)})$ obtained from the marginal distribution of $(\tilde{X}, \tilde{Y}, \vartheta_{post})$ can be used to assess the (posterior) explanatory potential of a subvector of covariates. The (conditional) coefficient of determination $I(\tilde{X}, \tilde{Y}|\theta)$ in frequentist theory as well as $I(\tilde{X}, \tilde{Y})$ in Bayesian theory can be decomposed by the chain rule, now applied to \tilde{X} , for example,

$$\begin{aligned} I(\tilde{X}, \tilde{Y}) &= I(\tilde{Y}, (\tilde{X}_1, \dots, \tilde{X}_p)) \\ &= I(\tilde{Y}, (\tilde{X}_1, \dots, \tilde{X}_{p-1})) + I(\tilde{Y}, \tilde{X}_p | (\tilde{X}_1, \dots, \tilde{X}_{p-1})). \end{aligned}$$

Hence $I(\tilde{X}_{(s)}, \tilde{Y})$ is increasing with the number of covariates. Within an encompassing model and under joint sampling therefore typically submodels are ordered applying forward selection or backward elimination. Alternatively a subset of covariates with explanatory power $I(\tilde{X}_{(s)}, \tilde{Y})$ may be chosen such that a certain high percentage of the explanatory power $I(\tilde{X}, \tilde{Y})$ of all covariates is obtained.

Example

For illustration a simulated data set based on an example introduced in (GEORGE and MCCULLOCH, 1993) is analyzed. $n = 60$ data points were generated from a joint Gaussian distribution of Y and $X = (X_1, \dots, X_5)$ with zero mean vector and a covariance matrix Σ corresponding to the linear regression model

$$Y = X_4 + 1.2X_5 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \quad \sigma^2 = 6.25.$$

Referring to joint sampling the quantity of interest is $I(\tilde{X}_{(s)}, \tilde{Y})$ respectively $H(\tilde{Y}|\tilde{X}_{(s)})$ for subvectors $X_{(s)}$ of X . The prior was chosen as a conjugate Normal-Wishart distribution such that the posterior predictive distribution of $U = (Y, X)^T$ is multivariate Student-t with 66 degrees of freedom and can be approximated by a

multivariate Normal distribution in order calculate the conditional entropies of \tilde{Y} in terms of conditional variances. Formally,

$$\begin{aligned} U &\sim N(\mu, \Sigma), \\ \mu &\sim N(\mu_0, n_0^{-1}\Sigma_0), \quad \mu_0 = 0, \quad n_0 = 2, \\ \Sigma &\sim W(\alpha, \Sigma_0), \quad \alpha = 5.5, \quad \Sigma_0 = \begin{bmatrix} \Sigma_{0,11} & \Sigma_{0,12} \\ \Sigma_{0,21} & \Sigma_{0,22} \end{bmatrix}, \\ \Sigma_{0,11} &= 6, \quad \Sigma_{0,12} = (1, 1, 1, 1, 1) = \Sigma_{0,21}^T, \quad \Sigma_{0,22} = I_5, \end{aligned}$$

such that the prior is moderately informative. The results of forward selection are reported and compared to those based on $I(X_{(s)}, Y|\theta)$ and $I(\tilde{X}_{(s)}, \tilde{Y}|\hat{\theta}_{ML})$, where $\hat{\theta}_{ML}$ denotes the maximum likelihood estimate of θ , in table 1. All forward selections correctly build up the sequence $\{X_5\} \rightarrow \{X_5, X_4\} \rightarrow (\{X_5, X_4, X_1\} \rightarrow \{X_5, X_4, X_1, X_2\}) \rightarrow X$.

Table 1

Values of mutual information and conditional entropies corresponding to forward selection

			(s)		
	{5}	{5, 4}	{5, 4, 1}	{5, 4, 1, 2}	X
$I(\tilde{X}_{(s)}, \tilde{Y})$	0.0854	0.1846	0.1933	0.1976	0.1980
$I(\tilde{X}_{(s)}, \tilde{Y} \hat{\theta}_{ML})$	0.0852	0.1845	0.1919	0.1983	0.1983
$I(X_{(s)}, Y \theta)$	0.0906	0.1648	0.1648	0.1648	0.1648
$H(\tilde{Y} \tilde{X}_{(s)})$	0.9308	0.8316	0.8229	0.8186	0.8182
$H(\tilde{Y} \tilde{X}_{(s)}, \hat{\theta}_{ML})$	0.9583	0.8590	0.8516	0.8452	0.8452
$H(\tilde{Y} \tilde{X}_{(s)}, \theta)$	0.9905	0.9163	0.9163	0.9163	0.9163

In this example $I(\tilde{X}_{(s)}, \tilde{Y}|\hat{\theta}_{ML})$ and $I(\tilde{X}_{(s)}, \tilde{Y})$ are pretty close. Necessarily all quantities are monotone in the number of covariates. The true submodel $\{5, 4\}$ achieves $(0.1846/0.1980=)$ 93.2% of the estimated explanatory power of the full model in terms of the mutual information and $(0.8182/0.8316=)$ 98.4% in terms of the conditional entropy. \triangle

3.2 Analysis within an encompassing model: conditional sampling

In this case variable selection is posed as a problem of hypothesis testing. An adequate quantity of interest for testing $\theta = (\theta_{(s)}, 0)$ is (a variant of a of) the directed divergence from the encompassing model $E_{\vartheta_{post}}(D(p(y_d|x_d, \theta)||p(y_d|x_{(s)d}, \theta_{(s)})))$. See BERNARDO and RUEDA (2002). The quantity $E_{\vartheta_{post}}(D(p(y_d|x_d, \theta)||p(y_d|x_{(s)d}, \theta_{(s)})))$ is a statistic for testing if $X \setminus X_{(s)}$ is not in the model, that is one aims at excluding variables such that the explanatory power of the remaining covariates is not much smaller than that of the full model, the maximum reference model. In contrast, in

a constructive approach, building up a model including more and more covariates, one would quantify the improvement using $X_{(s)}$ with respect to a minimum reference model, $p(y_d|\theta_0)$, say, without any covariates. The corresponding statistic is $D(p(y_d|x_{(s)d}, \theta_{(s)})||p(y_d|\theta_0))$. This comes close in spirit to the coefficient of determination used under joint sampling $I(X_{(s)d}, Y_d|\theta) = D(p(y_d|x_{(s)d}, \theta_{(s)})||p(y_d|\theta))$, where θ is the parameter of the joint distribution of X and Y . Again, such a coefficient of determination would in turn be compared to the maximum achievable one determined by the full model. Hence for comparisons within an encompassing model, model complexity usually is of minor concern. For further discussion see (VAN DER LINDE and TUTZ).

A different approach to the problem of variable selection under conditional sampling is based on indicator variables such that posterior probabilities for the inclusion of single covariates are obtained. See (GEORGE and MCCULLOCH, 1993).

3.3 Analysis across models

Different regression models are typically specified if a sampling distribution is combined with differently informative priors, or if different sampling distributions, for example corresponding to $p(y|x_{(s)}, \theta_{(s)})$ are combined with standard weakly informative priors. Across models the quantities of interest like $I(\tilde{X}, \tilde{Y})$ or $H(\tilde{Y}|\tilde{X})$ are not necessarily monotone if more and more covariates are included in the model. Technically ranking of models according to forward selection or backward elimination may again be tried as search strategy but is not justified by a monotonicity property of the quantity of interest, reflected in its estimate, any more. DIC is a method for model comparison devised for such a set-up, particularly accounting for different prior distributions. It is discussed and compared to the posterior predictive entropy in the next section.

4 Posterior predictive comparisons across models

4.1 Applications of the chain rule

In order to put DIC into the context of information theoretic criteria for posterior predictive model comparison, first the chain rule (3) is further examined. The chain rule induces the typical decomposition of a criterion into a measure of ‘model adequacy’ and a measure of ‘model complexity’. If there is only one condition under which Y (\tilde{Y}) is observed either because there are no covariates in the model or because the response is already defined to be the whole vector Y_d (\tilde{Y}_d) of observations and the one condition is determined by the experimental design, the chain rule (3) rewritten as

$$H(\tilde{Y}|\tilde{X}) = H(\tilde{Y}|\tilde{X}, \vartheta_{post}) + I(\tilde{Y}, \vartheta_{post}|\tilde{X})$$

simplifies to

$$H(\tilde{Y}_d) = H(\tilde{Y}_d|\vartheta_{post}) + I(\tilde{Y}_d, \vartheta_{post}). \quad (5)$$

In this form further readings can be suggested.

(3) *Posterior Bayes factors*

Analogously to (4), the left hand side of (5) is spelled out as

$$-H(\tilde{Y}_d) = E_{\tilde{Y}_d}(\log p(\tilde{y}_d)) = E_{\tilde{Y}_d}(\log E_{\vartheta_{post}}(p(\tilde{y}_d|\theta))).$$

The posterior log-Bayes factor as ratio of log- (posterior mean likelihoods) $\log E_{\vartheta_{post}}(p_i(y_d|\theta))$, $i = 1, 2$, may be interpreted as a comparison of estimates of $-H(\tilde{Y}_d)$ corresponding to two different models. See also (LAUD and IBRAHIM, 1995). In contrast to a cross-validatory estimate the problem of using the data twice is severe here. \square

(4) *Model adequacy*

Estimation of

$$H(\tilde{Y}_d|\vartheta_{post}) = -E_{\vartheta_{post}}E_{\tilde{Y}_d|\theta}(\log p(\tilde{y}_d|\theta))$$

by $-E_{\vartheta_{post}}(\log p(y_d|\theta))$ yields half of the posterior expected deviance, $\bar{D}/2$, a measure of model adequacy introduced by DEMPSTER (1974). Compare the discussion in (SPIEGELHALTER et al., 2002, p.601). As a criterion for model choice \bar{D} has been regarded as ‘not penalizing enough for model complexity’. Choosing instead $H(\tilde{Y}_d)$ as a more appropriate criterion according to (5) the estimate $\bar{D}/2$ is to be penalized by (an estimate of) $I(\tilde{Y}_d, \vartheta_{post})$. \square

(5) *Model complexity*

$I(Y_d, \vartheta) = D(p(\theta|y_d)||p(\theta))$, the directed divergence between posterior and prior density has been interpreted by DEGROOT (1962) as a term of model complexity. For Gaussian distributions it takes the form of the difference between posterior and prior variance. $I(\tilde{Y}_d, \vartheta_{post})$ is the predictive version of that quantity. \square

The symmetrized mutual information $J(\tilde{Y}_d, \vartheta_{post})$ was identified in (VAN DER LINDE, 2004) as the measure of model complexity in DIC. $J(\tilde{Y}_d, \vartheta_{post})$ is nicely interpretable and seems to be preferable to $I(\tilde{Y}_d, \vartheta_{post})$.

4.2 DIC

DIC has been derived as an approximation to $-2E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\bar{\theta}))$, where $\bar{\theta} = E(\vartheta_{post})$. With $\bar{D} = -2E_{\vartheta_{post}}(\log p(y_d|\theta))$, $D(\bar{\theta}) = -2\log p(y_d|\bar{\theta})$ and $p_D = \bar{D} - D(\bar{\theta})$,

$$DIC = D(\bar{\theta}) + 2p_D = \bar{D} + p_D$$

is an empirical version of

$$-2E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\bar{\theta})) = -2E_{\vartheta_{post}}E_{\tilde{Y}_d|\theta}(\log p(\tilde{y}_d|\theta)) + 2E_{\vartheta_{post}}(D(p(\tilde{y}_d|\theta)||p(\tilde{y}_d|\bar{\theta}))). \quad (6)$$

Thus $DIC = \bar{D} + p_D$ follows a structure like that in (5) with \bar{D} being model adequacy and p_D estimating model complexity quantified as $2E_{\vartheta_{post}}(D(p(\tilde{y}_d|\theta)||p(\tilde{y}_d|\bar{\theta})))$. In exponential families p_D estimates

$$2E_{\vartheta_{post}}(D(p(\tilde{y}_d|\theta)||p(\tilde{y}_d|\bar{\theta}))) \approx J(\tilde{Y}_d, \vartheta_{post}), \quad (7)$$

the symmetrized mutual information between future observations \tilde{Y}_d and (posterior) parameters ϑ_{post} . In general $J(\tilde{Y}_d, \vartheta_{post}) \neq 2I(\tilde{Y}_d, \vartheta_{post})$. In exponential families $2I(\tilde{Y}_d, \vartheta_{post}) \leq 2E_{\vartheta_{post}} D(p(\tilde{y}_d|\theta)||p(\tilde{y}_d|\bar{\theta})) \approx J(\tilde{Y}_d, \vartheta_{post})$ is to be expected because

$$\begin{aligned} & I(\tilde{Y}_d, \vartheta_{post}) - E_{\vartheta_{post}}(D(p(\tilde{y}_d|\theta)||p(\tilde{y}_d|\bar{\theta}))) \\ &= E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\bar{\theta})) - E_{\tilde{Y}_d}(\log p(\tilde{y}_d)) \\ &= -D(p(\tilde{y}_d)||p(\tilde{y}_d|\bar{\theta})) \leq 0, \end{aligned}$$

and hence $J(\tilde{Y}_d, \vartheta_{post})$ penalizes more strongly than $2I(\tilde{Y}_d, \vartheta_{post})$, the complexity term in (5) corresponding to the posterior predictive entropy.

Keeping $2H(\tilde{Y}_d|\vartheta_{post})$ and $J(\tilde{Y}_d, \vartheta_{post})$ as measures of model adequacy and model complexity the target that DIC approximates can be expressed independently of any estimate. Without referring to $\bar{\theta}$ we have instead of (6)

$$-E_{\vartheta_{post}}[E_{\tilde{Y}_d|\theta}(\log p(\tilde{y}_d|\theta)) + E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\theta))] = 2H(\tilde{Y}_d|\vartheta_{post}) + J(\tilde{Y}_d, \vartheta_{post}). \quad (8)$$

(7) induces $-2E_{\tilde{Y}_d} \log p(\tilde{y}_d|\bar{\theta})$, the target DIC was derived from by SPIEGELHALTER et al. (2002), as an approximation (in exponential families) to the left hand side in (8), but in general (8) is equivalent to

$$-2E_{\vartheta_{post}} E_{\tilde{Y}_d} \log p(\tilde{y}_d|\theta) = 2H(\tilde{Y}_d|\vartheta_{post}) + 2J(\tilde{Y}_d, \vartheta_{post}). \quad (9)$$

Estimating the right hand side by $\bar{D} + 2p_D = D(\bar{\theta}) + 3p_D$ (in exponential families) yields a criterion that penalizes ‘fit’ ($D(\bar{\theta})$) more strongly. Thus the penalty for model complexity used in DIC is intermediate between that corresponding to the posterior predictive entropy ((5) multiplied by 2) and that of (9). Modifications of DIC could be estimates of the right hand side in (8) or (9) where reference to $\bar{\theta}$ in order to approximate $J(\tilde{Y}_d, \vartheta_{post})$ is appropriate in exponential families (due to bi-affinity of the log-odds ratio function). Whether any estimate works similarly for different sampling distributions, particularly mixtures, is to be investigated. A sampling estimate directly of the left hand side of (8) or (9) might also be generated.

AIC originally was derived from the target $E_{Y_d^t}(D(p_t(\tilde{y}_d)||p(\tilde{y}_d|\hat{\theta}_{ML}(y_d))))$ (to be minimized) assuming a ‘true’ density $p_t(y_d)$ corresponding to the random vector Y_d^t . This is reduced to $-E_{Y_d^t} E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\hat{\theta}_{ML}(y_d)))$, and eventually in estimation the ‘true’ distribution (under the ‘good model assumption’) is replaced by the model specific distribution. Substituting in the targets yields $-E_{Y_d} E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\hat{\theta}_{ML}(y_d)))$. In DIC the maximum likelihood estimate is replaced by $\bar{\theta}$, the expectation over y_d eliminated, and the model specific distribution again is the posterior predictive distribution resulting in $-E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\bar{\theta}))$. The variability of parameters is thus partly ignored. Taking into full account the variability of θ , not as variability of estimates but as induced by the posterior distribution leads to $-E_{\vartheta_{post}} E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\theta))$, essentially the left hand side of (9). The comparison of penalty terms supports preference of DIC and its modifications to the posterior predictive entropy as criterion for model comparison across models.

Example (continued)

We resume the example introduced in section 3.1 and illustrate comparisons across models under conditional sampling. The quantities of interest are chosen as $nH(\tilde{Y}|\tilde{X}_{(s)})$ for (approximate) analytical evaluation and as $-2E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\bar{\theta}))$ for sampling evaluation with DIC. The prior distributions are slightly informative Normal-Gamma distributions, that is,

$$\begin{aligned} Y|x_{(s)}, \beta_{(s)}, \sigma^2 &\sim N(x_{(s)}^T \beta_{(s)}, \sigma^2) \quad \text{independently,} \\ \beta_{(s)} &\sim N(\beta_{0,(s)}, n_0 I_{(s)}), \quad n_0 = 2, \\ \sigma^2 &\sim IG(a, b), \quad a = 0.002, \quad b = 0.012. \end{aligned}$$

First three models are considered corresponding to the regression model with all five covariates yielding the same likelihood but priors focussing on different means for $\beta = \beta_{(1,2,3,4,5)}$.

1. model: $\beta_0 = (0, 0, 0, 1, 1)^T$ supporting the true structure,
2. model: $\beta_0 = (1, 1, 1, 1, 1)^T$ indifferent,
3. model: $\beta_0 = (1, 1, 1, 0, 0)^T$ misleading.

The results are summarized in table 2.

Table 2
comparison of models 1-3 based on $nH(\tilde{Y}|\tilde{X}_{(s)})$ and DIC

	$nH(\tilde{Y} \tilde{X}_{(s)})$	DIC	p_D
1. model	97.56	277.42	5.04
2. model	97.62	278.32	5.02
3. model	97.65	279.14	5.0

Both criteria correctly suggest model 1 to be preferred although the differences of the conditional posterior predictive entropies are not at all pronounced.

Secondly, forward selection based on DIC for models with the same parameters as above but $\beta_{0,(s)} = 0_{(s)}$ was evaluated. The following sequence was obtained:

$X_{(s)}$	$\{X_5\}$	$\rightarrow \{X_5, X_4\}$	$\rightarrow \{X_5, X_4, X_2\}$	$\rightarrow \{X_5, X_4, X_2, X_1\}$	$\rightarrow X$
DIC	285.3	275.1	275.8	276.7	278.2
p_D	1.78	2.68	3.44	4.27	5.02.

Again DIC picks up the correct model with X_4, X_5 but the value for the model including additionally X_2 is very close, illustrating that DIC may not always penalize enough for model complexity. $DIC + p_D$ emphasizes the difference more strongly taking the values 277.8 for $\{X_5, X_4\}$ and 279.22 for $\{X_5, X_4, X_2\}$. \triangle

5 Discussion

An information theoretic review of some criteria for posterior predictive model assessment has been sketched. We discussed the mutual information $I(\tilde{Y}, \tilde{X}|\vartheta_{post})$,

$I(\tilde{Y}, \tilde{X})$ and the conditional entropy $H(\tilde{Y}|\tilde{X})$ as criteria for variable selection within an encompassing model under joint sampling. For variable selection posed as a problem of model comparison across models we focused on the posterior predictive entropy $H(\tilde{Y}|\tilde{X})$ respectively $H(\tilde{Y}_d)$ and DIC (with modifications). Summarizing several points are to be emphasized.

5.1 Choice of the criterion

$I(\tilde{Y}, \tilde{X}|\vartheta_{post})$ is a conditional coefficient of determination. Like $H(\tilde{Y}_d|\vartheta_{post})$ across models, this criterion assesses model adequacy.

$I(\tilde{X}, \tilde{Y})$ and $H(\tilde{Y}|\tilde{X})$ directly quantify the strength of dependence between covariates and response and are intuitively appealing quantities of interest for the comparison of (regression) models. Within an encompassing model under joint sampling $I(\tilde{X}, \tilde{Y})$ and $H(\tilde{Y}|\tilde{X})$ are monotone with the number of included covariates and thus may be regarded as unconditional posterior coefficients of determination. Across models $I(\tilde{X}, \tilde{Y})$ and $H(\tilde{Y}|\tilde{X})$ respectively $H(\tilde{Y}_d)$ do induce a penalty for model complexity. However, frequently this penalty is not sufficient.

In comparisons across models the posterior predictive entropy $H(\tilde{Y}_d)$ rather than being an expected utility represents the information theoretic idea of uncertainty about future observations. DIC, with the slightly different target of maximizing the *marginally* posterior expected likelihood of replicate vectors penalizes more strongly for model complexity than $H(\tilde{Y}_d)$ but partly neglects the variability of the parameter in the posterior distribution. It eventually points to the criterion $-2E_{\vartheta_{post}}E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\theta))$ which again is not an expected utility. This is worth exploring as it has been observed that AIC, a special case of DIC under a vague prior, tends to underpenalize fit.

5.2 Invariance

The criterion $-2E_{\vartheta_{post}}E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\theta))$ is invariant under re-parameterizations, unlike DIC. Transforming \tilde{Y} to $\tilde{Z} = h(\tilde{Y})$, say, yields a model dependent shift,

$$-2E_{\vartheta_{post}}E_{\tilde{Z}_d}(\log p(\tilde{z}_d|\theta)) = -2E_{\vartheta_{post}}E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\theta)) - 2E_{\tilde{Y}_d}(\log p(h'(\tilde{y}_d))).$$

In the original relative comparison according to $E_{\vartheta_{post}}E_{\tilde{Y}_d}(D(p_t(\tilde{y}_d)||p(\tilde{y}_d|\theta)))$ analogous to AIC the true shift cancels. Hence formally the occurrence of lack of invariance is due to giving up the reference model and has to be fixed by definition. Either one could claim under the good model assumption that all model dependent shifts are sufficiently similar to be neglected or one has to fix the representation of observations as \tilde{Y} or \tilde{Z} .

5.3 Approximation and estimation

Given an adequate criterion, a proposed estimate needs to be validated. If criteria for model choice are not explicitly derived, the choice of the quantity of interest and the quality of an approximation or an estimate are indistinguishable. DIC estimates $-E_{\tilde{Y}_d}(\log p(\tilde{y}_d|\hat{\theta}))$, well in exponential families but otherwise to be validated.

The posterior Bayes factor can be interpreted as a poor estimate of the posterior predictive entropy. The classical coefficient of determination R^2 in regression illustrates all points already made: it is a poor estimate of a criterion ($H(\tilde{Y}_d|\theta)$), that can be adequate for a comparison of submodels within an encompassing model but is inadequate for posterior predictive comparisons across models. Estimates based on likelihoods (to evaluate the integrals) often use the data twice. Numerous attempts have been made to correct for this deficiency (overfitting) introducing ideas of cross-validation. The chain rule or the decomposition of mutual information points to an alternative approach in estimation. The chain rule induces the form ‘adequacy+complexity’ (Occam’s razor) taken by many criteria for model comparison.

References

- BERNARDO, J.M. and J.D. BERMUDEZ (1985), The choice of variables in probabilistic classification, in: J.M. BERNARDO et al. (eds.), *Bayesian Statistics 2*, North-Holland, Amsterdam, 67-82.
- BERNARDO, J.M. and R. RUEDA (2002), Bayesian hypothesis testing: a reference approach, *International Statistical Review* **70**, 351-372.
- COVER, T.M. and J.A. THOMAS (1991), *Information Theory*. Wiley, New York.
- DEGROOT, M. (1962), Uncertainty, information and sequential experiments, *Annals of Statistics* **33**, 404-419.
- DEMPSTER, A.P. (1974), The direct use of likelihood for significance testing, in: O. BARNDORFF-NIELSEN et al. (eds.), *Proceedings of the conference on foundational questions in statistical inference*, University of Aarhus, Aarhus, 335-352.
- GEORGE W.I. and R.E. MCCULLOCH (1993), Variable selection via Gibbs sampling, *Journal of the American Statistical Association* **88**, 881-889.
- KULLBACK, S. (1968), *Information Theory and Statistics*, Dover Publications, Mineola, New York (2nd ed.).
- LAUD, P.W. and J.G. IBRAHIM (1995), Predictive model selection, *Journal of the Royal Statistical Society B* **57**, 247-262.
- VAN DER LINDE, A. (2004), On the association between a random parameter and an observable, *Test* **13**, 85-111.
- VAN DER LINDE, A. and G. TUTZ (2004), On association in regression: the coefficient of determination revisited, submitted.
- SPIEGELHALTER, D.J., N.G. BEST, B.P. CARLIN, and A. VAN DER LINDE (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B* **64**, 583-639 (with discussion).