*Title Page

A Prior Predictive Loss Function for the Evaluation of Inequality Constrained Hypotheses

Rens van de Schoot *†, Herbert Hoijtink
Department of Methods and Statistics, Utrecht University

Daniel Brugman
Department of Developmental Psychology, Utrecht University,

Jan-Willem Romeijn
Department of Philosophy, Groningen University

July 8, 2011

Abstract

In many types of statistical modeling inequality constraints are imposed between the parameters of interest. As we will show in this paper, the DIC (i.e., posterior Deviance Information Criterium as proposed as a Bayesian model selection tool by Spiegelhalter et al., 2002) fails when comparing inequality constrained hypotheses. In this paper we will derive the prior DIC and show that it also fails when comparing inequality constrained hypotheses. However, it will be shown that a modification of the prior predictive loss function that is minimized by the prior DIC renders a criterion that does have the properties needed in order to be able to compare inequality constrained hypotheses. This new criterion will be called the Prior Information Criterion (PIC) and will be illustrated and evaluated using simulated data and examples. The PIC has a close connection with the marginal likelihood in combination with the encompassing prior approach and both methods will be compared. All in all, the main message of the current paper is: (1) do not use the classical DIC when evaluating inequality constrained hypotheses, better use the PIC; and (2) the PIC is considered a proper

*Correspondence should be addressed to Rens van de Schoot: Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508TC, Utrecht, The Netherlands; Tel.: +31 302534468; Fax: +31 2535797; E-mail address: a.g.j.vandeschoot@uu.nl.

†This research is financed by The Netherlands Organization for Scientific Research (NWO-VICI-453-05-002).

model selection tool in the context of evaluating inequality constrained hypotheses.

keywords: Bayesian Model Selection, Inequality Constrained Hypothesis, Deviance Information Criterion, DIC.

*Manuscript

A Prior Predictive Loss Function for the Evaluation of Inequality Constrained Hypotheses

July 8, 2011

Abstract

In many types of statistical modeling inequality constraints are imposed between the parameters of interest. As we will show in this paper, the DIC (i.e., posterior Deviance Information Criterium as proposed as a Bayesian model selection tool by Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) fails when comparing inequality constrained hypotheses. In this paper we will derive the prior DIC and show that it also fails when comparing inequality constrained hypotheses. However, it will be shown that a modification of the prior predictive loss function that is minimized by the prior DIC renders a criterion that does have the properties needed in order to be able to compare inequality constrained hypotheses. This new criterion will be called the Prior Information Criterion (PIC) and will be illustrated and evaluated using simulated data and examples. The PIC has a close connection with the marginal likelihood in combination with the encompassing prior approach and both methods will be compared. All in all, the main message of the current paper is: (1) do not use the classical DIC when evaluating inequality constrained hypotheses, better use the PIC; and (2) the PIC is considered a proper model selection tool in the context of evaluating inequality constrained hypotheses.

keywords: Bayesian Model Selection, Inequality Constrained Hypothesis, Deviance Information Criterion, DIC.

1 Introduction

In many types of statistical modeling inequality constraints are imposed between the parameters of interest (Barlow, Bartholomew, Bremner, & Brunk, 1972; Hoijtink, Klugkist, & Boelen, 2008; Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2004; Van de Schoot, Hoijtink, & Deković, 2010). For an overview of literature about inequality constrained hypotheses see

Van de Schoot, Romeijn, and Hoijtink (2011). More specifically, the current paper considers model parameters such as means or regression coefficients that can be constrained to being greater or smaller than either a fixed value or other means or regression coefficients. Phrases like "The mean outcome in both experimental groups is expected to be larger than in the control group" and "women score higher than men in each condition" can be found in many applied papers. These specific expectations may be derived from theories, or empirical evidence, and are translated into statistical hypotheses formulated with inequality constraints. For applications, see, for example Kammers, Mulder, De Vignemont, and Dijkerman (2009); Meeus, Van de Schoot, Keijsers, Schwartz, and Branje (2010); Van de Schoot and Wong (2010); Van Well, Kolk, and Klugkist (2009). Evaluating such inequality constrained hypotheses can be done using model selection procedures. For an overview of literature about inequality constrained hypotheses see Van de Schoot et al. (2011). There is a variety of such model selection tools commonly used in practical applications, most notably Akaike's Information Criterium (AIC; Akaike, 1973), the Bayesian Information Criterium (BIC; Schwarz, 1978), minimal description length (MDL, see, for example Grnwald, Myung, & Pitt, 2005), Bayes factors (BF; see, e.g., Kass & Raftery, 1995) and the recently developed Deviance Information Criterium (DIC; Spiegelhalter et al., 2002).

However, all these tools are not equipped to properly deal with inequality constrained hypotheses. Klugkist, Laudy, and Hoijtink (2005) showed that the Bayes factor can only be used in combination with an encompassing prior approach (see also, Mulder, Hoijtink, & Klugkist, 2009). Both the AIC and BIC fail when evaluating inequality constrained hypotheses because these criteria are not equipped to deal with inequality constraints between the parameters of a model. Alternatives are the order restricted information criterion (ORIC; Anraku, 1999; Kuiper & Hoijtink, 2010) which is limited to analysis of variance, and the prior-adapted-BIC (Romeijn, Van de Schoot, & Hoijtink, 2011), respectively. The MDL in relation to a reduction of the parameter space is discussed in Balasubramanian (2005). The DIC is up till now not discussed in relation to its behavior in the context of evaluating inequality constraints and this is exactly what we do in the current paper.

The DIC has an important role in statistical model comparison, see for example its availability in software like WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), MlwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) or Mplus (Muthen & Muthen, 2010). However, as we will show, the DIC fails when evaluating inequality constraint hypotheses. The plan of this paper is as follow. After introducing some examples in Section 2, we introduce in Section 3 the original DIC and we show that, it can not be used to choose between a set of inequality constrained hypotheses. In Section 4 we provide an alternative for the classical DIC, namely the prior Deviance Information Criterium (*prior* DIC). Unfortunately, also the *prior* DIC does not work well in the context of inequality constrained hypotheses. To accommodate for this, we propose a new loss function in Section 5, which is minimized by the Prior Information Criterion (PIC). The PIC can be used to evaluate a set of inequality constrained hypothesis. We evaluate its performance, see Section 6, and we show that it is connected to the marginal likelihood and thus to the Bayes factor approach of, for example, Klugkist et al. (2005).

2 Examples

In this section we provide three different situations where inequality constrained hypotheses can be of interest and we describe two real-life examples where the hypotheses of interest are specified using inequality constraints. We will use Example 1 as a case study throughout the paper to investigate the performance of the *posterior* DIC, *prior* DIC and the PIC. In Section 6 we briefly reconsider all other examples. Note that the scope of our proposed method is limited to the multivariate normal linear model.

2.1 Example 1

First, consider an example of a univariate model with where persons from two groups receive a score on one dependent variable, y_i $(i = 1, \ldots, N)$:

$$y_i = \mu_1 d_{i1} + \mu_2 d_{i2} + \epsilon_i \ , \tag{2.1}$$

where μ_1 and μ_2 denote the mean score on y for group 1 and 2 respectively and where the residuals ϵ_i are assumed to be normally distributed $N(0, \sigma^2)$. The group membership of a person is denoted by $d_{ig} \in 0, 1$, where 1 and 0 denote that a person is either a member or not a member of group g. Suppose we want to evaluate two hypotheses: $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_1 < \mu_2$.

This example has its counterparts in applied papers, see, for example, Van Well et al. (2009) about the relationship between sex, gender role identification, and the gender relevance of a stressor. The authors examined mean scores for eight groups on the dependent variable stress responses, to investigate sex and gender (mis)match effects. They formulated several hypotheses by imposing inequality constraints upon group means (i.e., one or more group means are expected to be larger or smaller than one or more other group means).

2.2 Example 2

Next, consider a second example of a multivariate model with two dependent variables (denoted by y_{1i} and y_{2i} for $i = 1, \ldots, N$),

$$\begin{aligned} y_{1i} &= \mu_1 + \epsilon_{i1} \\ y_{2i} &= \mu_2 + \epsilon_{i2} \ , \end{aligned} \tag{2.2}$$

where the residuals are assumed to be normally distributed

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \sim N\Big(0, \Sigma\Big) \ , \Sigma = \begin{bmatrix} \sigma^2_{y_1} & \rho\sigma_{y_1}\sigma_{y_2} \\ \rho\sigma_{y_1}\sigma_{y_2} & \sigma^2_{y_2} \end{bmatrix} . \tag{2.3}$$

Suppose we want to evaluate two hypotheses: $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_1 > 0;\ \mu_2 > 0$.

Also this multivariate model has its counterparts in applied papers, see, for example Kammers et al. (2009) about the number of body representations in the brain. The authors examined the main problems that are encountered when trying to dissociate multiple body representations in healthy individuals with the use of bodily illusions. Several models were specified within a multivariate normal model using (in)equality constraints between five repeated measurements.

2.3 Example 3

Finally, consider an example of a non-linear regression model with one dependent variable, y_i ($i = 1, \ldots, N$) with a linear, i.e. X_i, and a non-linear predictor, i.e. X_i^2:

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \ , \tag{2.4}$$

where β_0 denote the intercept and where the residuals ϵ_i are assumed to be normally distributed with $N(0, \sigma^2)$.

Suppose we want to evaluate two hypotheses: $H_0 : \beta_2$ and $H_1 : \beta_2 > 0$. Such expectations are of interest in, for example, the research of retention memory (see, e.g., Myung, 2003). See Section 2.5 for a real life example.

2.4 Real-life Example 1: Moral Judgment Competence

Leenders and Brugman (2005) investigated whether moral judgment competence and attitude towards delinquent behavior create a domain shift in young adolescents. That is, a certain behavior which in society as a whole is considered to be not moral (e.g. aggression, violence), might be a group convention in certain adolescent groups. In total 135 pupils of intermediate secondary schools in the Netherlands were asked to report whether the respondent had committed such behaviour (never, once, more than once). They were also asked to judge aggressive acts and vandalistic acts in hypothetical situations on how moral they thought the behavior was. For each hypothetical situation, questions were asked (on a 4-point scale) about the acceptability (Is it wrong or right to do such a thing?'), the seriousness (How bad is it to do such a thing?), the generalizability (If everybody were doing such things, would they then be wrong or right?) and the rule/authority contingency (If nobody saw it, would it then be wrong or right?) of the transgression. Just like in the original article, for each category the sum scores were computed in a way that a high criterion score indicated a more non-moral (conventional/personal) judgment. The researchers had specific ideas about differences in the level of morality in these hypothetical situations between pupils that did or did not report to conduct aggressive acts themselves.

The model under investigation is given by

$$\begin{aligned} y_{1i} &= \mu_{11} d_{ig1} + \mu_{12} d_{ig2} + \epsilon_{1i} \\ y_{2i} &= \mu_{21} d_{ig1} + \mu_{22} d_{ig2} + \epsilon_{2i}\ , \end{aligned} \tag{2.5}$$

where $\mu_{1\cdot}$ and $\mu_{2\cdot}$ denote the mean score on the hypothetical construct vandalism (denoted by y_1) and the hypothetical construct aggression (denoted by y_2) and where $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ denote the mean for the group reported not to conduct aggressive acts and the group that did report to conduct aggressive acts, respectively. Again, group membership of a person is denoted by $d_{ig} \in 0, 1$ and the residuals are assumed to be normally distributed with

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \sim N\Big(0, \Sigma\Big)\ , \Sigma = \begin{bmatrix} \sigma^2_{y_1} & \rho\sigma_{y_1}\sigma_{y_2} \\ \rho\sigma_{y_1}\sigma_{y_2} & \sigma^2_{y_2} \end{bmatrix}\ . \tag{2.6}$$

Note that this example is a combination of (2.2) and (2.4).

There are three hypotheses of interest:

$$\begin{aligned} &H_0 : \mu_{12}, \mu_{11} \text{ and } \mu_{22}, \mu_{21} \\ &H_1 : \mu_{12} > \mu_{11} \text{ and } \mu_{22} > \mu_{21} \\ &H_2 : \mu_{12} = \mu_{11} \text{ and } \mu_{22} > \mu_{21}\ . \end{aligned} \tag{2.7}$$

The first hypothesis, is an unconstrained hypothesis (H_0). A second hypothesis, H_1, postulates that the aggressive group ($\mu_{\cdot 2}$) also judge the same behavior in all hypothetical situations to be more conventional and as such morally more appropriate than their peers who do not report such behavior ($\mu_{\cdot 1}$). The third hypothesis, H_2, is that there is a domain shift in the judgement about hypothetical situations. That is, for pupils that reported to have conducted some delinquent behavior (i.e. aggression), in the same hypothetical situation, they will judge it to be more morally accepted compared to adolescents that did not report to conduct the same behavior. However, in hypothetical situations concerning other delinquent behavior that was not reported by these same adolescents (i.e. vandalism), they will judge the hypothetical situation to be equally morally condemnable as adolescents that did not report any antisocial behavior. In Section 6.3 the data of Leenders and Brugman (2005) will be used to re-evaluate these hypotheses.

2.5 Real-life Example 2: Ph.D. delays

Sonneveld, Yerkes, and Van de Schoot (2009) report on Ph.D. trajectories and employment outcomes of recent Dutch Ph.D. recipients at four universities in the Netherlands. The report provides detailed information on the background of Ph.D. candidates, their Ph.D. trajectory, including supervision and the performance of Ph.D. candidates, as well as their initial employment after obtaining their Ph.D.

In the Netherlands it is possible to differentiate between three different types of Ph.D. status, including: (a) a Ph.D. candidate that is employed by the university, (b) scholarship recipients and (c) external and/or dual Ph.D. candidates. Full employment contracts for Ph.D. candidates are the exception and not the rule throughout Europe. Only the Netherlands, Finland and Turkey have doctoral educational structures in which different types of Ph.D. status exist simultaneously. The majority of respondents surveyed (71.1%) reported that their main formal status was 'employee'. In the current paper we will only focus on employees (n = 304).

Among many other questions, the researchers asked the Ph.D. recipients how long it took them to finish their Ph.D thesis. It appeared that Ph.D. recipients took an average of 59.8 months (five years and four months) to complete their Ph.D. trajectory. In the current paper we will answer the question why some Ph.D. recipients took longer than other by using age as a predictor (M = 30.7, SD = 4.48, min-max = 26-69). The relation between completion time and age is expected to be non-linear. This might be due to the fact that at a certain point in your life (i.e., mid-thirties), family life takes up more of your time than when you are in your twenties or when you are older.

However, we expect that if you are in your mid-thirties and you are doing a Ph.D. you also take this extra time into account. The researchers asked to Ph.D. candidates about their planned graduation day according to the original contract and their actual graduation day. The average gap between the two data appeared to be 9.6 months (SD = 14.4, min-max = -3 − 69). We expect that the lag between planned and actual time spent on the trajectory is less prone to non-linearity compared to actual project time.

If y_{1i} denotes actual project time and y_{2i} denotes the lag between actual and planned project time, the model under investigation is given by

$$\begin{aligned} y_{1i} &= \beta_{01} + \beta_1 Age_i + \beta_2 Age_i^2 + \epsilon_{1i} \ , \\ y_{2i} &= \beta_{02} + \beta_3 Age_i + \beta_4 Age_i^2 + \epsilon_{2i} \ . \end{aligned} \tag{2.8}$$

To avoid multicollinearity Age will be centered. The residuals are assumed to be normally distributed

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \sim N\Big(0, \Sigma\Big) \ , \Sigma = \begin{bmatrix} \sigma_{y_1}^2 & \rho\sigma_{y_1}\sigma_{y_2} \\ \rho\sigma_{y_1}\sigma_{y_2} & \sigma_{y_2}^2 \end{bmatrix} \ . \tag{2.9}$$

The following hypotheses are of interest:

$$\begin{aligned} &H_0 : \beta_2, \beta_4 \\ &H_1 : \beta_2 > 0 \text{ and } \beta_4 > 0 \\ &H_2 : \beta_2 > \beta_4 > 0 \ . \end{aligned} \tag{2.10}$$

In Section 6.4 the data of Sonneveld et al. (2009) will be used to evaluate these hypotheses.

3 Posterior DIC

One way of evaluating hypotheses, is to use a model selection approach. This is not a test of the model in the sense of hypothesis testing, rather it is an evaluation between models using a trade-off of model fit and model complexity. The likelihood of an hypothesis is a measure of model fit, and the number of parameters involved in the hypothesis is a measure of complexity. The greater the number of parameters, the larger the compensation for model complexity becomes. So, adding a parameter should be accompanied by an increase in model fit to accommodate for the increase in complexity. Several competing statistical models may be ranked according to their value on the model selection tool used and the one with the best trade-off is the winner of the model selection competition.

The Deviance Information Criterion (DIC), is proposed in Spiegelhalter et al. (2002) as a Bayesian criterion for minimizing the posterior predictive loss. In this section we briefly introduce the DIC, thereafter we show with our running example that the DIC fails when comparing inequality constrained hypotheses.

Note that from now on, we will use the *posterior* DIC whenever we refer to the DIC of Spiegelhalter et al. (2002) and we will use *prior* DIC whenever we refer to our adjustment of the DIC, that will be introduced in Section 4.

3.1 Definition

The *posterior* DIC minimizes the posterior expectation of the expected loss (Gelman, Carlin, Stern, & Rubin, 2004). It is defined as the error that is expected when a statistical model estimated by the observed data set $\mathbf{y}$ is applied to a future data set $\mathbf{x}$. Let $f(\cdot)$ denote the likelihood, then the expected loss is given by

$$\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta}^*)}[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)] \ , \tag{3.1}$$

where $-2\log f(\cdot)$ is the loss function of a future data set $\mathbf{x}$ in which $\bar{\boldsymbol{\theta}}_y$ is the expected a-posteriori estimate of the model parameters $\boldsymbol{\theta}$ based on the observed data set $\mathbf{y}$. If we would know the true parameter value $\boldsymbol{\theta}^*$, the expectation in (3.1) could be computed. However, since these are unknown, the *posterior* DIC takes the posterior expectation of (3.1). Let $\mathrm{E}_{g(\theta|\mathbf{y})}$ denotes the expectation with respect to the posterior distribution $g(\boldsymbol{\theta} \mid \mathbf{y})$, then

$$\begin{gathered}\mathrm{E}_{g(\boldsymbol{\theta}|\mathbf{y})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} \approx \\ -2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + 2\big[-2\overline{\log f(\mathbf{y} \mid \boldsymbol{\theta})} + 2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)\big] \ , \end{gathered} \tag{3.2}$$

where (3.2) is the definition of the *posterior* DIC. The term $-2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)$ in (3.2) is often interpreted as model (mis)fit and the term $\big[-2\overline{\log f(\mathbf{y} \mid \boldsymbol{\theta})} + 2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)\big]$ in (3.2) is often interpreted as the effective number of parameters and is considered a penalty term.

3.2 Estimation

The *posterior* DIC can be computed using Monte Carlo simulation and is available in several software packages, for example, WinBUGS (Lunn et al., 2000), MLwiN (Rasbash et al., 2009) and Mplus (Muthen & Muthen, 2010).

Let $\boldsymbol{\theta}^1 \ldots \boldsymbol{\theta}^L$ be L draws from the posterior distribution $g(\boldsymbol{\theta} \mid \mathbf{y})$, then $-2\overline{\log f(\mathbf{y} \mid \boldsymbol{\theta})}$ can be estimated by

$$\sum_{l=1}^{L} \frac{-2 \log f(\mathbf{y} \mid \boldsymbol{\theta}^l)}{L} , \tag{3.3}$$

and $-2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)$ can be estimated by

$$-2 \log f(\mathbf{y} \mid \sum_{l=1}^{L} \frac{\boldsymbol{\theta}_1^l}{L}, \ldots, \sum_{l=1}^{L} \frac{\boldsymbol{\theta}_k^l}{L}) , \tag{3.4}$$

where k is an index for the parameters in $\boldsymbol{\theta}$ ($k = 1, \ldots, K$).

An important issue when computing the *posterior* DIC is the specification of the prior distribution. A default approach is to specify a vague or low-informational prior distribution. In that case, the computation of the *posterior* DIC is independent of the specified prior because the posterior distribution, $g(\boldsymbol{\theta} \mid \mathbf{y})$, is dominated by the data.

3.3 Behavior of the Posterior DIC in Constrained Model Selection

To inspect the behavior of the *posterior* DIC in the context of evaluating inequality constraint hypotheses, we consider Example 1 with $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_1 < \mu_2$. According to Mulder, Hoijtink, and Klugkist (2009), the prior distribution for Example 1 is given by

$$h_0(\mu_1, \mu_2, \sigma^2) = N(\mu_1|\mu_0, \tau_0^2) \times N(\mu_2|\mu_0, \tau_0^2) \times Inv\chi^2(\sigma^2|\upsilon_0, \sigma_0^2), \tag{3.5}$$

where μ_0 is the prior mean and τ_0^2 is the prior variance. Hypothesis H_1 is nested in H_0, therefore $h_1(\cdot)$ is proportional to $h_0(\cdot)$, with

$$h_1(\cdot) : \begin{cases} c^{-1} h_0(\cdot) & \text{if}(\mu_1, \mu_2) \in H_1 \\ 0 & \text{otherwise} \end{cases} , \tag{3.6}$$

where c is a normalization constant given by

$$c = \int_{(\mu_1, \mu_2) \in H_1} h_0(\mu_1, \mu_2) d(\mu_1, \mu_2) . \tag{3.7}$$

Using this encompassing prior approach only the prior distribution for H_0 needs to be specified. Note that the encompassing prior approach has been used in computing Bayes factors, which will not be considered in the current paper, but see for more information Mulder, Hoijtink, and Klugkist (2009).

Now let $g_0(\cdot)$ denote the posterior distribution of the unconstrained hypotheses and $g_1(\cdot)$ the posterior distribution of H_1, then $g_0(\cdot) \propto f(\cdot) \times h_0(\cdot)$ and $g_1(\cdot) \propto f(\cdot) \times h_1(\cdot)$. Then, $g_1(\cdot) = d^{-1} g_o(\cdot)$ where

$$d = \int_{(\mu_1,\mu_2)\in H_1} g_0(\mu_1, \mu_2) \partial(\mu_1, \mu_2) \; . \tag{3.8}$$

For $\mu_2 - \mu_1 \rightarrow \infty$, $g_0(\mu_1, \mu_2, \sigma^2|\mathbf{y})$ - $g_1(\mu_1, \mu_2, \sigma^2|\mathbf{y}) \rightarrow 0$. That is, if the population from which the data are generated is strongly in agreement with H_1, the difference between the posterior distributions for H_0 and H_1 goes to zero. Since the *posterior* DIC is computed using samples of μ_1, μ_2 and σ^2 obtained from the posterior distribution, see Equations (3.3) and (3.4), for $\mu_2 - \mu_1 \rightarrow \infty$, samples obtained under H_0 and H_1 are exchangeable. Consequently, DIC_{H_0} and DIC_{H_1} have the same values. This result is counterintuitive and unwanted because H_1 is more parsimonious than H_0 and hence it contains more information (cf. Sober, 2006), so it should be preferred by the DIC.

A simulation study was performed to illustrate the failure of the *posterior* DIC. You can also derive analytic expressions for the behavior of the posterior probability distribution, on which the behavior of the *posterior* DIC hinges. For more details see Romeijn et al. (2011). Here we discuss this behavior merely for the purpose of illustration. Seven data sets from seven populations were considered. Data were constructed in such a way that the sample means and variance are exactly equal to the population parameters (with $\sigma^2 = 1$ and $n = 20$ for each group). The population means for the seven data sets are displayed on the x-axis in Figure 1. Note that the first four data sets are in agreement with the constraints of H_1, whereas the last three data sets are constructed in such a way that they violate the constraints of H_1. The difference between the seven data sets is that the size of the difference between the two group means varies from small to large. We also considered an equality constrained hypothesis, $H_2 : \mu_1 = \mu_2$, to investigate the performance of the *posterior* DIC. For this hypothesis, μ_1 and μ_2 can be replaced by μ. For each data set we used WinBUGS to compute the *posterior* DIC.

Next, the hypotheses of interest were evaluated for all seven data sets with the *posterior* DIC. The results are presented in Figure 1. When looking at populations 1-5 in Figure 1, it can be seen that the values of the *posterior* DIC for H_0 and H_1 are equal. Hence, the *posterior* DIC can not distinguish

H_0 and H_1. This is counterintuitive because the population values satisfy the constraints of H_1 and H_1 is more parsimonious than H_0. For population 4 and 5, the two data sets with the smallest difference in sample means, the value of the *posterior* DIC for H_2 is lowest. This result is in line with what would be expected because the means are approximately equal. When the population means do not fit the constraints imposed by H_1 (i.e. populations 6 and 7) the values for the *posterior* DIC for H_0, H_1 and H_2 are in line with what would be expected: the lowest value for H_0 followed by H_2 and H_1, respectively. In sum, the *posterior* DIC fails to distinguish between hypotheses H_0 and H_1 when the data are strongly in agreement with the most constrained hypothesis, H_1.

4 Prior DIC

Within the Bayesian framework, there are two perspectives on model selection: a prior predictive approach (e.g. Box, 1980; Kass & Raftery, 1995) and a posterior predictive approach (e.g. Gelman et al., 2004; Gelman, Meng, & Stern, 1996). Spiegelhalter, Best, Carlin, and Van Der Linde (2002) derived the posterior Deviance Information Criterium (*posterior* DIC) to choose between a set of competing hypotheses. As we have seen in the previous section, the *posterior* DIC failed to choose between a set of inequality constrained hypotheses. In this section we will derive the prior Deviance Information Criterium (*prior* DIC).

4.1 Definition

The point of departure for the *prior* DIC is the same as for the *posterior* DIC, namely the expected loss given in (3.1). However, to deal with the unknown parameters $\boldsymbol{\theta}^*$, for the *prior* DIC, we take the expectation of the expected loss with respect to the prior distribution, $h(\boldsymbol{\theta})$, instead of the posterior distribution, $g(\boldsymbol{\theta} \mid \mathbf{y})$, as was the case for the *posterior* DIC:

$$\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} \tag{4.1}$$

The major difference between (3.2) and (4.1) is that $g(\boldsymbol{\theta} \mid \mathbf{y})$ is replaced by $h(\boldsymbol{\theta})$. As will be shown, using $h(\boldsymbol{\theta})$ instead of $g(\boldsymbol{\theta} \mid \mathbf{y})$ is a final step towards an IC that does not suffer from the drawbacks discussed in the previous section.

The main problem now, is to find an expression for $\mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)\big]$ and this is what we do in Appendix A resulting in the definition of the *prior*

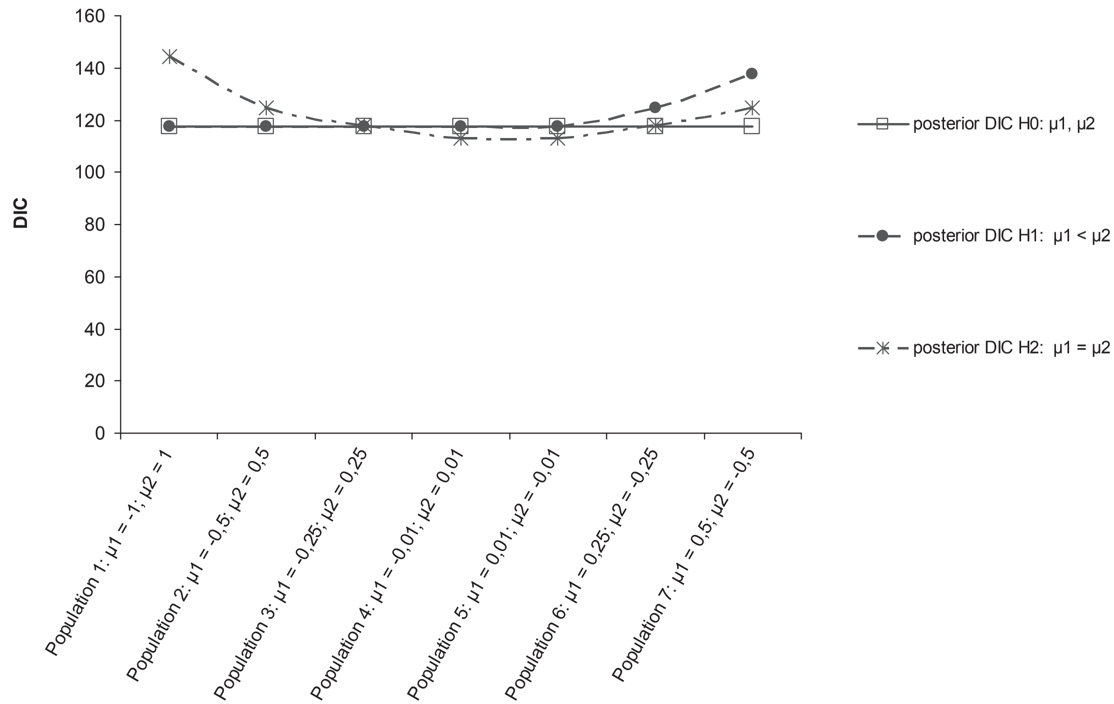

Figure 1: Values of the Posterior DIC for H_0, H_1 and H_2; and for populations 1-7 of Example 1.

DIC:

$$\mathrm{E}_{h(\boldsymbol{\theta})}\left\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\left[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\right]\right\} \approx \\ C + 2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\left[-2\log f(\mathbf{y} \mid \boldsymbol{\theta})\right] , \tag{4.2}$$

where $C = \mathrm{E}_{h(\boldsymbol{\theta})}\left\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\left[-2\log f(\mathbf{x} \mid \boldsymbol{\theta})\right]\right\}$ is constant when comparing inequality constrained hypotheses, see Appendix B, and consequently can be ignored.

Note the two major differences between the *prior* and *posterior* DIC: the first term of (4.2) (i.e. C) does not have a corresponding part in the definition of the *posterior* DIC, see (3.2) and the third term on the right hand side of (4.2) (i.e. $\mathrm{E}_{h(\boldsymbol{\theta})}\left[-2\log f(\mathbf{y} \mid \boldsymbol{\theta})\right]$) is the expectation with respect to the prior distribution whereas the corresponding term in (3.2) is the expectation with respect to the posterior distribution.

4.2 Estimation

The *prior* DIC can be computed using Monte Carlo simulation, for example using R (R Development Core Team, 2006). Let $\boldsymbol{\theta}^1 \ldots \boldsymbol{\theta}^L$ be L draws from the posterior distribution, then $2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)$, in Equation (4.2) can be estimated by

$$2 \log f(\mathbf{y} \mid \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\theta}_1^l, \ldots, \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\theta}_k^l). \tag{4.3}$$

Furthermore, let $\boldsymbol{\theta}^1 \ldots \boldsymbol{\theta}^K$ be K draws from the prior distribution, then $\mathrm{E}_{h(\boldsymbol{\theta})}\left[-2\log f(\mathbf{y} \mid \boldsymbol{\theta})\right]$ in Equation (4.2) can be estimated by

$$\frac{1}{K}\sum_{k=1}^{K}\text{-}2 \log f(\mathbf{y} \mid \boldsymbol{\theta}^k) . \tag{4.4}$$

Just like for the *posterior* DIC, the specification of the prior distribution is of importance. For the *prior* DIC it is even essential that the prior distribution is specified correctly because only then background knowledged in the form of inequality constraints between the parameters of interested can be incorporated. In order to incorporate the constraints in the prior distribution, we use the encompassing prior approach as was discussed before. The prior is given in Equation (3.5) and we assume the same prior distribution for each parameter that is subjected to constraints, $h_0(\mu_1) = h_0(\mu_2)$. Specifying the parameters of the prior distribution in constrained hypotheses selection is further explained in Mulder, Hoijtink, and Klugkist (2009) and Mulder, Klugkist, et al. (2009). The actual computation of the second term of the *prior* DIC for H_0 and H_1 can be done using samples from $g_0(\mu_1, \mu_2, \sigma^2|\mathbf{y})$ and $g_1(\mu_1, \mu_2, \sigma^2|\mathbf{y})$, respectively. These samples can be obtained using the Gibbs sampler for $g_0(\cdot)$ (see, Gelman et al., 2004) and the constrained Gibbs sampler for $g_1(\cdot)$ (see, Klugkist et al., 2005). The third term of the *prior* DIC can be computed using a sample from the prior distribution of the hypotheses under investigation.

4.3 Behavior of The prior DIC in Constrained Model Selection

To show that the *prior* DIC can be used to choose between a set of constrained hypotheses if the population from which the data are generated is fully in agreement with the most constrained hypothesis, whereas the *posterior* DIC fails to do so, we reconsider Example 1.

If we would compare H_0 and H_1 with the *prior* DIC, the first term of the *prior* DIC given in Equation (4.2) is constant (see Appendix B). Now, consider the same situation as in the beginning of Section 3.3 where the population from which the data was generated is strongly in agreement with H_1. In this case, the second term in Equation (4.2) does also not differ between H_0 and H_1, because for $\mu_1 - \mu_2 \to \infty$, $\bar{\mu}_1|H_0 \to \bar{\mu}_1|H_1$ and $\bar{\mu}_2|H_0 \to \bar{\mu}_2|H_1$. So, the third term, $\mathrm{E}_{h(\mu_1,\mu_2,\sigma^2)}\left[\cdot\right]$, should make the difference between H_0 and H_1.

Since samples of μ_1 and μ_2 are taken from the prior distribution $h_0(\mu_1, \mu_2, \sigma^2)$ and since $h_0(\mu_1, \mu_2, \sigma^2) \neq h_1(\mu_1, \mu_2, \sigma^2)$ because of the normalization of the prior distribution according to Equation (3.6), samples from the prior distribution are different for H_0 and H_1. For $\mu_1 - \mu_2 \to \infty$, the third term of (4.2) when computed for H_0 is based on more large values of $-2\log\ f(\mathbf{y} \mid \mu_1, \mu_2, \sigma^2)$ than when it is computed for H_1. Consequently, the third term of (4.2) for H_1 is smaller than the third term of (4.2) for H_0.

Again, a simulation study was performed where data sets from the seven populations of Section 3.3 were considered. The exact specification of the parameters of the prior distribution for population 1 with $\mu_1 = -1$ and $\mu_2 = 1$, are $\mu_0 = 0$, $\tau_0^2 = 0.97$, $\upsilon_0 = 2$ and $\sigma_0^2 = 1.95$.

In contrast to the *posterior* DIC, the *prior* DIC is able to correctly distinguish between H_0 and H_1 when the data are in agreement of the constraints of H_1, see populations 1-3 in Figure 2, where the *prior* DIC is lowest for H_1. For the data with the smallest differences in sample means (population 4 and 5), the *prior* DIC is lowest for H_2. When the constraints are not supported by the data, populations 6-7, the value for H_0 should be the lowest value, but as can be seen in Figure 2, this is not the case! So, when the data are fully in agreement with H_1 the *prior* DIC outperforms the *posterior* DIC, but when the data do not support H_1, the *prior* DIC fails to correctly distinguish the three hypotheses.

What goes wrong? Consider the prior expectation of the expected loss given in (3.1), which is approximated by the *prior* DIC as was shown in Appendix A:

$$\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} \tag{4.5}$$

$$\approx 2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\big[-2\log f(\mathbf{y} \mid \boldsymbol{\theta})\big] \ . \tag{4.6}$$

The loss function in Equation (4.5) captures how well replicated data fit a certain hypothesis, that is, how good $\bar{\boldsymbol{\theta}}_y$ is a summary of $\mathbf{x}$. However, this loss function does not accommodate 'bad' fitting hypotheses, that is, if for a hypothesis $\bar{\boldsymbol{\theta}}_y$ is not a good summary of $\mathbf{y}$, this will not be detected by the loss function in (4.5). Note that it might appear the correction for 'bad' fitting hypotheses is done by the first term of the approximation of the loss function, see Equation (4.6). However, the second term cancels the influence of the first term because the second term can be written as a Taylor expansion around the first term, see Appendix A and Equation (A.7).

Let us return to the loss function in Equation (4.6) and consider the situation of Example 1. Suppose that a population is not in agreement with the inequality constrained hypothesis, $H_1 : \mu_1 < \mu_2$, for example Population 7 with population means $\mu_1 = 0.5; \mu_2 = -0.5$. In this situation the *prior* DIC chooses H_1 as the best hypothesis, see Figure 2. This result is unwanted because the means in the data satisfy $\mu_1 > \mu_2$.

Under the assumption $\mu_1 < \mu_2$ in the data were $\mu_1 = 0.5; \mu_2 = -0.5$, the prior mean that fits these constraints will have a mean of zero because it is set at the boundary of the admissible parameter space. For the computation of (4.5), data are replicated based on $\boldsymbol{\theta}$ from a prior distribution with $\mu_0 = 0$. These replicated data are adequately summarized by $\overline{\mu}_1$ and $\overline{\mu}_2$. However, what is not accounted for in (4.5) is that the observed data $\mathbf{y}$ are not adequately summarized by $\overline{\mu}_1$ and $\overline{\mu}_2$. This leads to situations where the loss function in (4.5) has a preference for 'bad' fitting inequality constrained hypotheses.

In conclusion, neither the *prior* DIC, nor the *posterior* DIC are proper model selection tools for the evaluation of inequality constrained hypotheses. In the next section the prior predictive loss function will be adjusted such that its estimate, the PIC, can be used to select the best of a set of equality and inequality constrained hypotheses.

5 A New Loss Function for the Evaluation of Inequality Constrained Hypotheses

The solution of the aforementioned problem (i.e. that neither the *prior* DIC, nor the *posterior* DIC are proper model selection tools for the evaluation of inequality constrained hypotheses) is to adjust the loss function that is used to select the best hypothesis such that it also accounts for the agreement between $\bar{\boldsymbol{\theta}}_y$ and $\mathbf{y}$. The loss function in (4.5) can be rewritten as

$$-2 \ \mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} + \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) \tag{5.1}$$

$$\approx \mathrm{E}_{h(\boldsymbol{\theta})}\big[-2 \log f(\mathbf{y} \mid \boldsymbol{\theta})\big]. \tag{5.2}$$

The new loss function determines not only how well replicated data fit with a certain hypothesis (the term between accolades in 5.1), but it also determines how well a hypothesis fits the data (the second term between accolades in 5.1). It is approximated by the third term of the *prior* DIC and is our final model selection tool, to be called Prior Information Criterium (PIC) given by (5.2).

5.1 Behavior of the PIC in Constrained Model Selection

In Figure 3 the PIC values for populations 1-7 of Example 1 are shown. As can be seen, the PIC chooses for H_1 as the best hypothesis in situations where this hypothesis is true in the population, see populations 1-3. The PIC chooses for H_2 as the best hypothesis where this hypothesis is strongly supported by the population values, see populations 4 and 5. Finally, the PIC chooses for the unconstrained hypothesis, H_0, where the (in)equality constraints for both H_1 and H_2 are not supported by the data, see populations 6 and 7. These results makes the PIC outperform both the *posterior* and *prior* DIC in all situations.

5.2 Influence of Prior Specification

Since the specification of the prior has an impact on the results, we evaluated the influence of the prior specifications on the PIC. To do so, we performed a simulation study where μ_0, τ_0^2, υ_0 and σ_0 were varied across populations.

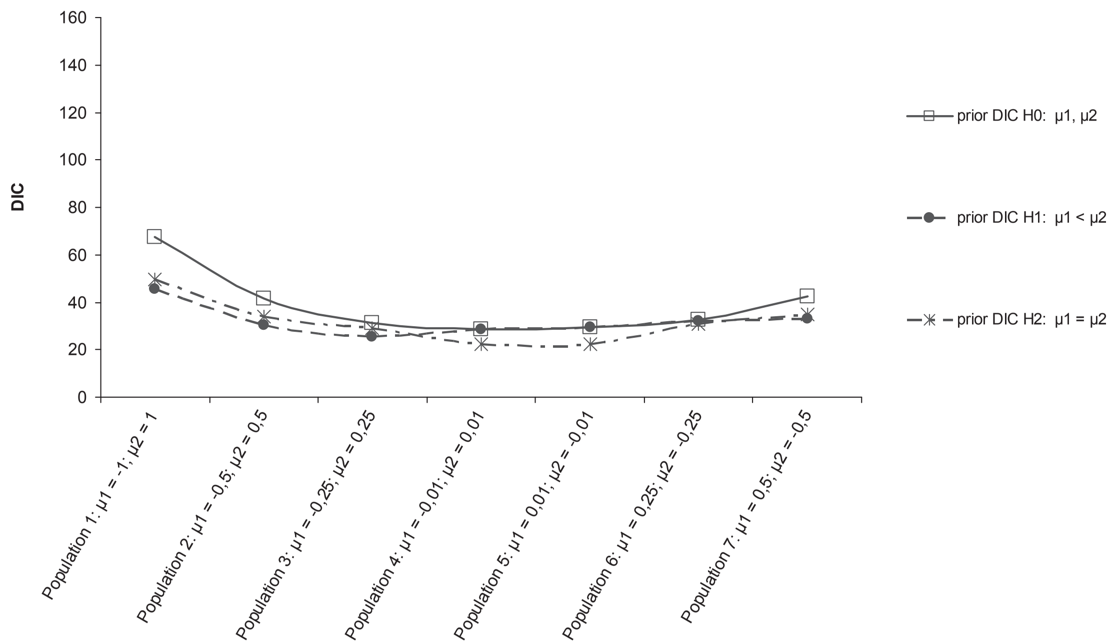

Figure 2: Values of the Prior DIC for H_0, H_1 and H_2; and for populations 1-7 of Example 1.

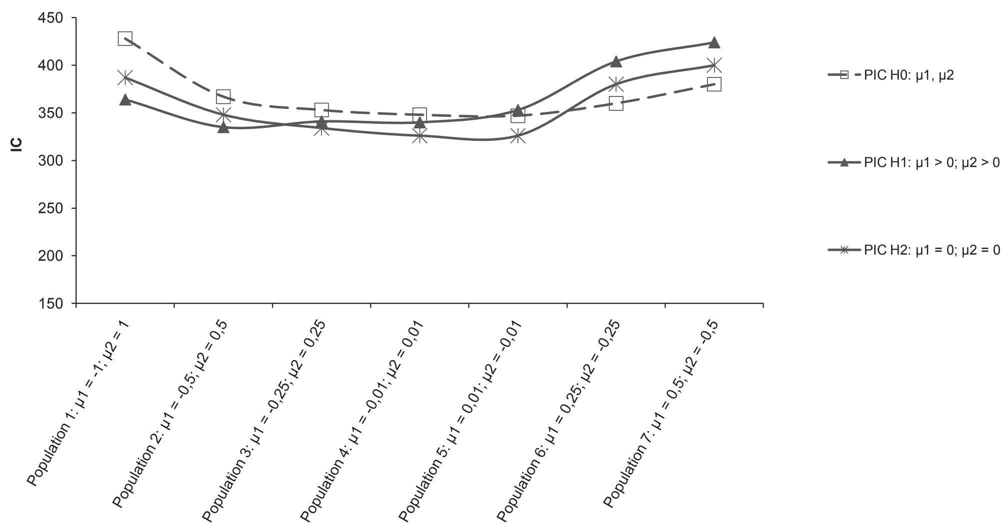

Figure 3: The PIC for populations 1-7 of Example 1.

We evaluated H_0, H_1 and H_2 for populations 1, 4, and 7 with: (1) $\mu_0 - 1$, $\mu_0 + 0$ and $\mu_0 + 1$; (2) $\tau_0 \times .5$, $\tau_0 \times 1$, and $\tau_0 \times 5$; (3) $\upsilon_0 = 2$ and $\upsilon_0 = 5$; (4) $\sigma_0 \times .5$, $\sigma_0 \times 1$, and $\sigma_0 \times 5$.

The results are presented in Table 1 with in bold the correct conclusions. As can be seen, the specification of the prior influences the results. However, as can be seen for different prior specifications the influence is mainly on the height of PIC and not the relative ordering of PIC_{H_0}, PIC_{H_1}, and PIC_{H_2}.

5.3 PIC versus Marginal Likelihood

The PIC is related to the marginal likelihood (ML) which is given by

$$\mathrm{ML} \approx -2 \log \mathrm{E}_{h_t(\boldsymbol{\theta})} \big[f(\mathbf{y} \mid \boldsymbol{\theta}) \big] \tag{5.3}$$

The difference between (5.2) and (5.3) is the position of the log: inside (PIC) or outside (ML) the expectation. If within the constrained model $h_1(\cdot) = c \times h_0(\cdot)$, see Equation (3.7), and under the further assumptions made about encompassing and constrained priors made in this paper then the relation between the PIC and ML shows a monotone relation. To exemplify this relation, we performed a small simulation study. In Figure 4 $\mathrm{PIC}_1 - \mathrm{PIC}_2$ and $\mathrm{ML}_1 - \mathrm{ML}_2$ are displayed for populations 1-7 of Example 1. As can be seen, there is a monotone relation between both selection tools. So, the PIC is related to the marginal likelihood approach, which is often used for inequality constrained model selection (see for example, Klugkist et al., 2005; Mulder, Hoijtink, & Klugkist, 2009).

6 Examples Reconsidered

After we have evaluated the performance of the *posterior* DIC, the *prior* DIC, the PIC and the ML for Example 1, it is now time to reconsider the other examples. For Examples 2 and 3 we only consider two populations: one population in agreement with the inequality constrained hypothesis and one population not in agreement with the constraints.

6.1 Example 2 continued

Let us return to Example 2 with $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_2 > 0, \mu_1 > 0$. To evaluate H_0 and H_1, we performed a small simulation study where data sets from two different populations were considered. Population 1 satisfy the constraints of H_1 and population 2 is not in agreement with H_1. The

	$\mu_0 - 1$			μ_0			$\mu_0 + 1$		
	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2
Population 1	203	**180**	203	182	**164**	183	202	**181**	202
Population 4	186	185	**178**	144	187	**136**	187	187	**178**
Population 7	**189**	200	228	**155**	167	187	**188**	199	228

	$\tau_0 \times .5$:			$\tau_0 \times 1$			$\tau_0 \times 5$		
	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2
Population 1	185	**160**	186	182	**164**	183	292	**270**	195
Population 4	150	151	**143**	144	143	**136**	215	214	**207**
Population 7	**162**	174	175	**155**	167	162	**236**	257	238

	$\upsilon_0 = 2$			$\upsilon_0 = 5$		
	H_0	H_1	H_2	H_0	H_1	H_2
Population 1	182	**160**	186	168	**145**	186
Population 4	144	143	**136**	130	129	**127**
Population 7	**155**	167	162	**141**	152	146

	$\sigma_0^2 \times .5$			$\sigma_0^2 \times 1$			$\sigma_0^2 \times 5$		
	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2
Population 1	213	**167**	246	183	**161**	186	200	**196**	177
Population 4	167	165	**156**	144	143	**136**	169	169	**159**
Population 7	**180**	203	218	**177**	181	179	**177**	181	179

Table 1: PIC values for different prior specifications. The bold numbers represent the hypothesis that is preferred by the PIC.

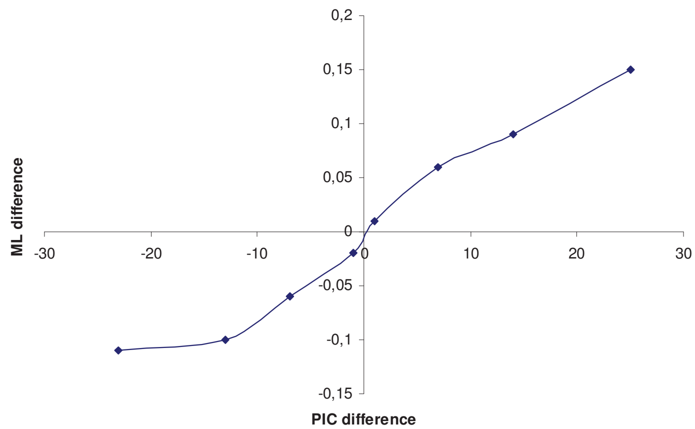

Figure 4: The differences between H_1 and H_2 are displayed for both the PIC and the ML for populations of Example 1.

two data sets were constructed in such a way that the sample means and variance-covariance matrix are exactly equal to the population parameters ($\rho = .4; \sigma_1^2 = 1; \sigma_2^2 = 1; n = 40$). For each of these data sets, we computed the *posterior* DIC, the *prior* DIC, and the PIC for H_0 and H_1.

According to Mulder, Hoijtink, and Klugkist (2009) the prior distribution, $\boldsymbol{\theta}_n) = h_0(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})\ h_0(\Sigma)$, can be given by a multivariate normal distribution for the means and an inverse Wishart distribution for the variance-covariance matrix

$$h_0(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \Sigma) = MVN(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\tau}_0^2) \times W^{-1}(\Sigma|\upsilon_0, \boldsymbol{\Sigma}_0), \tag{6.1}$$

where $\boldsymbol{\mu} = \{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$ and $\boldsymbol{\mu}_0 = \{\mu_0, \mu_0, \mu_0, \mu_0\}$. For the Inverse Wishart, we used $\upsilon_0 = 3$ and for $\boldsymbol{\Sigma}_0$, which is the scale matrix, we used

$$\begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix}, \tag{6.2}$$

For population 1 with $\mu_1 = 1$ and $\mu_2 = 1$, the priors are $\mu_0 = 0$, $\tau_0 = 0.98$, $\upsilon_0 = 3$ and $\sigma_0^2 = 3.95$. The results are shown in Table 2. As is illustrated in Table 2, the situation for this example is analogous to Example 1. Analogously to Example 1, the *prior* DIC does not correctly distinguish H_0 and H_1 because the loss function does not take 'bad' fitting hypotheses into account.

			post. DIC	*prior* DIC	PIC
Example 2	Population 1: $\mu_1 = 1, \mu_2 = 1$	H_0	234	489	428
		H_1	234	437	364
	Population 2: $\mu_1 = -.5, \mu_2 = -.5$	H_0	360	296	370
		H_1	404	294	406
Example 3	Population 1: $\beta_2 = 0.2$	H_0	275	919	820
		H_1	275	916	754
	Population 2: $\beta_2 = -0.2$	H_0	275	922	819
		H_1	411	924	960

Table 2: Results for Example 2 and 3.

6.2 Example 3 continued

For Example 3 we compared, $H_0 : \beta_2$ and $H_1 : \beta_2 > 0$. Analogously to Example 1 and 2, the *posterior* and *prior* DIC do not correctly distinguish H_0 and H_1 for Example 3. We also performed a small simulation study to evaluate H_0 and H_1. Data sets from two different populations were considered, see Table 2, where population 1 satisfy the constraints of H_1 and population 2 is not in agreement with H_1. The two data sets were constructed in such a way that the sample means and variance-covariance matrix are exactly equal to the population parameters ($\beta_0 = 1.0; \beta_1 = 0.5; \beta_2 = 0.2; n = 50$). The prior parameters we used are $\beta_1 = 0$, $\beta_2 = 0$, and $\sigma_0^2 = 1.95$. For each of these data sets, we computed the *posterior* DIC, the *prior* DIC, and the PIC for H_0 and H_1. The results are shown in Table 2 and it can be seen that the PIC outperforms the *posterior* and *prior* DIC.

6.3 Real-life Example 1

We re-evaluated the hypotheses given in (2.7). In Table 3 group means and standard deviations (SD) are provided. We computed the *posterior* DIC, the *prior* DIC, and the PIC for H_0, H_1 and H_2. The results of the model selection procedure are presented in Table 4. As can be seen in this table the *posterior* DIC is indifferent for all hypotheses, whereas both the *prior* DIC and the PIC choose for H_2. This result can be confirmed when looking at the group means in Table 3 where μ_{22} is larger than μ_{21} and μ_{11} is close to μ_{12}.

The theoretical conclusion is that there is support for a domain shift in the judgement about hypothetical situations. That is, for pupils that

Table 3: Descriptive Statistics for real-life example 1 ($n_1 = 38; n_2 = 97; \rho = .52$)

	Mean	SD
μ_{11}	5.37	1.23
μ_{12}	5.68	1.62
μ_{21}	5.27	1.27
μ_{22}	6.71	2.14

Table 4: Model Selection Results for the Real-life data 1.

Hypothesis	*post.* DIC	*prior* DIC	PIC
H_0	935	1976	1044
H_1	935	1952	1023
H_2	935	1803	872

reported to have conducted some delinquent behavior (i.e. aggression), in the same hypothetical situation, they will judge it to be more morally accepted compared to adolescents that did not report to conduct the same behavior. However, in hypothetical situations concerning other delinquent behavior that was not reported by these same adolescents (i.e. vandalism), they will judge the hypothetical situation to be equally morally condemnable as adolescents that did not report any antisocial behavior.

6.4 Real-life Example 2

We evaluated the hypothesis given in (2.10) using the *posterior* DIC, the *prior* DIC, and the PIC. The results are shown in Table 5. As can be seen in this table the *posterior* DIC fails to correctly distinguish the hypotheses of interest, whereas both the *prior* DIC and the PIC choose for H_2 as the best hypothesis. This result can be confirmed when looking at the group means in Table 6 where both $\beta_2 Age^2$ as well as $\beta_4 Age^2$ are both smaller than zero and $\beta_2 Age^2$ is smaller than $\beta_4 Age^2$.

The theoretical conclusion is that the relation between on the one hand age, and on the other hand either time to complete a Ph.D. trajectory or the gap between planned and actual project time are both non-linear. Moreover, this non-linear effect is stronger for time to complete a Ph.D. trajectory compared to the gap. This might be due to the fact that Ph.D candidates in the middle thirties take more time to finish their Ph.D thesis, but they also plan extra time.

Table 5: Model Selection Results for the Real-life data 2.

Hypothesis	*post.* DIC	*prior* DIC	PIC
H_0	4869	1862	1896
H_1	4872	1839	1855
H_2	4884	1834	1840

Table 6: Descriptive Statistics real-life example 2.

	Mean	SD
β_{01}	61.58	0.79
$\beta_1 Age$	2.88	0.27
$\beta_2 Age^2$	-0.96	0.01
β_{02}	10.54	0.82
$\beta_3 Age$	1.43	0.29
$\beta_2 Age^2$	-0.04	0.01

7 Conclusion

The main message of the current paper is: (1) although the DIC (Spiegelhalter et al., 2002) is often used in model selection, do not use it when evaluating inequality constrained hypotheses, better use the PIC which is derived in the current paper; and (2) the PIC is related to the marginal likelihood approach, which is often used for inequality constrained model selection (see for example, Klugkist et al., 2005; Mulder, Hoijtink, & Klugkist, 2009). We showed how to obtain the *prior* DIC based on the derivation of the *posterior* DIC presented in Spiegelhalter et al. (2002). The point of departure for the *prior* DIC is the same as for the *posterior* DIC, namely the expected loss. The derivation of the *prior* DIC is provided and the choice for the prior distribution, which is based on training data is motivated (see also Mulder, Hoijtink, & Klugkist, 2009). Its performance is illustrated using examples and we showed that the *prior* DIC can be used to choose between a set of constrained hypotheses if the population from which the data are generated is fully in agreement with the most constrained hypothesis, where the *posterior* DIC failed to do so. However, the *prior* DIC fails to choose between a set of inequality constrained hypotheses if the population is *not* in agreement with the constrained hypothesis.

In conclusion, neither the *prior* DIC, nor the *posterior* DIC are proper model selection tools for the evaluation of inequality constrained hypotheses. To accommodate for this, the loss function that is minimized by the *prior*

DIC was adjusted. The proposed loss function determines not only how well replicated data fit with a certain hypothesis, but it also determines how well a hypothesis fits the data. It is approximated by a new model selection tool, the Prior Information Criterium (PIC). We demonstrated with examples that the PIC is able to select the best of a set of (in)equality constrained hypotheses. More research is needed to evaluate under what conditions the PIC is expected to work well and under what other conditions is it expected to fail. However, since we showed that the marginal likelihood is highly related to the PIC, we expect that the PIC behaves similar as the marginal likelihood approach. The current paper adds to the growing body of evidence that classical model selection tools, like AIC, BIC, MDL and now also the DIC, are not equipped to deal with inequality constraints and offers a viable alternative.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267 - 281). Budapest: Akademiai Kiado.

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Journal of the Royal Statistical Society, series B*, *86*, 141-152.

Balasubramanian, V. (2005). Minimum description lenght. theory and applications. In P. J. Grunwald, I. J. Myung, & M. A. Pitt (Eds.), (p. 81-99). MIT Press: Boston.

Barlow, R. E., Bartholomew, D. J., Bremner, H. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions.* New York: Wiley.

Box, G. E. P. (1980). Sampling and bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, series A*, *143*, 383 - 430.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman&HallCRC.

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733 -807.

Grnwald, P. D., Myung, I. J., & Pitt, M. a. (2005). *Advances in minimum description lenght. theory and applications.* The MIT Press: Cambridge.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses*. New-York: Springer.

Kammers, M., Mulder, J., De Vignemont, F., & Dijkerman, H. (2009). The weight of representing the body: Addressing the potentially indefinite number of body representations in healthy individuals. *Experimental Brain Research*, *Published on-line, 22 sept. 2009*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*, 477 - 493.

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*, 69-86.

Leenders, I., & Brugman, D. (2005). Moral/non-moral domain shift in young adolescents in relation to delinquent behaviour. *British Journal of Developmental Psychology*, *23*, 65 - 79.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs - a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325 - 337.

Meeus, W., Van de Schoot, R., Keijsers, L., Schwartz, S. J., & Branje, S. (2010). On the progression and stability of adolescent identity formation. a five-wave longitudinal study in early-to-middle and middle-to-late adolescence. *Child Development*.

Mulder, J., Hoijtink, H., & Klugkist, I. (2009). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887-906.

Mulder, J., Klugkist, I., Van de Schoot, R., Meeus, W., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530-546.

Muthen, L. K., & Muthen, B. (2010). *Mplus users guide. sixth edition.* Los Angeles: Muthen & Muthen.

Myung, J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology 00*, *47*, 90-100.

Press, S. J. (2005). *Applied multivariate analysis: Using bayesian and frequentist methods of inference (2nd ed).* Malabar, FL: Krieger.

R Development Core Team. (2006). R:a language and environment for statistical computing (no. isbn 3-900051-07-0) [Computer software manual].

Rasbash, J., Charlton, C., Browne, W., Healy, M., & Cameron, B. (2009). *Mlwin version 2.1.* Centre for Multilevel Modelling, University of Bristol.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference.* New York : Wiley.

Romeijn, J.-W., Van de Schoot, R., & Hoijtink, H. (2011). One size does not fit all: Derivation of an adapted bic. In D. Dieks et al. (Eds.), *Probabilities, laws, and structures* (p. xxx-xxx). Berlin: Springer.

Schwarz. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Silvapulle, M. J., & Sen, P. K. (2004). *Constrained statistical inference: Order, inequality, and shape constraints.* London: John Wiley Sons.

Sober, E. (2006). Parsimony. In J. Pfeifer & S. Sarkar (Eds.), *The philosophy of science: An encyclopedia* (Vol. 2, p. 530-541). New York: Routledge.

Sonneveld, H., Yerkes, M., & Van de Schoot, R. (2009). *Ph.d. trajectories and labor market mobility: A survey of doctoral graduates in the netherland.* (Tech. Rep.). Report for Netherlands Centre for Graduate and Research Schools in the Netherlands and was subsidized by the Dutch Ministry of Education, Culture and Science.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, series B*, *64*, 583-639.

Van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, *17*, 443-463.

Van de Schoot, R., Romeijn, J.-W., & Hoijtink, H. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Quantitative Psychology and Measurement*.

Van de Schoot, R., & Wong, T. (2010). Do antisocial young adults have a high or a low level of self-concept? *Self and Identity*.

Van Well, S., Kolk, A. M., & Klugkist, I. (2009). The relationship between sex, gender role identification, and the gender relevance of a stressor on physiological and subjective stress responses: Sex and gender (mis)match effects. *International Journal of Psychophysiology*, *32*, 427-449.

A Derivation of Prior Predictive DIC

In this appendix we show how to obtain the *prior* DIC based on the derivation of the *posterior* DIC presented in Spiegelhalter et al. (2002). The point of departure for the *prior* DIC is the same as for the *posterior* DIC, namely the expected loss given in (3.1). However, to deal with the unknown parameters $\boldsymbol{\theta}^*$, we take the expectation with respect to the *prior* distribution, $h(\boldsymbol{\theta})$, instead of the *posterior* expectation of the expected loss:

$$\begin{aligned}&\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x}\mid\bar{\boldsymbol{\theta}}_y)\big]\Big\} = \\ &\qquad -2\log f(\mathbf{y}\mid\bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y},\boldsymbol{\theta},\bar{\boldsymbol{\theta}}_y)\big]\ .\end{aligned}\tag{A.1}$$

The main problem now, is to find an expression for the second term on the right hand side in (A.1). Using $D(\mathbf{a},\mathbf{b}) = -2\log f(\mathbf{a}\mid\mathbf{b})$, $c(\mathbf{y},\boldsymbol{\theta},\bar{\boldsymbol{\theta}}_y)$ in (A.1) can be rewritten to

$$\begin{aligned}c(\mathbf{y},\boldsymbol{\theta},\bar{\boldsymbol{\theta}}_y) &= \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\bar{\boldsymbol{\theta}}_y) - D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)\big] \\ &= \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\bar{\boldsymbol{\theta}}_y) - D(\mathbf{x},\boldsymbol{\theta})\big] \\ &+ \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\boldsymbol{\theta}) - D(\mathbf{y},\boldsymbol{\theta})\big] \\ &+ D(\mathbf{y},\boldsymbol{\theta}) - D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)\ .\end{aligned}\tag{A.2}$$

Now, $D(\mathbf{x},\bar{\boldsymbol{\theta}}_y)$ in (A.2) can be approximated by taking a second order Taylor expansion about $\boldsymbol{\theta}$,

$$\begin{aligned}D(\mathbf{x},\bar{\boldsymbol{\theta}}_y) \approx\ & -2\log f(\mathbf{x}\mid\boldsymbol{\theta}) - 2\left\{\frac{\partial \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\}^T\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big) - \\ & -\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)^T\left\{\frac{\partial^2 \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\}\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)\ .\end{aligned}\tag{A.3}$$

Since $-2\log f(\mathbf{x}\mid\boldsymbol{\theta})$ is equal to $D(\mathbf{x},\boldsymbol{\theta})$ and the expectation of the second term on the right hand side of (A.3) with respect to $f(\mathbf{x}\mid\boldsymbol{\theta})$ is zero (p. 604 Spiegelhalter et al., 2002),

$$\begin{aligned}&\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\bar{\boldsymbol{\theta}}_y) - D(\mathbf{x},\boldsymbol{\theta})\big] \approx \\ &\qquad \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\left[-\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)^T\left\{\frac{\partial^2 \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\}\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)\right]\ .\end{aligned}\tag{A.4}$$

The expression on the right hand side of (A.4) can be rewritten as $\mathrm{tr}\big\{\mathbf{I}(\boldsymbol{\theta})\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)^T\big\}$ and since $\mathbf{x}$ and $\mathbf{y}$ stem from the same data generating mechanism, the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ can be approximated by the observed Fisher information matrix, $\mathbf{I}(\bar{\boldsymbol{\theta}}_y)$ (p. 604 Spiegelhalter et al., 2002), where $\mathbf{I}(\bar{\boldsymbol{\theta}}_y) = -\partial^2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. Using $\mathrm{E}\{\mathrm{tr}(\cdot)\} = \mathrm{tr}\{\mathrm{E}(\cdot)\}$, the prior expectation of $c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)$ can now be approximated by:

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)\big] &\approx \mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\} + \\ &\quad + E_{h(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \boldsymbol{\theta}) - D(\mathbf{y}, \boldsymbol{\theta})\big]\Big\} + d \text{ ,} \end{aligned} \tag{A.5}$$

where $\boldsymbol{\Lambda} = E_{h(\boldsymbol{\theta})}[(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta})(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta})^T]$ denotes the variation in the prior distribution around $\bar{\boldsymbol{\theta}}_y$. The last term on the right hand side of (A.5) is defined as

$$\begin{aligned} d &= \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)\big] \\ &= \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) \text{ .} \end{aligned} \tag{A.6}$$

To show that $\mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\}$ is approximately equal to d, we use a second order Taylor expansion about $\bar{\boldsymbol{\theta}}_y$:

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] &\approx D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\Bigg[-2\Big\{\frac{\partial \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)}{\partial\boldsymbol{\theta}}\Big\}^T(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y) - \\ &\quad - (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y)^T\Big\{\frac{\partial^2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big\}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y)\Bigg] \text{ .} \end{aligned} \tag{A.7}$$

Since, $\bar{\boldsymbol{\theta}}_y \to \bar{\boldsymbol{\theta}}_{ML}$ for $n \to \infty$, $-2\Big\{\frac{\partial \log f(\mathbf{y}|\bar{\boldsymbol{\theta}}_y)}{\partial\boldsymbol{\theta}}\Big\}^T$ is asymptotically zero (Gelman et al., 2004). This way, $\mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big]$ can now be approximated by

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] &\approx D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\Bigg[\mathrm{tr}\Big\{-\frac{\partial^2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y)^T\Big\}\Bigg] \\ &\approx D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) + \mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\} \end{aligned} \tag{A.8}$$

To show that $\mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\}$ is approximately equal to d, $D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)$ is subtracted from both sides of (A.8)

$$\mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\} \approx \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) = d \text{ .} \tag{A.9}$$

Equation (A.5) then becomes

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)\big] &\approx \mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \boldsymbol{\theta}) - D(\mathbf{y}, \boldsymbol{\theta})\big]\Big\} + \\ &\quad + 2\Big\{\mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)\Big\} \text{ .} \end{aligned} \tag{A.10}$$

The *prior* DIC can now be written as

$$\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x}\mid\bar{\boldsymbol{\theta}}_y)\big]\Big\}\approx \\ \mathrm{E}_{h(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\boldsymbol{\theta})\big]\Big\}-D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)+\mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y},\boldsymbol{\theta})\big] \quad \text{(A.11)}$$

whereas, using the same notation, the *posterior* DIC can be written as

$$\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x}\mid\bar{\boldsymbol{\theta}}_y)\big]\Big\}\approx \\ D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)+2\Big\{\mathrm{E}_{g(\boldsymbol{\theta}|\mathbf{y})}\big[D(\mathbf{y},\boldsymbol{\theta})\big]-D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)\Big\}\ . \quad \text{(A.12)}$$

B Simplifying the prior DIC for constrained hypotheses

Let $H_t (t = 1, \ldots, T)$ denote a hypothesis specified using constraints and let H_0 denote an unconstrained hypothesis. All hypotheses H_t are nested in H_0, As we will prove in this section, $\mathrm{E}_{h_t(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\ \big[D(\mathbf{x},\boldsymbol{\theta})\big]\Big\}$ in (A.11) is constant between constrained hypotheses. In this context the *prior* DIC reduces to

$$\textit{prior}\ \mathrm{DIC}=C+2\log f(\mathbf{y}\mid\bar{\boldsymbol{\theta}}_y)+\mathrm{E}_{h_t(\boldsymbol{\theta})}\big[-2\log f(\mathbf{y}\mid\boldsymbol{\theta})\big]\ , \quad \text{(B.1)}$$

where $C=\mathrm{E}_{h_t(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x}\mid\boldsymbol{\theta})\big]\Big\}$ and can be ignored for all H_t.

B.1 Example 1 Continued

For Example 1, $h_t(\boldsymbol{\theta}_c)h_t(\boldsymbol{\theta}_u) = h_t(\mu_1,\mu_2)h_t(\sigma^2)$ where $h_t(\sigma^2)$ is the same, but $h_t(\mu_1,\mu_2)$ differs across hypotheses because of the normalization of the prior distribution in Equation (3.6). In the remainder of this subsection we drop the subscript t to simplify the notation. We will prove that $E_{h(\sigma^2)h(\mu_1,\mu_2)}\Big\{E_{f(\mathbf{x}|\mu_1,\mu_2,\sigma^2)}\ \big[-2\log f(\mathbf{x}\ \mid\ \mu_1,\mu_2,\sigma^2)\big]\Big\}$ is constant over all hypotheses under consideration. When comparing constrained hypotheses we have to prove that the term within accolades is independent of μ_1, μ_2, and σ^2. First using

$$f(\mathbf{x}\mid\mu_1,\mu_2,\sigma^2)=\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^N\exp\Big[-\frac{1}{2}\frac{\sum_{i=1}^N(x_i-\mu_1d_1-\mu_2d_2)^2}{\sigma^2}\Big]\ , \quad \text{(B.2)}$$

the term being constant can be written as

$$\int_{\sigma^2}\int_{\mu_1,\mu_2}\int_{\mathbf{x}} 2N\log\sqrt{2\pi\sigma^2}\ \partial f(\mathbf{x}\mid\mu_1,\mu_2,\sigma^2)\partial h(\mu_1,\mu_2)\partial h(\sigma^2)+ \tag{B.3}$$
$$+\int_{\sigma^2}\int_{\mu_1,\mu_2}\int_{\mathbf{x}}\sum_{i=1}^{N}\frac{(x_i-\mu_1 d_1-\mu_2 d_2)^2}{\sigma^2}\ \partial f(\mathbf{x}\mid\mu_1,\mu_2,\sigma^2)\partial h(\mu_1,\mu_2)\partial h(\sigma^2)\ .$$

The first term of (B.3) is independent of μ_1,μ_2, and since $h(\sigma^2)$ is the same for each hypothesis, the second term integrated over σ^2 in (B.3) should be constant for every value for σ^2 to render (B.3) constant. Let $\mathbf{x}=\{\mathbf{x}_1,\mathbf{x}_2\}$ denote subgroups with sample sizes N_1 and N_2 for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Omitting the integral over σ^2, we can now rewrite the second term in (B.3) to

$$\int_{\mu_1}\int_{\mathbf{x}_1}\sum_{i=1}^{N_1}\frac{(x_i-\mu_1)^2}{\sigma^2}\ \partial f(\mathbf{x}_1\mid\mu_1,\sigma^2)\partial h(\mu_1)+ \tag{B.4}$$
$$+\int_{\mu_2}\int_{\mathbf{x}_2}\sum_{i=1}^{N_2}\frac{(x_i-\mu_2)^2}{\sigma^2}\ \partial f(\mathbf{x}_2\mid\mu_2,\sigma^2)\partial h(\mu_2)\ .$$

Note, that for the first group in (B.4) $x_i\sim N(\mu_1,\sigma^2)$ and for the second group $x_i\sim N(\mu_2,\sigma^2)$. Using $x_i^*=\frac{x_i-\mu_1}{\sigma^2}$ with $x_i^*\sim N(0,1)$ in the first group, and $x_i^*=\frac{x_i-\mu_2}{\sigma^2}$ with $x_i^*\sim N(0,1)$ in the second group, the integral over μ_1 and μ_2 drop out of (B.4):

$$\int_{\mathbf{x}_1^*}\sum_{i=1}^{N}(x_i^*)^2\ \partial f(\mathbf{x}_i^*\mid 0,1)+\int_{\mathbf{x}_2^*}\sum_{i=1}^{N}(x_i^*)^2\ \partial f(\mathbf{x}_i^*\mid 0,1)\ . \tag{B.5}$$

Consequently, for every value of σ^2, (B.4) is independent of μ_1,μ_2. That is, for this example, $E_{h(\sigma^2)h(\mu_1,\mu_2)}\Big\{E_{f(\cdot)}\big[-2\log f(\cdot)\big]\Big\}$ is constant over constrained hypotheses.

B.2 Example 2 Continued

For Example 2, $h_t(\boldsymbol{\theta}_c)h_t(\boldsymbol{\theta}_u)=h_t(\mu_1,\mu_2)h_t(\Sigma)$ where $h_t(\Sigma)$ is the same, but $h_t(\mu_1,\mu_2)$ differs across hypotheses because of the normalization of the prior distribution in Equation (3.6). In the remainder of this subsection we drop the subscript t to simplify the notation. We now have to prove that the term between accolades in

$$E_{h(\mu_1,\mu_2)h(\sigma_{x1},\sigma_{x2},\rho)}\Big\{E_{f(\cdot)}\big[-2\log f(\mathbf{x}_1,\mathbf{x}_2\mid\mu_1,\mu_2,\sigma_{x1},\sigma_{x2},\rho)\big]\Big\} \tag{B.6}$$

is constant over hypotheses for μ_1, μ_2, and Σ. Using

$$\begin{aligned} f(\mathbf{x}_1, \mathbf{x}_2 \mid \mu_1, \mu_2, \sigma_{x1}, \sigma_{x2}, \rho) \quad = \quad & \Big(\frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}}\Big)^N \exp\Big[-\frac{1}{2(1-\rho^2)} \\ & \Big\{\frac{\sum_{i=1}^N (x_{1i}-\mu_1)^2}{\sigma_{x_1}^2} + \frac{\sum_{i=1}^N (x_{2i}-\mu_2)^2}{\sigma_{x_2}^2} - \\ & - \frac{2\rho\sum_{i=1}^N (x_{1i}-\mu_1)(x_{2i}-\mu_2)}{\sigma_{x_1}\sigma_{x_2}}\Big\}\Big], \end{aligned} \quad \text{(B.7)}$$

(B.6) can be written as the sum of

$$\begin{aligned} & \int_{\sigma_{x1},\sigma_{x2},\rho}\int_{\mu_1,\mu_2}\int_{\mathbf{x}_1,\mathbf{x}_2} 2N\log 2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2} \\ & \qquad \partial f(\mathbf{x}_1, \mathbf{x}_2 \mid \mu_1, \mu_2, \sigma_{x1}, \sigma_{x2}, \rho)\partial h(\mu_1,\mu_2)\partial h(\sigma_{x1},\sigma_{x2},\rho) \ , \end{aligned} \quad \text{(B.8)}$$

and

$$\begin{aligned} & \int_{\sigma_{x1},\sigma_{x2},\rho}\int_{\mu_1,\mu_2}\int_{\mathbf{x}_1,\mathbf{x}_2} \frac{1}{(1-\rho^2)} \\ & \quad \Big\{\frac{\sum_{i=1}^N (x_{1i}-\mu_1)^2}{\sigma_{x_1}^2} + \frac{\sum_{i=1}^N (x_{2i}-\mu_2)^2}{\sigma_{x_2}^2} - \frac{2\rho\sum_{i=1}^N (x_{1i}-\mu_1)(x_{2i}-\mu_2)}{\sigma_{x_1}\sigma_{x_2}}\Big\} \\ & \quad \partial f(\mathbf{x}_1, \mathbf{x}_2 \mid \mu_1, \mu_2, \sigma_{x1}, \sigma_{x2}, \rho)\partial h(\mu_1,\mu_2)\partial h(\sigma_{x1},\sigma_{x2},\rho) \ . \end{aligned} \quad \text{(B.9)}$$

Since $h(\Sigma)$ is the same for each hypothesis, the integrals in (B.9) integrated over $\sigma_{x1}, \sigma_{x2}, \rho$ should be constant for every value of $h(\Sigma)$ to render (B.9) constant. Also, in this situation (B.8) is constant over constrained hypotheses. Using $x^*_{1i} = \frac{x_{1i}-\mu_1}{\sigma}$ and $x^*_{2i} = \frac{x_{2i}-\mu_2}{\sigma}$, (B.9) can be rewritten into

$$\begin{aligned} & \int_\rho\int_{\mathbf{x}^*_1,\mathbf{x}^*_2}\sum_{i=1}^N \frac{1}{(1-\rho^2)}\Big\{(x^*_{1i})^2 + (x^*_{2i})^2 - 2\rho^2 x^*_{1i}x^*_{2i}\Big\} \\ & \qquad \partial f(\mathbf{x}^*_1, \mathbf{x}^*_2 \mid 0, 0, 1, 1, \rho)\partial(\rho) \ . \end{aligned} \quad \text{(B.10)}$$

Consequently, for every Σ, (B.9) is independent of μ_1 and μ_2. That is, for this example, $E_{h(\mu_1,\mu_2)h(\sigma_{x1},\sigma_{x2},\rho)}\Big\{E_{f(\cdot)}\big[-2\log f(\cdot)\big]\Big\}$ is constant over constrained hypotheses.

B.3 Multivariate Models

Finally, consider a multivariate example with two groups with mean scores on two dependent variables:

$$\begin{aligned} y_{1i} &= \mu_{11}d_{ig1} + \mu_{12}d_{ig2} + \epsilon_{1i} \\ y_{2i} &= \mu_{21}d_{ig1} + \mu_{22}d_{ig2} + \epsilon_{2i} \ , \end{aligned} \quad \text{(B.11)}$$

where $\mu_{1\cdot}$ and $\mu_{2\cdot}$ denote the mean score on y_1 and y_2 respectively and where $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ denote the mean for group 1 and 2 respectively. Again, group membership of a person is denoted by $d_{ig} \in 0, 1$ and the residuals are assumed to be normally distributed with

$$\left[\begin{array}{c} \epsilon_{i1} \\ \epsilon_{i2} \end{array} \right] \sim N\Big(0, \Sigma\Big) \ , \Sigma = \left[\begin{array}{cc} \sigma^2_{y_1} & \rho\sigma_{y_1}\sigma_{y_2} \\ \rho\sigma_{y_1}\sigma_{y_2} & \sigma^2_{y_2} \end{array} \right] . \tag{B.12}$$

Note that this example is a combination of (2.4) and (2.2). Also for constrained hypotheses in this multivariate example it can be proved that $E_{h_t(\mu_{11},\mu_{12},\mu_{21},\mu_{22})h_t(\Sigma)}\Big\{E_{f(\cdot)}\big[-2\log f(\mathbf{y}_1, \mathbf{y}_2 \mid \mu_{11}, \mu_{12}, \mu_{21},\ \mu_{22}, \Sigma)\big]\Big\}$ is constant over constrained hypotheses. Even so, using the same steps as presented in Section B.1 and B.2, it can be proved for the general multivariate normal linear model (Press (2005), pp. 252-257) that $\mathrm{E}_{h_t(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}$ $\big[-2\log f(\mathbf{x} \mid \boldsymbol{\theta})\big]\Big\}$ is constant over constrained hypotheses.