Chapter 3

Marginalization for DIC – Part I

After beginning with a discussion of statistical model selection, this chapter will present technical details regarding the deviance information criterion (DIC) and explore its behavior in the mixed modeling setting. We discuss the mathematical and philosophical differences between using marginalized vs. unmarginalized DIC computations, and we offer two schemes for numerical approximation of the DIC in the linear mixed model (LMM) setting.

3.1 Background

In this section, we provide an introduction to the topic of statistical model selection and we review a number of important developments therein. We focus primarily on the class of model selection criteria known as information criteria, based on their connection to information theory and their interpretation as the information lost through modeling—a notion that arises from Kullback and Leibler (1951).

3.1.1 Philosophy of Model Selection

George Box famously said, "All models are wrong, but some are useful" (Box and Draper 1987). While this is a valuable dictum, in the area of model selection we must recognize that stochastic objects such as experimental data must, necessarily, arise from some stochastic process. Even if that process is unknowable, the fundamental goal of model selection is to identify—subject to certain constraints—which class of models best matches the unknown data-generating process.

Conceptually, we can consider that somewhere in the space of all probability distributions there exists a unique distribution by which our data were generated. If we consider a restricted subspace of probability distributions such as a parametric family, model selection seeks to find some distribution within this subset that comes as close as possible to replicating the behavior of the data-generating distribution. Following the conventions of D. J. Spiegelhalter et al. (2002), hereafter SBCV, we refer to the former generating distribution as the true distribution and the latter approximating distribution as a pseudo-true distribution. For a given true distribution there may be many different pseudo-true distributions, each corresponding to a different subspace of probability distributions. A pseudo-true distribution is often a parametric distribution, and in this case we refer to the parameter of a pseudo-true distribution as a pseudo-true parameter. For example, a set of data might be modeled by either a Weibull or a log-normal distribution. A pseudo-true Weibull distribution and a pseudo-true log-normal distribution would both exist for that data, each with their own pseudo-true parameters. These distributions would be the closest fit in each class to the true data-generating distribution, but neither would necessarily be that true distribution.

The fundamental goal of statistical model selection is to identify a model with good predictive accuracy for some set of response data y. Model selection tools often marry a measure of goodness-of-fit to a measure of desirability. For example, Akaike's information criterion

(Akaike 1974) uses cross entropy as a measure of goodness-of-fit and adds a complexity penalty equal to the number of parameters in the model. In this way, the criterion tends to pick models with fewer parameters when they yield comparable cross entropies, but if additional complexity can result in appreciably better goodness-of-fit, a larger model may be preferred. Note, however, that when models are picked purely through comparison of criteria such as this, one cannot guarantee that the resulting model fits the data well—only that it fits the data better than the other models considered.

Before we turn to our own work with the deviance information criterion (DIC), we review the basis for information criteria beginning with the Kullback-Leibler (KL) divergence and early information criteria. We develop and explain the ideas behind the DIC, as well as discuss two newer information criteria that have been created to address issues related to selection in hierarchical models. We do this to provide a fuller understanding of the framework surrounding DIC: both its historical place and how it relates to other criteria that are frequently used in model selection.

3.1.2 Kullback-Leibler (KL) Divergence, 1951

Of particular interest here is what it means for two distributions to be similar to one another. Traditionally, one would use a distance metric to express this. A common method has been to use the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951).

Formally, consider two probability measures μ_0 and μ_1, both absolutely continuous with respect to a third measure λ. By the Radon-Nikodym theorem, for $j \in \{0,1\}$ there exist measurable functions f_j such that for a measurable set E under λ,

$$\mu_j(E) = \int_E f_j(y) d\lambda(y).$$

Kullback and Leibler define $\log \frac{f_0(y)}{f_1(y)}$ to be the information in y for discriminating between the hypothesis that y was selected from a population with probability measure μ_0 and the hypothesis that y was selected from a population with probability measure μ_1. The μ_0-directed KL divergence is defined as the expected amount of information for discriminating between these hypotheses contained in an observation from μ_0, namely

$$KL_{(\mu_0:\mu_1)} = \int \log\left(\frac{f_0(y)}{f_1(y)}\right) f_0(y) d\lambda(y).$$

The directed KL divergence is not a distance metric—it neither satisfies symmetry nor the triangle inequality. Even so, it is a premetric, satisfying the properties

$$KL_{(\mu_0:\mu_1)} \geq 0 \,\forall\, \mu_0, \mu_1 \qquad \text{and} \qquad KL_{(\mu_0:\mu_1)} = 0 \iff \mu_0 = \mu_1.$$

We can rewrite the formula for the directed divergence as

$$KL_{(\mu_0:\mu_1)} = \mathrm{E}_{\mu_0}[\log f_0(y)] - \mathrm{E}_{\mu_0}[\log f_1(y)]\,. \tag{3.1}$$

The term $-\mathrm{E}_{\mu_0}[\log f_0(y)]$ is known as the entropy of μ_0. The term $-\mathrm{E}_{\mu_0}[\log f_1(y)]$ is called the cross entropy of μ_0 and μ_1. Thus, the KL divergence is often conceptualized as the difference between the cross entropy of (μ_0, μ_1) and the entropy of μ_0 alone. The cross entropy forms the basis for many model selection procedures as we demonstrate below.

3.1.3 Akaike's Information Criterion (AIC), 1974

Akaike (1974) proposed an information-criterion-based model selection procedure, the AIC, that derives its validity from the interpretation of the negative cross entropy as a measure of the proximity between an inexact model and a true generating distribution, as demonstrated

with the KL divergence above. Akaike considers a scenario where μ_0 is the true generating distribution and μ_1 is an inexact modeling distribution whose fit we want to assess. He rewrites (3.1) as

$$KL_{(\mu_0:\mu_1)} = \int \log[f_0(y)] f_0(y) d\lambda y - \int \log[f_1(y)] f_0(y) d\lambda(y),$$

additively separating the component involving the modeling distribution from the component that involves only the generating distribution. Akaike shows how the above form can be used to compare different models, by recognizing that they share the unknown constant

$$c(\mu_0) = \int \log[f_0(y)] f_0(y) d\lambda y$$

that only involves the data-generating distribution.

What we are left with is an expectation over the data-generating distribution. Consider a set of alternate modeling distributions, $\mu_1, ..., \mu_M$. Then the directed KL divergence for μ_j relative to the generating distribution is given by

$$KL_{(\mu_0:\mu_j)} = c(\mu_0) - \mathrm{E}_{\mu_0}\big[\log f_j(y)\big] \qquad\qquad j \in \{1, ..., M\}.$$

Approximating this expectation with a sample mean of values that are generated by the unknown distribution μ_0 allows us to use the negative cross entropy as a measure of fit that can be compared across a range of modeling distributions.

Now assume our modeling distributions $\mu_1, ..., \mu_M$ are parametric, with parameter vectors $\theta_1, ..., \theta_M$ respectively. Assume further that the true distribution μ_0 is parametric, with θ_0. Then let a sample of data $y_i \overset{\text{iid}}{\sim} f(y_i \mid \theta_0)$, $i \in \{1, ..., n\}$ and a consider μ_j with density $f_j(y \mid \theta_j)$, $\theta_j \in \Theta_j$. There exists a maximum likelihood estimate (MLE) for θ_j based on these data, $\hat{\theta}_j$.

As Cavanaugh (1997) explains,

$$-2\sum_{i=1}^{n}\log f_j(y_i \mid \hat{\theta}_j)$$

is a biased estimator for twice the negative directed KL divergence between the true and fitted models,

$$-2\,\mathrm{E}_{\theta_0}\left[KL_{(\theta_0:\hat{\theta_j})}\right] = -2c(\mu_0) + 2\,\mathrm{E}_{\theta_0}\left[\sum_{i=1}^{n}\log f_j(y_i \mid \hat{\theta}_j)\right].$$

Further, this bias is asymptotically equal to twice the dimension of $\hat{\theta}_j$.

Then for this collection of models, Akaike writes

$$AIC_j = -2\sum_{i=1}^{n}\log f_j(y_i \mid \hat{\theta_j}) + 2k_j, \tag{3.2}$$

where k_j is the dimension of θ_j—the number of parameters θ_j includes. He is able to ignore $c(\mu_0)$ because it constitutes a fixed adjustment to each model, and is thus not useful in making comparing among those models.

Equation (3.2) marries a goodness-of-fit measure, the sample average cross entropy between θ_0 and θ_j, to a measure of model complexity, $2k_j$. This marriage is standard for information criteria, and the complexity penalty guards against overfitting. Because the MLE $\hat{\theta}_j$ is a function of the observed data, models with more parameters will tend to fit better than submodels that include only a subset of those parameters. Together, these measures converge to the cross entropy as long as θ_j is sufficiently close to θ_0. Model selection is performed by comparing AIC_j among a collection of models and choosing the model with the smallest AIC_j.

3.1.4 Bayes Factors (BF)

A common approach to model selection in the Bayesian setting is the Bayes factor (BF), the ratio of marginal likelihoods for the data under two distinct models. It is most easily understood in the context of hypothesis testing, where one of two models μ_1 or μ_2 is assumed to be the true distribution for some observed data, y. Our explanation below is derived from R. R. Christensen, Johnson, et al. (2010).

We define θ_1, θ_2 to be the parameters associated with μ_1, μ_2, and $f_1(y \mid \theta_1), f_2(y \mid \theta_2)$ their associated pdfs. In the Bayesian setting, we are also concerned with prior distributions on these parameters, $P_1(\theta_1 \mid \mu_1), P_2(\theta_2 \mid \mu_2)$; and prior probabilities for each model, q_1, q_2 where $q_1 + q_2 = 1$. We will use μ_T to denote the true model, whichever one it is.

The Bayes factor is based on the marginal predictive density for y,

$$f_j(y) = \int f_j(y \mid \theta_j) P_j(\theta_j \mid \mu_j) d\theta_j \qquad j \in \{1, 2\},$$

and the associated marginal likelihood $L(\mu_i \mid y) \propto f_j(y)$. The posterior probability of $\mu_1 = \mu_T$ is

$$\Pr[\mu_1 = \mu_T \mid y] = \frac{q_1 f_1(y)}{q_1 f_1(y) + q_2 f_2(y)}.$$

Then the posterior odds for $\mu_1 = \mu_T$ are

$$\begin{aligned} \frac{\Pr[\mu_1 = \mu_T \mid y]}{\Pr[\mu_2 = \mu_T \mid y]} &= \frac{\frac{q_1 f_1(y)}{q_1 f_1(y) + q_2 f_2(y)}}{\frac{q_2 f_2(y)}{q_1 f_1(y) + q_2 f_2(y)}} \\ &= \frac{q_1 f_1(y)}{q_2 f_2(y)} \\ &\equiv \frac{q_1}{q_2} BF \end{aligned}$$

Thus the Bayes factor comparing μ_1 to μ_2 is defined as

$$BF_{1:2} = \frac{f_1(y)}{f_2(y)}. \tag{3.3}$$

Since $\frac{q_1}{q_2}$ is the prior odds for $\mu_1 = \mu_T$, we can understand the Bayes factor as the degree to which our data change our prior beliefs about the odds. A Bayes factor above one means that the data favor the conclusion that $\mu_1 = \mu_T$, while a Bayes factor below one means that the data favor $\mu_2 = \mu_T$.

3.1.5 Bayesian Information Criterion (BIC), 1978

Schwarz (1978) provides the next major advance in the development of statistical information criteria. He begins by giving a concise summary of the criterion proposed by Akaike—quoted below[1]:

> An extension of the maximum likelihood principle is suggested... for the slightly more general problem of choosing among different models with different numbers of parameters. His suggestion amounts to maximizing the likelihood function separately for each model j, obtaining, say, $f_j(y_1, ..., y_n \mid \hat{\theta}_j)$, and then choosing the model for which $\log f_j(y_1, ..., y_n \mid \hat{\theta}_j) - k_j$ is largest, where k_j is the dimension of the model.

In contrast to this, Schwarz proposes that a model should instead minimize

$$BIC_j = -2 \sum_{i=1}^{n} \log f_j(y_i \mid \hat{\theta}_j) + k_j \log n, \tag{3.4}$$

[1]It is our standard practice in this dissertation, when quoting from other sources, to match their statements to our notation for the ease of the reader. We have endeavored to render the quoted material here and elsewhere as accurately as possible.

a value that what would later come to be called the Bayesian Information Criterion. Schwarz reasons asymptotically from the Bayesian strategy of picking the *a posteriori* most probable model from a class of models that are all given positive probability. This is very reminiscent of Equation (3.2). The only difference is the change in penalty term from $2k_j$ for AIC to $k_j \log n$ for BIC. The BIC penalty scales with the number of observed data values; and as more data are observed, BIC more strongly prefers a parsimonious model.

Formally Schwarz considers exponential family data with density $h(y)\exp\left(\theta t(y) - b(\theta)\right)$ where $\theta \in \Theta$ is multidimensional. Modeling distributions for these data, $\mu_1, ..., \mu_M$ depend on parameters $\theta_1, ..., \theta_M$ where θ_j lives in a k_j-dimensional subspace of Θ. Again, let q_j be the prior probability that model μ_j is correct, and let $P_j(\theta_j \mid \mu_j)$ be the prior distribution for θ_j conditional on μ_j. Then the Bayesian choice should select j to maximize

$$\begin{aligned}
S(j) &= \log\left(q_j f_j(y_1, ..., y_n)\right) \\
&= \log \int q_j \left(\prod_{i=1}^{n} h(y_i)\right) \exp\left(\theta_j \sum_{i=1}^{n} t(y_i) - nb(\theta_j)\right) dP_j(\theta_j \mid \mu_j) \\
&= \log\left(q_j \prod_{i=1}^{n} h(y_i) \int \exp\left(\theta_j \sum_{i=1}^{n} t(y_i) - nb(\theta_j)\right) dP_j(\theta_j \mid \mu_j)\right) \\
&= \log\left(\int \exp\left(\theta_j \sum_{i=1}^{n} t(y_i) - nb(\theta_j)\right) dP_j(\theta_j \mid \mu_j)\right) + \log q_j + \sum_{i=1}^{n} \log h(y_i)
\end{aligned}$$

Schwarz shows that

$$S(j) = \left(\hat{\theta}_j \sum_{i=1}^{n} t(y_i) - nb(\hat{\theta}_j)\right) - \frac{1}{2} k_j \log n + R_j,$$

where R_j is a remainder term bounded in n. As n grows, the boundedness of R_j means it goes away relative to $S(j)$ as a whole. This gives an asymptotic justification for using $k_j \log n$ as a complexity penalty in Equation (3.4) rather than Akaike's $2k_j$ in Equation (3.2). Of note is the fact that Schwarz's BIC relies on an asymptotic justification to eliminate prior beliefs about model preference—the prior model probability q_j is absorbed into the remainder R_j that is asymptotically eliminated.

Although the argument in Schwarz (1978) assumes the data come from an exponential family distribution, Cavanaugh and Neath (1999) show this result more generally. They let $Y = \{y_1, ..., y_n\}$ and let $f(Y)$ be the marginal density for the data over all models $\mu_1, ..., \mu_M$,

$$f(Y) = \sum_{l=1}^{M} q_l f_l(y_1, ..., y_n \mid \mu_l).$$

Then Cavanaugh and Neath show that

$$\begin{aligned} \log \Pr\left[\mu_j = \mu_T \mid Y\right] + \log f(Y) &\simeq \sum_{i=1}^{n} \log f_j(y_i \mid \hat{\theta}_j) - \frac{1}{2} k_j \log n \\ &= -\frac{1}{2} BIC_j \end{aligned}$$

Since $\log f(Y)$ does not depend on j, choosing a model based on $BIC(\theta_j)$ is asymptotically the same as choosing a model based on the posterior model probability.

From this, we can also see another interesting feature of the BIC: its asymptotic relation to the Bayes factor and the posterior odds. Observe that

$$\begin{aligned} \log \Pr\left[\mu_j = \mu_T \mid Y\right] + \log f(Y) &= \log\left(\frac{q_j f_j(Y \mid \mu_j)}{f(Y)}\right) + \log f(Y) \\ &= \log q_j + \log f_j(Y \mid \mu_j) \end{aligned}$$

Then using our definition of the BIC, (3.4), and the result from Cavanaugh and Neath (1999), we have

$$\begin{aligned} \log BF_{j:j'} &= \log f_j(Y \mid \mu_j) - \log f_{j'}(Y \mid \mu_{j'}) \\ &= \left(\log f_j(Y \mid \mu_j) + \log q_j - \log q_j\right) - \left(\log f_{j'}(Y \mid \mu_{j'}) + \log q_{j'} - \log q_{j'}\right) \\ &\simeq -\frac{1}{2}(BIC_j - BIC_{j'}) + (\log q_{j'} - \log q_j) \end{aligned}$$

If we assume equal prior probabilities for μ_j and $\mu_{j'}$, then, $\log BF_{j:j'} \simeq -\frac{1}{2}(BIC_j - BIC_{j'})$.

3.1.6 Log Pseudo-Marginal Likelihood (LPML), 1979

Geisser and Eddy (1979) argue that a fixed-penalty decision approach for choosing the wrong model, as Schwarz uses in his justification of the BIC, "may be reasonable for a selection procedure, but if the ultimate goal is prediction, then the penalty should depend both on sample size and type of error made." They build a predictive criterion, the log pseudo-marginal likelihood (LPML), using conditional predictive densities where each datapoint is fit knowing the rest of the data vector.

Let $Y_{(i)} = \{y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\}$. This is the collection of all the elements in Y except y_i. Then for a model μ_j, Geisser and Eddy define a conditional predictive ordinate

$$CPO_{ij} = f_j(y_i \mid Y_{(i)}, \mu_j)$$

and a pseudomarginal likelihood

$$\begin{aligned} L_j &= \prod_{i=1}^{n} CPO_{ij} \\ &= \prod_{i=1}^{n} f_j(y_i \mid Y_{(i)}, \mu_j) \end{aligned}$$

A. Gelfand and Dey (1994) provide an easy method for calculating the inverses of the conditional predictive ordinates from a posterior sample $\theta_j^{(1)}, ..., \theta_j^{(B)}$,

$$\begin{aligned} CPO_{ij}^{-1} &= \mathrm{E}_{\theta_j|Y}\left[\frac{1}{f(y_i \mid \theta_j, \mu_j)}\right] \\ &\doteq \frac{1}{B}\sum_{s=1}^{B} \frac{1}{f(y_i \mid \theta_j^{(s)}, \mu_j)} \end{aligned}$$

The log pseudo-marginal likelihood itself, as the name implies, is given by

$$\begin{aligned} LPML_j &= \log L_j \\ &= \sum_{i=1}^{n} \log CPO_{ij}. \end{aligned}$$

Although not created as an information criterion, the LPML shares many of the same model selection uses as the common information criteria when comparing two models. It also has a standard interpretation as a "pseudo Bayes factor" (R. R. Christensen, Johnson, et al. 2010), and thus shares a connection with BIC in that they both provide approximations to the same quantity. Watanabe (2010a) also proves the asymptotic equivalence of LPML and the widely applicable information criterion (WAIC) presented in Section 3.1.8.

3.1.7 Deviance Information Criterion (DIC), 2002

Whereas Akaike and Schwarz are concerned with the deviation between a known model (expressed through $\hat{\theta}$) and the truth, the approach of SBCV considers an average deviation from the truth for a possible model. As before, let $Y = \{y_1, ..., y_n\}$ be a set of observed data, and let $\mu_1, ..., \mu_M$ be a collection of models with associated parameters $\theta_1, ..., \theta_M$ and pdfs $f_1(Y \mid \theta_1), ..., f_M(Y \mid \theta_M)$. Instead of using the cross entropy as defined in Section 3.1.2, SBCV consider the posterior expectation of the log density. They define

$$DIC_j = -2\, \mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] + p_{Dj}, \tag{3.5}$$

where p_{Dj} is a penalization term that will be defined below.

SBCV frame their method around a quantity $D(\theta_j)$ defined as

$$D(\theta_j) = -2\left[\log f_j(Y \mid \theta_j) - \log f(Y)\right],$$

where SBCV describe $f(Y)$ as "some fully specified standardizing term that is a function of the data alone." Observe that $f(Y)$ will play a similar role to $c(\mu_0)$ in the AIC development above—as a constant term not based on the model being evaluated. Although it appears in the formal development of DIC, because it is an empirical function of the data alone, it will be irrelevant in comparing DICs for different models.

Based on $D(\theta_j)$, SBCV then define the quantities

$$\overline{D(\theta_j)} = -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] + 2\log f(Y)$$
$$D(\hat{\theta}_j) = -2\log f_j\left(Y \mid \hat{\theta}_j\right) + 2\log f(Y).$$

where $\hat{\theta}_j$ is some posterior summary for θ_j—most commonly a posterior mean, median, or mode. SBCV describe $\overline{D(\theta_j)}$ as a "Bayesian measure of fit [or] perhaps better considered a measure of 'adequacy'." It it the posterior expectation of their $D(\theta_j)$, and it quantifies how well the model fits on average, across the posterior distribution for θ_j. Meanwhile, they describe $D(\hat{\theta}_j)$ as a "classical 'plug-in' measure of fit," akin to the quantities used in the AIC and BIC.

This classical measure will tend fit be better[2] than SBCV's Bayesian measure—and in fact this is guaranteed to be the case when $f_j(Y \mid \theta_j)$ is log-concave in θ_j and we choose $\hat{\theta}_j$ to be the posterior mean of θ_j as SBCV recommend. Then Jensen's inequality guarantees

$$\log f_j\left(Y \mid \mathrm{E}_{\theta_j|Y,\mu_j}\left[\theta_j\right]\right) \geq \mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right],$$

which is the same as saying that the log-likelihood evaluated at the estimate $\hat{\theta}_j$ is larger than the posterior expectation of the log-likelihood function for θ_j.

[2]Because SBCV have structured their work around the idea of deviances, "better" can be difficult to follow here. $D(\hat{\theta}_j)$ is better than $\overline{D(\theta_j)}$ when $D(\hat{\theta}_j) < \overline{D(\theta_j)}$.

The penalization term, p_{Dj}, is then defined as

$$\begin{aligned} p_D &= \overline{D(\theta_j)} - D(\hat{\theta}_j) \\ &= -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] + 2\log f_j\left(Y \mid \hat{\theta}_j\right) \\ &= 2\left[\log f_j\left(Y \mid \hat{\theta}_j\right) - \mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right]\right], \end{aligned} \tag{3.6}$$

SBCV interpret this quantity as the degree of overfitting when a classical measure of fit, $D(\hat{\theta}_j)$, is used in place of a Bayesian measure of fit, $\overline{D(\theta_j)}$.

In models where the likelihood admits a normal approximation, SBCV argue that p_D is approximately the number of free parameters in the model. We give a more general argument below that also suggests an asymptotic equivalency between DIC_j and AIC_j under the conditions that $L_j(\theta_j \mid Y)$, the likelihood for θ_j, admit a normal approximation; and the prior $p_j(\theta_j \mid \mu_j)$ is sufficiently diffuse.

To begin, let $\hat{\theta}_j^{ML}$ be the MLE for θ_j and let $\hat{\theta}_j^{B}$ be the posterior mode for θ_j. We repeat Equations (3.2) and (3.5):

$$\begin{aligned} AIC_j &= -2\sum_{i=1}^{n}\log f_j\left(y_i \mid \hat{\theta}_j^{ML}\right) + 2k_j \\ DIC_j &= -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] + p_{Dj} \end{aligned}$$

Now observe that since

$$p_{Dj} = -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] + 2\log f_j\left(Y \mid \hat{\theta}_j^{B}\right),$$

we can also write

$$-2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] = p_{Dj} - 2\log f_j\left(Y \mid \hat{\theta}_j^{B}\right).$$

Then

$$DIC_j = -2\sum_{i=1}^{n} \log f_j\left(y_i \mid \hat{\theta}_j^B\right) + 2p_{Dj}.$$

To show $AIC_j \simeq DIC_j$, it is sufficient to show that $\hat{\theta}_j^{ML} \doteq \hat{\theta}_j^B$ and $k_j \doteq p_{Dj}$.

It is well known that when the likelihood admits a normal approximation and the prior is sufficiently diffuse; $\hat{\theta}_j^{ML}$, $\hat{\theta}_j^B$, and $\mathrm{E}_{\theta_j|Y,\mu_j}[\theta_j]$ are consistent estimators of the same quantity (Gelman et al. 2013, p.92), so asymptotic equivalence is clear. Further, the large sample approximation to the posterior is known to be

$$\theta_j \mid Y, \mu_j \dot{\sim} \mathrm{N}\left(\hat{\theta}_j^B, 2\ddot{D}(\hat{\theta}_j^B)^{-1}\right),$$

as discussed in Gelman et al. (2013, p.93).

What remains to be shown is that p_{Dj} is asymptotically equal to the number of parameters in the model, k_j. Using a second-order Taylor expansion, we observe that

$$D(\theta_j) \simeq D(\hat{\theta}_j^B) + \frac{1}{2}\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right),$$

where the first-order Taylor term is zero because our choice of $\hat{\theta}_j^B$ ensures that $\dot{D}(\hat{\theta}_j^B) = 0$. Then as David Hinkley would argue,

$$\begin{aligned}
\overline{D(\theta_j)} &= \mathrm{E}_{\theta_j|Y,\mu_j}[D(\theta_j)] \\
&\simeq \mathrm{E}_{\theta_j|Y,\mu_j}\left[D(\hat{\theta}_j^B) + \frac{1}{2}\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right)\right] \\
&= D(\hat{\theta}_j^B) + \frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right)\right]
\end{aligned}$$

Then since $\overline{D(\theta_j)} = D(\hat{\theta}_j^B) + p_{Dj}$, this indicates that

$$p_{Dj} \simeq \frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right)\right].$$

We now show that

$$\frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]\doteq k_j.$$

Since $\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]$ is a scalar, we can also write

$$\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]=\mathrm{tr}\left(\ddot{D}(\hat{\theta}_j^B)\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)\left(\theta_j-\hat{\theta}_j^B\right)^T\right]\right).$$

But our large sample posterior approximation gives

$$\begin{aligned}\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)\left(\theta_j-\hat{\theta}_j^B\right)^T\right]&\doteq\mathrm{Cov}_{\theta_j|Y,\mu_j}\left[\theta_j\right]\\&\doteq 2\ddot{D}(\hat{\theta}_j^B)^{-1}.\end{aligned}$$

Then

$$\begin{aligned}p_{Dj}&\simeq\frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]\\&\doteq\frac{1}{2}\,\mathrm{tr}\left(\ddot{D}(\hat{\theta}_j^B)\left(2\ddot{D}(\hat{\theta}_j^B)^{-1}\right)\right)\\&=\mathrm{tr}\left(\ddot{D}(\hat{\theta}_j^B)\ddot{D}(\hat{\theta}_j^B)^{-1}\right)\\&=k_j\end{aligned}$$

The interpretation of p_D as the number of free parameters in a model is widely considered to hold more generally than we argue above, but not without debate. The choice of a posterior summary $\hat{\theta}_j$ can affect p_{Dj}, even to the point of making it negative. There is also no guarantee that p_{Dj} will be positive when the posterior mean is chosen in cases where the density is not log-concave in θ_j, as is often the case with mixture models. Further, the DIC is not well defined in hierarchical models. Celeux et al. (2006) present eight possible DIC constructions for hierarchical models, differing in how the latent parameters are handled by

$\overline{D(\theta_j)}$ and $D(\hat{\theta}_j)$. We discuss these issues further in the following sections, and we present three DIC constructions pertaining to mixed modeling in Sections 3.3.1 below.

It is important to note also that DIC_j as it has been defined is not analytically tractable, although it is simple to numerically approximate it given an MCMC sample from the posterior distribution, θ_j, $\{\theta_j^{(1)}, ..., \theta_j^{(B)}\}$. Below is the standard computational form of DIC_j when $\hat{\theta}_j$ is taken to be the posterior mean.

$$DIC_j \simeq -\frac{4}{B}\sum_{s=1}^{B} \log f_j\left(Y \mid \theta_j^{(s)}\right) + 2\log f_j\left(Y \mid \frac{1}{B}\sum_{s=1}^{B}\theta_j^{(s)}\right)$$

3.1.8 Other Information Criteria

The Bayesian Predictive Information Criterion (BPIC; Ando 2007) and the Widely Applicable Information Criterion (WAIC; Watanabe 2010b) are two newer information criteria created to deal with the problems inherent in selection involving hierarchical models.

The first of these, the BPIC proposed by Ando in 2007, takes the form

$$\begin{aligned} BPIC_j &= -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] + 2n\hat{b}_{\theta_j} \\ &= \left(\overline{D(\theta_j)} - 2\log f(Y)\right) + 2n\hat{b}_{\theta_j} \end{aligned}$$

where μ is a parametric model for y with parameter vector θ. Computation of the penalization term $\hat{b}_\theta$ is quite involved, but it is meant to approximate the bias created by using $\overline{D(\theta_j)}$ rather than $\mathrm{E}_{\mu_0}[D(\theta_j)]$ as a measure of model fit, where μ_0 is the true generating distribution for the data y as before.

Formally,

$$b_{\theta_j} = \int \left(\frac{1}{n}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] - \mathrm{E}_z\left[\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(z \mid \theta_j)\right]\right]\right) d\mu_0(y),$$

where μ_0 is the true data-generating distribution, and where $z \sim \mu_0$. BPIC is, thus, using the same measure of fit—or as SBCV say, "adequacy"—as DIC, but with a different choice of penalization term to approximate the asymptotic bias in $\overline{D(\theta_j)}$ when the data are generated under an unknown distribution.

WAIC, proposed by Watanabe in 2010, is an attempt to build a Bayesian criterion that does not rely on plug-in point estimates of parameters. It depends on what Gelman et al. (2013) call the log pointwise predictive density (LPPD) for a model μ_j,

$$\begin{aligned} LPPD_j &= \sum_{i=1}^{n} \log f_j(y_i \mid Y, \mu_j) \\ &= \sum_{i=1}^{n} \log \int f_j(y_i \mid \theta_j) P_j(\theta_j \mid Y, \mu_j) d\theta_j \end{aligned}$$

where $P_j(\theta_j \mid Y, \mu_j)$ is the posterior density for θ_j. If a sample $\{\theta_j^{(1)}, ..., \theta_j^{(B)}\}$ is available from this posterior, then $LPPD_j$ can be numerically approximated with

$$LPPD_j \simeq \sum_{i=1}^{n} \log \left(\frac{1}{B} \sum_{s=1}^{B} f_j(y_i \mid \theta_j^{(s)}) \right).$$

Then $WAIC_j$ is defined as

$$WAIC_j = LPPD_j + p_{WAIC_j},$$

where p_{WAIC_j} is an overfitting penalty. Although Gelman et al. (2013, p.173) give two versions of this penalty, we do not intend to do an exhaustive review of the criterion here and will

only report the first.

$$
\begin{aligned}
p_{WAIC_j} &= 2\sum_{i=1}^{n}\Big(\log f_j(y_i \mid Y, \mu_j) - \mathrm{E}_{\theta_j|Y,\mu_j}\big[\log f_j(y_i \mid \theta_j)\big]\Big) \\
&= 2\sum_{i=1}^{n}\Big(\log \mathrm{E}_{\theta_j|Y,\mu_j}\big[f_j(y_i \mid \theta_j)\big] - \mathrm{E}_{\theta_j|Y,\mu_j}\big[\log f_j(y_i \mid \theta_j)\big]\Big) \\
&\simeq 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{B}\sum_{s=1}^{B} f_j(y_i \mid \theta_j^{(s)})\right) - \frac{1}{B}\sum_{s=1}^{B}\log f_j(y_i \mid \theta_j^{(s)})\right)
\end{aligned}
$$

WAIC has a number of nice properties. Principally, it does not rely on point estimation as AIC, BIC, and DIC do. It has also been shown to be asymptotically equivalent to LPML (Watanabe 2010a), as well as to Bayesian leave-one-out cross-validation (Gelman et al. 2013, p.176).

3.2 Model Selection in Mixed Models

As we mentioned in Section 3.1.7, mixed modeling is an area where the DIC is not well defined and many competing constructions have been offered (Celeux et al. 2006). In this section we introduce the mixed modeling framework. We then discuss how the definition of p_D in particular is complicated by these models. We close the section with an example to demonstrate the behavior of p_D in a simple random effects model.

3.2.1 The Mixed Modeling Framework

Mixed models, models that incorporate both fixed and random effects, are commonly used in statistical analysis. To understand their appeal, consider a simple linear regression of a response y on a covariate x.

Let y and x both be measured at multiple times on multiple individuals. Denote as y_{ij} the response on individual i at time j, and define x_{ij} analogously. In a simple linear regression we may write $y_{ij} = \beta_0 + x_{ij}\beta_1 + e_{ij}$, where β_0 is the intercept of a regression line, β_1 is the slope of that line, and e_{ij} is the amount by which y_{ij} differs from the value that would be predicted for it based on the regression line. The standard assumption is that $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for some constant variance σ^2 across all observations.

A linear mixed model (LMM) with random intercepts for each individual would be written as $y_{ij} = \beta_0 + x_{ij}\beta_1 + \gamma_i + e^*_{ij}$. Note that both of these models can apply to the same set of response and covariate data. In the LMM, we assume that $\gamma_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ and $e^*_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_r)$, with the γ's and e^*'s independent of each other. The total variability in the response data around the regression line is unchanged, but now we are splitting it into two terms: one that represents variability shared by observations on the same individual (γ_i) and another that represents leftover error that can't be attributed to individuals (e^*_{ij}). Since the total variability is the same, it is easy to see that $\sigma^2_r \leq \sigma^2$ and $\tau^2 \leq \sigma^2$. Because much of statistical inference depends on the amount of error in a dataset, using mixed models to account for between-subject variability allows statisticians to obtain more precise results when such variability exists. When such variability does not exist, $\tau^2 = 0$ and $\sigma^2_r = \sigma^2$, and the modeling cost incurred is simply that of estimating one extra parameter.

We proceed to give a mathematical definition for the mixed model that we use throughout the next three chapters.

Let $\mathbf{Y} = \{Y_i\} = \{y_{ij}\}$ be a $kn \times 1$ vector of response data on individuals $i \in \{1, ..., k\}$, with $j \in \{1, ..., n\}$ observations per individual. We use a balanced design with common n for all individuals to simplify some of the following linear algebra, but the results we obtain do not require this balance.

Let β be a $p \times 1$ vector of regression parameters. Let $\mathbf{X}$ be the $kn \times p$ design matrix for the regression parameters, $\mathbf{X}_i$ be the $n \times p$ block of the $\mathbf{X}$ matrix corresponding to cluster i, and X_{ij} be the $1 \times p$ row vector corresponding to the j^{th} observation on cluster i.

Let $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1^T & ... & \gamma_k^T \end{bmatrix}^T$ be a $kq \times 1$ vector of random effects, with γ_i the $q \times 1$ vector of random effects corresponding to cluster i. Let $\mathbf{Z}$ be the $kn \times kq$ block diagonal design matrix for the random effects. Let $\mathbf{Z}_i$ be the $n \times q$ submatrix of $\mathbf{Z}$ corresponding to its i^{th} diagonal block, and Z_{ij} be the $1 \times q$ row vector corresponding to the j^{th} row of the $\mathbf{Z}_i$ matrix.

Let $\boldsymbol{\psi} = \begin{bmatrix} \psi_1^T & ... & \psi_k^T \end{bmatrix}^T$ be the mean of the random effects vector $\boldsymbol{\gamma}$, and let $\boldsymbol{\Sigma}$ be block diagonal Σ_i, $i \in \{1, ..., k\}$ be the covariance matrix of the random effects. We assume that $\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma})$, or equivalently here that $\gamma_i \overset{\text{indep}}{\sim} \mathrm{N}_q(\psi_i, \Sigma_i)$. We use θ to refer to the collection of parameters $\{\beta, \boldsymbol{\psi}, \boldsymbol{\Sigma}\}$.

Then the linear mixed model can be written as

$$\begin{aligned} &y_{ij} = X_{ij}\beta + Z_{ij}\gamma_i + e_{ij}, \\ &\gamma_i \overset{\text{indep}}{\sim} \mathrm{N}_q(\psi_i, \Sigma_i), \\ &e_{ij} \overset{\text{iid}}{\sim} \mathrm{N}\Big(0, \sigma^2\Big). \end{aligned} \tag{3.7}$$

Or equivalently

$$\begin{aligned} &\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \\ &\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma}), \\ &\mathbf{e} \sim \mathrm{N}_{kn}\Big(0_{kn}, \sigma^2 I_{kn}\Big). \end{aligned} \tag{3.8}$$

3.2.1.1 The role of y and x

Both y and x are observed values gathered by researchers. As Bayesians, we consider these values to be fixed—*statheric* nodes, from the Greek word for 'constant', as contrasted with

the *stochastic* (random) nodes θ and γ. In a mixed model, we always assume some sort of structure in the observations that allows us to account for part of the observed response variability by grouping "like" observations together. Usually this is indicative of some type of observational unit or cluster: individuals who have been observed repeatedly, hospitals where data were gathered on multiple patients, nations whose economic output is observed over a number of years.

3.2.1.2 The role of θ

We use θ here, and throughout the next two chapters, to refer to the collection of all parameters in a model. In the simple linear regression example described above, $\theta = \{\beta_0, \beta_1, \sigma^2\}$ before random effects are added and $\theta = \{\beta_0, \beta_1, \tau^2\sigma_r^2\}$ when they are included. In the general form, θ refers to the collection of parameters $\{\beta, \boldsymbol{\psi}, \boldsymbol{\Sigma}, \sigma^2\}$.

3.2.1.3 The role of γ

The random terms, γ, allow us to efficiently account for unexplained subject-level variability without having to add a fixed-effect parameter for each individual. When random effects, γ, are included in a model, this is tantamount to making a statement that there are individual-level differences in the response—baseline differences and/or differences in covariate effect on the response—that are not captured by the fixed covariate effects model. Random effects act as a catchall for structural elements that the statistician hasn't built into a model. They are expressed efficiently because the only parameter they add to the model is a variance term for the individual-level differences.

The elements of γ are not themselves parameters, but are more accurately thought of as latent random variables. They are unknown stochastic objects whose inclusion can help us better understand our response data. The linear mixed model can be written in such

a way that γ is never specified—see the section below on preprocessing marginalization. Generalized linear mixed models (GLMMs), where the response data are assumed to arise from some non-normal distribution, do not allow for this convenient marginalization; but the role of γ as a latent random vector is the same.

3.2.2 Complications with p_D in Mixed Models

In this section, we will explain why p_D is not well defined for mixed models and how this relates to SBCV's notion of a model "focus", the collection of stochastic objects one is interested in. We then give an example showing how p_D can differ considerably, even in a simple model, depending on the focus one chooses.

In their initial paper on the Deviance Information Criterion (DIC), SBCV identify a key concern in applying DIC to mixed- and hierarchical models:

> Since the complexity [penalty p_D] depends on the focus, a decision must be made whether nuisance parameters, e.g. variances, are to be included in [the collection of model parameters] Θ or integrated out before specifying the model $P(x \mid \theta, \mu)$. However, such a removal of nuisance parameters may create computational difficulties.

To prevent confusion with traditional statistical notion for parameters (that do not admit the "nuisance parameters" mentioned by SBCV), we again distinguish between stochastic and statheric objects in a Bayesian model. Data and statically defined parameters for priors in the model are statheric: fixed by the analyst and not subject to MCMC sampling. All other objects—parameters for the data distribution and latent variables—are stochastic. In the parlance of SBCV, stochastic objects include both focal parameters (objects that interest the researcher) and nuisance parameters (objects that do not).

The crux of this issue is embedded in the definition of θ in Equations (3.5) and (3.6). If θ is the collection of parameters in a model, then the *DIC* should provide a reasonable model selection criterion. However, if θ is defined more generally as the collection of all stochastic objects in a Bayesian model, this becomes problematic. We demonstrate this with the following example.

Consider a simple case discussed in Hodges and Sargent (2001), a traditional random effects model, and see how this issue manifests. Let $\mathbf{Y} = \{y_{ij} : i \in \{1, ..., k\}, j \in \{1, ..., n\}\}$, where y_{ij} represents the j-th measurement of some variable on an individual i, and let $N = k * n$. Here and throughout this work, we use I_n to refer to an n-dimensional identity matrix, J_n to refer to an $n \times 1$ vector of 1's, and J_n^n to refer to an $n \times n$ matrix of 1's. We write this random effects model as

$$\begin{aligned} y_{ij} &= \gamma_i + \varepsilon_{ij} \\ \gamma_i &\sim \mathrm{N}\left(\psi, \frac{1}{\tau_g}\right) \\ \varepsilon_{ij} &\sim \mathrm{N}\left(0, \frac{1}{\tau_e}\right), \end{aligned} \tag{3.9}$$

or equivalently

$$\mathbf{Y} \sim \mathrm{N}_N\left(\psi J_N, \left(\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right)\right).$$

The parameters in this model are $\theta = \{\psi, \tau_g, \tau_e\}$. As previously discussed, the $k \times 1$ vector of $\boldsymbol{\gamma}$'s can be thought of as latent variables—unknown random objects that model correlation and extra heterogeneity in the data. By Equation (3.6), we can define a marginal p_D construction for this model:

$$\begin{aligned} p_{Dm} &= -2\left(\mathrm{E}_{\theta|y}[\log f(y \mid \theta)] - \log f\left(y \mid \mathrm{E}_{\theta|y}[\theta]\right)\right) \\ &= -2\left(\mathrm{E}_{\theta|y}\left[\log\left(\int f(y \mid \theta, \boldsymbol{\gamma}) P(\boldsymbol{\gamma} \mid \theta) d\boldsymbol{\gamma}\right)\right] - \log\left(\int f\left(y \mid \mathrm{E}_{\theta|y}[\theta], \boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma} \mid \mathrm{E}_{\theta|y}[\theta]\right) d\boldsymbol{\gamma}\right)\right) \end{aligned}$$

We use the subscript m to denote a marginalized p_D and henceforward we use the subscript j to denote what we will call a "joint" (or naive) p_D construction—that is, a p_D where the focus includes $\boldsymbol{\gamma}$ among the stochastic nodes of interest. Unfortunately, the value of p_D that many software packages calculate assumes γ to be a stochastic vector of interest, leading to the construction

$$p_{Dj} = -2\left(\mathrm{E}_{\theta,\boldsymbol{\gamma}|y}[\log f(y \mid \theta, \boldsymbol{\gamma})] - \log f\left(y \mid \mathrm{E}_{\theta,\boldsymbol{\gamma}|y}[\theta, \boldsymbol{\gamma}]\right)\right)$$

As the above example makes clear, there is a fundamental difference between p_{Dm} and p_{Dj}. Succinctly, the issues related to the application of p_D and DIC in instances like this are referred to as "the marginalization problem"—so named because the differences depend on whether or not γ is marginalized out before calculating p_D and DIC.

How big is this marginalization problem? Let us assume—uncharacteristically for this model, but it helps us to see an analytic example of the effect—that τ_g and τ_e are known. We also assume that ψ has an improper flat reference prior. With these assumptions, it is well known that

$$\begin{aligned}
\mathrm{E}[\psi \mid \mathbf{Y}] &= \frac{1}{N} J_N^T \mathbf{Y} && \equiv \overline{y}_{\cdot\cdot} \\
\mathrm{Var}[\psi \mid \mathbf{Y}] &= \frac{1}{N\tau_e} + \frac{1}{k\tau_g} && \equiv b
\end{aligned}$$

Further, let us define

$$\begin{aligned}
\overline{y}_{i\cdot} &= \frac{1}{n}\sum_{j=1}^{n} y_{ij} \\
\overline{Y} &= \begin{bmatrix} \overline{y}_{1\cdot} & \ldots & \overline{y}_{k\cdot} \end{bmatrix}^T \\
&= \left(\frac{1}{n} I_k \otimes J_n\right)^T \mathbf{Y}
\end{aligned}$$

Then we can write the distribution of the random effects, $\boldsymbol{\gamma}$, when $\mathbf{Y}$ and θ are known.

$$\gamma \mid \mathbf{Y}, \psi \sim \mathrm{N}_k\left(\frac{\tau_g}{\tau_g + n\tau_e}\psi J_k + \frac{n\tau_e}{\tau_g + n\tau_e}\overline{Y}, \frac{1}{\tau_g + n\tau_e} I_k\right).$$

We now give names to two quantities from the above distribution, to help us simplify our work below:

$$w \equiv \tau_g/(\tau_g + n\tau_e)$$

$$v \equiv 1/(\tau_g + n\tau_e)$$

This allows us to rewrite the above distribution of $\boldsymbol{\gamma}$ as

$$\gamma \mid \mathbf{Y}, \psi \sim \mathrm{N}_k\big(w\psi J_k + (1-w)\overline{Y}, vI_k\big).$$

Then plugging into the formula for p_{Dm}, we have

$$\begin{aligned}
&\mathrm{E}_{\theta|\mathbf{Y}}[\log f(y \mid \theta)] \\
&\quad = \mathrm{E}_{\psi|\mathbf{Y}}\left[\log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right|^{-1/2}\right) - \frac{1}{2}\left(\mathbf{Y} - \psi J_N\right)^T \left[\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right]^{-1} \left(\mathbf{Y} - \psi J_N\right)\right] \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right|^{-1/2}\right) - \mathrm{E}_{\psi|\mathbf{Y}}\left[\frac{1}{2}\left(\mathbf{Y} - \psi J_N\right)^T \left[\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right]^{-1} \left(\mathbf{Y} - \psi J_N\right)\right]
\end{aligned}$$

for the first term, and

$$
\begin{aligned}
&\log f\left(y \mid \mathrm{E}_{\theta|\mathbf{Y}}[\theta]\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right|^{-1/2}\right) \\
&\qquad - \frac{1}{2}\left(\mathbf{Y} - \mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right|^{-1/2}\right) - \frac{1}{2}\left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right)
\end{aligned}
$$

for the second term. Since the constants of integration are the same and the logs of those constants cancel, combining these two terms leads to the following equation for p_{Dm}.

$$
\begin{aligned}
p_{Dm} &= \mathrm{E}_{\psi|\mathbf{Y}}\left[\left(\mathbf{Y} - \psi J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \psi J_N\right)\right] \\
&\qquad - \left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right) \\
&= \left(\mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y} - \mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad - \left(\mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad - \left(\mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N - \mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\overline{y}_{\cdot\cdot}J_N\right) \\
&\qquad + \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\psi J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\psi J_N\right] - \overline{y}_{\cdot\cdot}J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\overline{y}_{\cdot\cdot}J_N\right) \\
&= \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\psi^2\right] - \overline{y}_{\cdot\cdot}^2\right) J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1} J_N \\
&= \mathrm{Var}[\psi \mid \mathbf{Y}]\, J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1} J_N \\
&= b\tau_e \sum_{i=1}^{k} J_n^T \left[I_n - \frac{\tau_e}{\tau_g + n\tau_e}J_n^n\right] J_n \\
&= N\tau_e wb \\
&= 1
\end{aligned}
$$

So p_{Dm} in this setting is identically equal to 1, which is what we would like. The only free parameter in the model we have described is ψ—since τ_e and τ_g are both known; and although the random vector $\boldsymbol{\gamma}$ is a stochastic node in an MCMC sampler, it is not a parameter vector.

We now turn our attention to the calculation of p_{Dj}, the p_D construction with naive focus. Again, we start by specifying the elements of the p_{Dj} formula for this problem, given above.

The first term is given by

$$\begin{aligned}
&\mathrm{E}_{\theta,\boldsymbol{\gamma}|\mathbf{Y}}[\log f(y \mid \theta, \boldsymbol{\gamma})] \\
&\quad = \mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[\log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) - \frac{1}{2}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)\right] \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) - \mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[\frac{1}{2}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)\right].
\end{aligned}$$

The second term is more complicated, necessitating our use of the Law of Total Expectation.

$$\begin{aligned}
&\log f\left(y \mid \mathrm{E}_{\theta,\boldsymbol{\gamma}|\mathbf{Y}}[\theta,\boldsymbol{\gamma}]\right) \\
&\quad = \log f\left(y \mid \mathrm{E}_{\psi|\mathbf{Y}}[\psi], \mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{E}_{\boldsymbol{\gamma}|\mathbf{Y},\psi}[\boldsymbol{\gamma}]\right]\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) \\
&\qquad\quad - \frac{1}{2}\left(\mathbf{Y}-\mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{E}_{\boldsymbol{\gamma}|\mathbf{Y},\psi}[\boldsymbol{\gamma}]\right]\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{E}_{\boldsymbol{\gamma}|\mathbf{Y},\psi}[\boldsymbol{\gamma}]\right]\otimes J_n\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) \\
&\qquad\quad - \frac{1}{2}\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right).
\end{aligned}$$

Once again, we recognize that these terms have equal constants of integration, and that the logs of those constants cancel. Combining terms, we obtain the following equation for p_{Dj}.

$$\begin{aligned}
p_{Dj} &= \mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n)\right] \\
&\qquad -\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right) \\
&= \left(\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}-\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad -\left(\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}[\boldsymbol{\gamma}]\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad -\left(\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}[\boldsymbol{\gamma}]\otimes J_n\right)-\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right) \\
&\qquad +\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[(\boldsymbol{\gamma}\otimes J_n)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}(\boldsymbol{\gamma}\otimes J_n)\right] \\
&\qquad -\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)
\end{aligned}$$

Note that each of the first three lines of the foregoing equality are equal to 0. Then

$$\begin{aligned}
p_{Dj} &= \tau_e\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[(\boldsymbol{\gamma}\otimes J_n)^T\boldsymbol{\gamma}\otimes J_n\right]-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)^T\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right) \\
&= n\tau_e\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[\boldsymbol{\gamma}^T\boldsymbol{\gamma}\right]-\left(w\overline{y}_{\cdot\cdot}J_k+(1-w)\overline{Y}\right)^T\left(w\overline{y}_{\cdot\cdot}J_k+(1-w)\overline{Y}\right)\right) \\
&= n\tau_e\left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\sum_{i=1}^{k}\mathrm{E}_{\boldsymbol{\gamma}_i|\mathbf{Y},\psi}\left[\boldsymbol{\gamma}_i^2\right]\right]-\sum_{i=1}^{k}(w\overline{y}_{\cdot\cdot}+(1-w)\overline{y}_{i\cdot})^2\right) \\
&= n\tau_e\sum_{i=1}^{k}\left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{Var}_{\gamma_i|\mathbf{Y},\psi}[\gamma_i]+\mathrm{E}_{\gamma_i|\mathbf{Y},\psi}[\gamma_i]^2\right]-(w\overline{y}_{\cdot\cdot}+(1-w)\overline{y}_{i\cdot})^2\right)
\end{aligned}$$

Recalling that $\gamma \mid \mathbf{Y},\psi \sim \mathrm{N}_k\left(w\psi J_k+(1-w)\overline{Y},vI_k\right)$, we can now finish simplifying the equation for p_{Dj} in this setting.

$$
\begin{aligned}
p_{Dj} &= n\tau_e \sum_{i=1}^{k} \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[v + (w\psi + (1-w)\overline{y}_{i\cdot})^2 \right] - (w\overline{y}_{\cdot\cdot} + (1-w)\overline{y}_{i\cdot})^2 \right) \\
&= N\tau_e v + n\tau_e \sum_{i=1}^{k} \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[w^2\psi^2 + 2w(1-w)\overline{y}_{i\cdot}\psi + (1-w)^2\overline{y}_{i\cdot}^2 \right] - (w\overline{y}_{\cdot\cdot} + (1-w)\overline{y}_{i\cdot})^2 \right) \\
&= N\tau_e v + n\tau_e \sum_{i=1}^{k} \left(w^2 \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\psi^2\right] - \overline{y}_{\cdot\cdot}^2 \right) + 2w(1-w)\left(\overline{y}_{i\cdot}\,\mathrm{E}_{\psi|\mathbf{Y}}[\psi] - \overline{y_{i\cdot}}^2 \right) + (1-w)^2\left(\overline{y_{i\cdot}}^2 - \overline{y_{i\cdot}}^2 \right) \right) \\
&= N\tau_e v + n\tau_e \sum_{i=1}^{k} \left(w^2 \,\mathrm{Var}_{\psi|\mathbf{Y}}[\psi] \right) \\
&= N\tau_e(v + w^2 b) \\
&= 1 + (k-1)\frac{n\tau_e}{\tau_g + n\tau_e}
\end{aligned}
$$

As we saw, p_{Dm} is identically equal to 1 in this setting. Similarly, when τ_g is much larger than $n\tau_e$, p_{Dj} also approaches 1. When τ_g is much smaller than $n\tau_e$, however, p_{Dj} approaches k. This is reasonable given the choice of focus and the situation described. The quantity p_{Dm} identifies a single free parameter, ψ. When $\tau_g \gg n\tau_e$, there is very little variability in the random effects terms relative to the variability within individuals, and the data behave like they come from a common population and p_{Dj} is near 1. When $\tau_g \ll n\tau_e$ on the other hand, the data behave like k separate populations, each having their own effect.

This, then, is the marginalization problem. Although the marginal construction gives $p_D = 1$, as we would expect, the naive construction gives $1 \leq p_D \leq k$. In the next section, we discuss why we believe this inconsistency necessitates use of the marginal construction.

3.3 The Need for Marginalization

We will now endeavor to describe three distinct constructions for p_D and DIC in the mixed model setting. These are the joint and marginal constructions, as discussed above, and the

BUGS numerical approximations. After discussing these three constructions, we proceed to give our argument as to why we believe the marginal construction should be preferred. Finally, we discuss some counterarguments against a preference for marginalization in model selection.

3.3.1 Three DIC Constructions for Mixed Models

In the following three chapters, we make extended reference to three different numerical approximations to DIC: the joint DIC, the BUGS DIC, and the marginal DIC. As explained above, in mixed models the value of the DIC depends on the choice of focal objects. We now explain the difference between these three approximations, and why we consider the marginal DIC to be philosophically preferable for doing model selection in the mixed model setting.

3.3.1.1 The joint DIC

The first construction for DIC is what one might consider the naive construction. This construction assumes that all stochastic objects are of focal interest. It centers on what we call a joint likelihood for both θ and γ,

$$L(\theta, \gamma \mid y) \propto f(y \mid \gamma, \theta).$$

Under the joint construction, we have the following definitions:

$$\begin{aligned} p_{Dj} &= \mathrm{E}_{\theta,\gamma|y}[-2\log L(\theta,\gamma \mid y)] + 2\log L(\hat{\theta},\hat{\gamma} \mid y) \\ &= \overline{D(\theta,\gamma)} - D(\hat{\theta},\hat{\gamma}), \\ DIC_j &= 2\,\mathrm{E}_{\theta,\gamma|y}[-2\log L(\theta,\gamma \mid y)] + 2\log L(\hat{\theta},\hat{\gamma} \mid y) \\ &= 2\overline{D(\theta,\gamma)} - D(\hat{\theta},\hat{\gamma}). \end{aligned}$$

In other words, the joint DIC treats γ as if it were a model parameter alongside θ, and uses the posterior mean of both θ and γ to obtain the fitted deviance. Here, $\hat{\theta} = \mathrm{E}_{\theta,\gamma|y}[\theta]$ and $\hat{\gamma} = \mathrm{E}_{\theta,\gamma|y}[\gamma]$.

The joint construction above is similar to the construction for DIC_7 in Celeux et al. (2006), with two notable differences. First, they choose $\hat{\theta}$ and $\hat{\gamma}$ to be the joint maximum *a posteriori* (MAP) estimates conditional on y. We choose $\hat{\theta}$ and $\hat{\gamma}$ to be the posterior means. Their choice to use joint MAP estimates is based on the poor behavior of estimators of γ in latent random variable problems and their concern with how the DIC behaves in mixture models, where the posterior mean may live in an area of relatively low posterior density. Our use of the posterior mean stems from SBCV's recommendation to use posterior means and our belief that this construction is more likely to be used than DIC_7 by others who might encounter a mixed modeling scenario.

Second, while we call this a joint construction, Celeux et al. (2006) call it a conditional construction. This is a fundamental difference in how we regard these constructions. We call this a joint construction because θ and γ appear jointly in a likelihood statement, and are considered jointly by the DIC formulae. They call this a conditional construction because the density (as opposed to likelihood) of interest is $f(y \mid \gamma, \theta)$, where y is conditioned on γ.

3.3.1.2 The BUGS DIC

Our information on how OpenBUGS constructs the DIC is drawn from *The BUGS Book* (Lunn et al. 2013) and the *OpenBUGS User Manual* (D. Spiegelhalter et al. 2014). The manual describes how OpenBUGS obtains $\overline{D(\theta)}$ and $D(\hat{\theta})$ as follows:

> $\overline{D(\theta)}$: this is the posterior mean of the deviance, which is exactly the same as if the node 'deviance' had been monitored. This deviance is defined as $-2\log P(y|\theta)$, where y comprises all stochastic nodes given values (i.e. data), and θ comprises the stochastic parents of y – 'stochastic parents' are the stochastic nodes upon which the distribution of y depends, when collapsing over all logical relationships.
>
> $D(\hat{\theta})$: this is a point estimate of the deviance ($-2\log P(y|\theta)$) obtained by substituting in the posterior means $\frac{1}{B}\sum_{s=1}^{B}\theta^{(s)}$ of θ: thus $D(\hat{\theta}) = -2\log p\left(y \mid \frac{1}{B}\sum_{s=1}^{B}\theta^{(s)}\right)$

As this construction pertains to hierarchical models, *The BUGS Book* describes numerical approximations to DIC in the BUGS family of programs as follows:

> WinBUGS (and OpenBUGS) separately reports the contribution to $\overline{D(\theta)}$, p_D, and DIC for each differently named (scalar, vector, or array) node, together with a total. This enables the individual contributions from different portions of data to be assessed. In some circumstances some of these contributions may need to be ignored and removed from the total.

This is not, unfortunately, enough information to classify OpenBUGS's construction of the DIC for hierarchical models into the framework provided by Celeux et al. (2006). We do not have sufficient information on which nodes constitute stochastic parents of our data in the mixed model. We can state, however, that for all the models considered in this dissertation, OpenBUGS reports DIC contributions from only our data y and cannot be partialed out

as described above. We believe, based on the information presented in the manual and *The BUGS Book*, that OpenBUGS's construction of the DIC should match our DIC_j, but simulation results presented in Chapters 4 and 5 confirm that there are differences in the numerical approximation algorithms.

We nonetheless present results for DIC_b and related quantities because we consider it important to compare the numerical approximations under methods we develop to what is given by commercially available software, since numerical approximations to DIC given by commercially available software are what practitioners are most likely to use.

3.3.1.3 The marginal DIC

Our final construction focuses only on the stochastic node θ, treating γ as a latent random vector to be marginalized out. The marginal DIC can be expressed using three different likelihood functions, which we briefly clarify before giving the formulae for this construction.

$$
\begin{aligned}
L(\theta \mid y, \gamma) &\propto f(y, \gamma \mid \theta) \\
L(\theta, \gamma \mid y) &\propto f(y \mid \gamma, \theta) \\
L(\theta \mid y) &\propto f(y \mid \theta) \\
&= \int f(y, \gamma \mid \theta) d\gamma \\
&= \int f(y \mid \gamma, \theta) P(\gamma \mid \theta) d\gamma \\
&= \mathrm{E}_{\gamma \mid \theta}[L(\theta, \gamma \mid y)]
\end{aligned}
$$

Then we define the marginal construction with a few equivalent expressions:

$$\begin{aligned}
p_{Dm} &= \mathrm{E}_{\theta|y}[-2\log L(\theta \mid y)] + 2\log L(\hat{\theta} \mid y) \\
&= \mathrm{E}_{\theta|y}\left[-2\log \int L(\theta \mid y, \gamma) d\gamma\right] + 2\log \int L(\hat{\theta} \mid y, \gamma) d\gamma \\
&= \mathrm{E}_{\theta|y}\left[-2\log \mathrm{E}_{\gamma|\theta}[L(\theta, \gamma \mid y)]\right] + 2\log \mathrm{E}_{\gamma|\theta}\left[L(\hat{\theta}, \gamma \mid y)\right] \\
&= \overline{D(\theta)} - D(\hat{\theta}) \\
DIC_m &= 2\,\mathrm{E}_{\theta|y}[-2\log L(\theta \mid y)] + 2\log L(\hat{\theta} \mid y) \\
&= 2\overline{D(\theta)} - D(\hat{\theta}).
\end{aligned}$$

where

$$\begin{aligned}
\hat{\theta} &= \mathrm{E}_{\theta|y}[\theta] \\
&= \int \theta P(\theta \mid y) d\theta \\
&= \int \theta \int P(\theta, \gamma \mid y) d\gamma d\theta \\
&= \int \int \theta P(\theta, \gamma \mid y) d\gamma d\theta \\
&= \mathrm{E}_{\theta,\gamma|y}[\theta]
\end{aligned}$$

Note that, as shown above, the quantities in the marginal construction can be written both as integrals of likelihoods and as expectations over the distribution $P(\gamma \mid \theta)$. These are subtly different interpretations, and both will prove useful to us in our discussion of methods for approximating DIC_m in Section 4.2.2.

This construction is given by Celeux et al. (2006) as DIC_1, who refer to it as an "observed DIC" to match their terminology for $L(\theta \mid y)$, which they call an observed likelihood.

To simplify notation where necessary, we use the j subscript (e.g. p_{Dj}, DIC_j, $\overline{D}_j$, $D(\hat{\theta})_j$) to refer to the DIC calculations focusing on the joint distribution of θ and γ. Similarly, we use the b subscript for the BUGS calculations and the m subscript for the marginal calculations.

3.3.2 Why Do We Prefer the Marginal DIC?

Having established that p_D can depend the set of focal stochastic objects with our example in Section 3.2.2, we now consider whether this dependence is worth our concern. We believe so, and in the subsections below we make our case for using the marginal DIC. We give three reasons based on: (1) the conceptual difference between adopting the marginal or the joint foci, (2) the rise of automated model selection procedures, (3) and the interpretation of p_D in hierarchical models.

3.3.2.1 Conceptual differences between marginalized and joint foci

We begin with a discussion of what it means for the DIC to focus on θ and γ, rather than θ alone. An individual-level random effect can, in general, be thought of as a catch-all correction factor encapsulating all of the remaining differences among individuals that are germane, after conditioning on every covariate already measured and included in the model. For example, if we consider the human fertility study referenced in the preceding chapter, a model might suggest that the probability of ovulation during a particular cycle is a function of certain covariates: e.g. ethnicity, age, weight, average daily caffeine intake, etc. But information on these covariates alone may not be sufficient to describe the differences observed among individuals. There may also be additional random slopes—individual differences in the relationships between time-varying covariates and the response of interest.

Random effects are usually specified by a particular parametric distribution, and the random effect for each individual is assumed to be independently drawn from this distribution. Researchers may be interested in the parameters of this random effects distribution—measures that reflect how much inter-individual variability remains in the response data that hasn't been captured by the covariates[3]. These so-called variance components are included in θ. The (random) γ's are the latent random effects, which under such a distribution reflect how far an individual's response differs from the overall population mean, adjusted for observed covariates.

The choice of whether one is interested in making inferences for particular γ's is essentially a choice about whether one is concerned with the population of individuals who haven't been sampled, or concerned only with the individuals in the sample. Both choices can be reasonable—but when one is concerned only with understanding the individuals in the sample, this is more accurately reflected by considering a fixed effect for those individuals. The choice to consider random, rather than fixed, effects is essentially a choice to prioritize generalizability. Otherwise why would one be concerned with the distributional properties of the random effects?

We consider that the marginalized approach to be philosophically preferable. The real distinction between fixed and random effects is whether one wants to make specific inferences about the observed clusters in particular, or whether one wants to extrapolate to the general population from which those clusters were sampled. If one wants to make inferences about the observed clusters, then one should fit a fixed effects model. The usual DIC, in that case, requires no marginalization. If one doesn't care about observed clusters, then the only parameter of interest should be the covariance matrix for the random effects, $\mathbf{\Sigma}$. This leads

[3]Note that the example in Section 3.2.2 was developed under the assumption that these values were known. This was done in order to provide insight about the behavior of p_D as a function of the precisions for the random effects and error distributions. SBCV consider a similar example with ψ constant and τ_e unknown, obtaining analogous results.

to the need to marginalize over the γ and simply focus on the parameters of interest, which includes $\mathbf{\Sigma}$ as well as any other model parameters.

Suggestions have been made to us that a scientist might be concerned with both generalizability to a wider population, *and* with details of individual-level effects for the sampled clusters. Some may have such interests—but considering the role of DIC as a model selection criterion, for it to be meaningful, a prioritization must be made. We have observed that p_D can depend on the choice of focal stochastic objects, and that the θ and (θ, γ) foci can yield different results. A fixed-effects structure replacing random effects gives results that differ from either of these, since the parameters of the random effects distribution are not of focal interest when those effects are considered as fixed. What we are left with, then, is the choice between three possible DIC calculations. A scientist whose interests are only out-of-sample generalizability should use DIC_m. A scientist whose interests are only on the k sampled clusters should use the fixed effects model and its associated DIC. We believe that the third option, the DIC_j construct, is never preferable to these. It depends on the ratio of random effects and error variances, and its meaning in the mixed model setting remains unclear. Certainly, it does not appear possible to argue that DIC_j represents a principled reweighting of DIC_m and the fixed effects DIC that will always reflect the inferential priorities of the user.

3.3.2.2 Automated model selection requires carefully chosen tools

The issue of focal choice is further complicated by the increasing reliance on automated model selection procedures. As a thought example, consider how Google places advertisements on websites. The following information is condensed and summarized from Google's AdSense Help Center (Help 2017).

When a website using Google AdSense has adspace for sale, computers at Google classify the website according to "factors [such] as keyword analysis, word frequency, font size, and the overall link structure of the web." Then the Google systems search a database of advertisers and select those whose ads are deemed relevant to the content or users of the website. Google creates an automated auction where advertisers can bid on the available adspace in units of cost-per-click (CPC), which is how much the advertiser is willing to pay the website owner for each click their advertisement receives. Google combines CPC bids with a quality score—a measure of how likely an ad is to be clicked based on its past performance and how well its content matches the website—to decide which advertiser wins an auction. Further, Google estimates how likely it is that an ad click will lead to a business transaction for its advertisers, and dynamically reduces some advertiser bids. The rationale behind this practice is that it protects advertisers from overspending on advertisements that are unlikely to result in business transactions, and allows advertisers to bid more freely in the auctions.

A statistician will recognize many areas in this process where covariate modeling and model selection procedures are relevant. Which website factors will best predict click-through rate (CTR) for a particular ad or class of ads? How much should an advertiser bid in a certain situation if that advertiser wants its ads to be seen? Which ad and advertiser characteristics best predict that ad clicks will result in business transactions? Because of the speed and frequency necessary for these decision-making problems, however, direct supervision is difficult if not impossible. New advertisements, and new websites, enter the marketplace too quickly for individual analysts to study or classify them. Simplicity of classification will tend to result in less content-specific ad placement, reducing revenue for the website owner, the advertisers, and Google itself. Incentives are high, in this situation, to create model selection algorithms that do not need supervision. This is an example of the discipline of machine learning.

Modern statisticians must anticipate encountering situations where it is necessary to choose a principled model selection procedure that behaves in a desired fashion even without close monitoring. With the rise of "big data," it is now more important than ever that statisticians and scientists have a clear understanding of their model selection tools—especially how those tools may give different results than other model selection tools, and which tool selects models that are preferred for a given application. Even in relatively simple settings, the various model selection criteria discussed above can lead to considerably different model choices. R. R. Christensen (2017) has shown that when comparing two nested linear models, selection by Adjusted R^2 is equivalent to selecting the larger model when the F statistic is greater than 1, selection by C_p corresponds to $F > 2$, selection by AIC is asymptotically equivalent to $F > 2$, and selection by BIC corresponds asymptotically to $F > \log n$. For more complicated settings like LMMs and GLMMs, good understanding of available criteria is even more important. We consider this another reason why the *marginal* DIC calculations should be preferred to other DIC calculation methods. Marginal DIC calculations are more easily understood, because the theory surrounding them is relatively straightforward compared to the broader theory surrounding DIC calculations for hierarchical models (c.f. Celeux et al. 2006). The asymptotic equivalence we showed between DIC and AIC in Section 3.1.7 fails when random effects are included in the model.

3.3.2.3 Interpreting p_D in hierarchical models

The problem of using DIC_j for model selection in hierarchical models has received considerable attention, especially as it relates to p_D, and has already been discussed by us in Section 3.2.2. Brooks (2002) explains that "[s]adly, in many cases the calculation of p_D will be impossible for the focus of primary interest since the deviance will not be available in closed [form]," including in random effects and state-space models. To elucidate the behavior of p_D, Sahu (2002) provides a simpler version of our own example in Section 3.2.2 to discuss

the fact that $p_D \to \infty$ as $k \to \infty$ under the DIC_j construct. Although not related directly to the DIC, Su and Johnson (2006) provide contributions explaining the asymptotic behavior of random effects models with respect to the roles of large n and large k.

Celeux et al. (2006) provide a comprehensive review of a number of DIC constructions and associated issues. As we mentioned above, our DIC_m corresponds directly with their DIC_1, and our DIC_j roughly with their DIC_7. Celeux et al. are concerned with cases such as mixture models where $\mathrm{E}[\theta \mid y, \gamma]$ may result in poor performance of DIC leading to a negative p_D. They show that in certain problems, using the maximum *a posteriori* (MAP) estimates for θ and γ can result in better behavior than the posterior mean. Celeux et al. also concur with our assessment that constructions of the DIC_j form, those that treat the latent random variable γ like a parameter vector, are unsatisfactory. They state that "this approach has obvious asymptotic and coherency difficulties, as discussed in previous literature" and "in the random effect model... computing the p_D's and therefore the DIC's does not really make sense."

In this section, we have argued that when random effects are needed, DIC_m is the sensible construct to consider because it correctly treats the random effects as "nuisance". We have argued that understanding the behavior of model selection criteria is especially important in situations where model selection must be automated, and that we should avoid criteria whose behavior is difficult to understand. And we have discussed the concerns other researchers have expressed with the behavior of p_D in hierarchical models. Neither DIC_j nor—based on their reported numerical results—any of the other constructions considered by Celeux et al. (2006) provide estimates of the number of parameters in an hierarchical model that match the number we would expect from a marginalized model.

3.3.3 Arguments Against Marginalization

We have made our case for why we believe the marginal definition for DIC is to be preferred. We recognize, however, that our position is not universally held. Below, we discuss two critiques of the marginal preference that we have encountered.

3.3.3.1 Criterion instability with variance components

We have been made aware of inconsistencies in criterion behavior when selecting among models with different variance component structures. Specifically, Dr. Daniel Gillen of the University of California, Irvine, has mentioned that information criteria can exhibit a "skipping" behavior when variance components are added to or removed from a model. We believe this may be analogous to an effect we have previously observed in our own work with Dr. Gillen, which involved in part the simulation of a linear mixed effects models with a LASSO penalty. The simulation behavior of LASSO models is often evaluated in terms of out-of-sample prediction error as the LASSO penalty, λ, varies. In our work with Dr. Gillen, we observed that in mixed effects models, the prediction error for five-fold cross-validation as a function of λ was not a continuous function for linear mixed models; it generally does appear as a continuous function for fixed effect models. Figure 3.1, taken from this earlier simulation work, displays the skipping behavior we describe to help the reader envision the phenomenon.

This skipping behavior occurs when the LASSO adds or removes a covariate. When a covariate is added or removed, assuming this covariate relates to the response variable, the random effect appears to lose or gain (respectively) variability to account for the change in the fixed effects model. This assumes, of course, that the random effects are also related to the covariates.

Figure 3.1: "Skipping" behavior in a simulation study of LASSO use for linear mixed models. As the LASSO penalty (λ) increases, five-fold cross-validation prediction error makes distinct jumps at certain lambda values.

Dr. Gillen reports that similar phenomena can be observed when information criteria are used for model selection purposes and variance components are allowed to be added and removed as in the marginal selection setting we describe. Our examples in this dissertation all presuppose a cluster-level random effect and do not appear to be subject to these issues; but studies that involve multi-level clustering (c.f. the cow abortion data of Thurmond et al. (2005) that we will describe in Section 6.2.3) may require us to choose which clusters are and are not modeled with random effects distributions. Above, we advocated for the use of selection criteria whose behavior is well-understood and consistent, especially in automated selection settings. We continue to advocate this policy here as well, and believe that further investigation of the behavior of marginal selection criteria like DIC_m is warranted when selecting among variance components.

3.3.3.2 Inferential priorities

Some statisticians suggest that model selection for mixed models should take account of both the random effects for the observed clusters and the variance components for those random effects. They suggest this because scientific interest can reside in both areas simultaneously: how the model functions for new observations in the sampled clusters, as well as how it functions for new observations on new clusters. In the Bayesian setting, fixed effects and random effects have very similar specifications within a probability model; the primary difference between them is how they are handled in inference, once a posterior sample has been obtained. We agree that both the conditional effects on a response when cluster is known, and the marginal effects on a response are legitimate areas of scientific interest, but as discussed above we find it difficult to carefully define how model selection should proceed when both conditional and marginal inferences are desired.

Nonetheless, the argument has been made to us that, following from the example in Section 3.2.2 one should reasonably want to penalize a model as if it has k fixed effects if data are

sufficiently different from cluster to cluster; or that one should penalize a model as if it had only a grand mean if data are sufficiently homogenous across clusters. "The behavior of p_{Dj} is not a bug, it's a feature," one might say. This is a view we have encountered with some frequency, though we remain troubled by the fact that this argument presumes that the appropriate penalization of a model is in some sense dependent on the number of clusters one happens to select, even when inference for those clusters is not itself desired.

Other statisticians suggest that DIC may not be the preferred tool for situations such as the ones we describe. Gelman et al. (2013) place DIC in a hierarchy with AIC and BIC where they suggest that DIC should be preferred when inference is desired for individuals within the sampled clusters, AIC should be preferred when inference is desired for unsampled clusters of a similar character, and BIC should be preferred when inference is desired marginally on the population. This suggests that they consider DIC less useful for model selection relative to marginal population-level models, though we believe the marginalization techniques we develop in this dissertation broaden the scope of situations in which DIC may be usefully applied.

3.4 Marginalization in the Linear Mixed Model

Our arguments in the preceding section lead to the question of why marginalization is not performed more often when selecting a model. One answer is that marginalization is difficult, especially in the GLMM setting where closed-form marginal equations do not exist. Marginalization is both possible and practical in the LMM setting, however, and so we begin by explaining two marginalization approaches for the DIC. The methods explained here, particularly our approach to postprocessing marginalization, point the way toward the methods we develop in the next two chapters for marginalization of GLMMs.

3.4.1 Methods for Marginalization

In the linear mixed model, when a normal distribution is used for the random effects, there are two methods for getting the marginal likelihood $L(\theta \mid y)$ and thus the marginal DIC calculations. The first, which we call the preprocessing approach, involves expressing the model directly in its marginalized form. MCMC sampling directly using the marginal model will, obviously, yield the desired marginal DIC calculations. The second approach, which we refer to as postprocessing marginalization, uses the complete-the-square formula (Proposition 3.1, below) after re-expressing $f(y, \gamma \mid \theta)$ as $f(\gamma \mid y, \theta) f(y \mid \theta)$ when both $\gamma \mid \theta$ and $y \mid \gamma, \theta$ have multivariate normal distributions.

3.4.1.1 Preprocessing marginalization

In this section we discuss MCMC sampling that is based on using the marginal likelihood for θ, $L(\theta \mid y) \propto \int f(y \mid \gamma, \theta) P(\gamma \mid \theta) d\gamma$. When the marginal likelihood is available in closed form, as is the case for the LMM discussed above, it is relatively straightforward to implement MCMC sampling to obtain iterates $\{\theta^{(1)}, ..., \theta^{(B)}\}$ from the posterior, $P(\theta \mid y)$, with the help of packaged software like OpenBUGS, JAGS, STAN etc. This will involve monitoring $D(\theta) = -2\log(L(\theta \mid y))$ in one of these packages, to obtain $\overline{D(\theta)}$. Then, we use $\hat{\theta} = \frac{1}{B}\sum_{s=1}^{B} \theta^{(s)}$ to numerically approximate p_{Dm} and DIC_m with

$$p_{Dm} \simeq -\frac{2}{B}\sum_{s=1}^{B} \log f\left(Y \mid \theta^{(s)}\right) + 2\log f\left(Y \mid \hat{\theta}\right)$$
$$DIC_m \simeq -\frac{4}{B}\sum_{s=1}^{B} \log f\left(Y \mid \theta^{(s)}\right) + 2\log f\left(Y \mid \hat{\theta}\right)$$

We now proceed to analytically obtain the marginal likelihood.

Refer to Equation (3.9), the matrix specification for the linear mixed model. We can rewrite this model as $\mathbf{Y} - \mathbf{X}\beta = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$. Recall that $\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma})$ and $\mathbf{e} \sim \mathrm{N}_{kn}(0_{kn}, \sigma^2 I_{kn})$. Then we

can write the marginal for the data as

$$\mathbf{Y} \sim \mathrm{N}_{kn}\Big(\mathbf{X}\beta + \mathbf{Z}\boldsymbol{\psi}, \mathbf{Z}^T\boldsymbol{\Sigma}\mathbf{Z} + \sigma^2 I_{kn}\Big). \tag{3.10}$$

Since the matrix Σ_i is often relatively uncomplicated—in the case of a random intercepts model, it is the scalar variance of the random intercepts—it is thus easy to write the linear mixed model in terms of its induced marginal mean vector and covariance matrix, ignoring $\boldsymbol{\gamma}$ entirely.

If this approach is used, we should be cognizant of how MCMC sampling efficiency is affected. MCMC sampling for Bayesian models can be improved by including intermediate stochastic nodes like $\boldsymbol{\gamma}$, so avoiding them as we do in the preprocessing approach may lead to sampling behaviors we dislike. Preprocessing ensures that the numerical DIC approximations obtained from software like OpenBUGS are approximations to the desired DIC_m construct, but we must weigh this against the potential loss of sampling efficiency under this approach.

3.4.1.2 Postprocessing marginalization

Here we begin with the full description of the model that involves γ. We sample from this model involving γ, but then obtain an analytical form for the marginal into which we can plug our posterior iterates to obtain our own numerical approximation to DIC_m.

We start by writing the joint density for the data and γ conditional on θ, $f(y, \gamma \mid \theta)$, and then through a series of algebraic manipulations, we obtain an equivalent expression, namely $f(y, \gamma \mid \theta) = f(y \mid \theta) f(\gamma \mid y, \theta)$, where the conditional distribution in γ is normal with parameters depending on θ. Thus upon integrating over γ, we obtain an analytical expression for $f(y \mid \theta)$. Thus given a MC sample from the posterior for θ, which is easily obtained using BUGS or some other package, we are able to numerically approximate the marginal model based DIC.

Below, we develop a new expression for the marginal density $f(y \mid \theta)$. This is not necessary for the LMM setting—it should be clear that the marginal form in Equation (3.10) will serve this purpose, and in fact must be equivalent to the expression we develop below. The work we present here is crucial to subsequent work in our development of a marginalization approach for GLMMs during the next two chapters. In the GLMM setting, no closed-form marginalization exists and we are unable to use a preprocessing approach. We thus consider it preferable to introduce this work here where there are no complications.

Our alternate expression for the marginal density takes advantage of the fact that both $f(y \mid \gamma, \theta)$ and $P(\gamma \mid \theta)$ are normal densities. Mathematically, we can use the complete-the-square formula to combine the two and isolate the γ terms.

Following from the notation in Section 3.2.1, we assume $\mathbf{X}$ and $\mathbf{Z}$ are full rank and write the following:

$$
\begin{aligned}
f(y_{ij} \mid \gamma_i, \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_{ij} - (x_{ij}\beta + z_{ij}\gamma_i)}{\sigma}\right)^2\right) \\
f(Y_i \mid \gamma_i, \theta) &= (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2}\left[Y_i - (X_i\beta + Z_i\gamma_i)\right]^T \left(\sigma^2 I_n\right)^{-1} \left[Y_i - (X_i\beta + Z_i\gamma_i)\right]\right) \\
f(\gamma_i \mid \theta) &= (2\pi)^{-q/2} |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}\left[\gamma_i - \psi_i\right]^T \Sigma_i^{-1} \left[\gamma_i - \psi_i\right]\right)
\end{aligned}
$$

This gives the joint density

$$
\begin{aligned}
& f(Y_i, \gamma_i \mid \theta) \\
&\quad = (2\pi)^{-(k+q)/2} \sigma^{-k} |\Sigma_i|^{-1/2} \\
&\qquad \times \exp\left(-\frac{1}{2}\left[Y_i - (X_i\beta + Z_i\gamma_i)\right]^T \left(\sigma^2 I_n\right)^{-1} \left[Y_i - (X_i\beta + Z_i\gamma_i)\right]\right) \\
&\qquad \times \exp\left(-\frac{1}{2}\left[\gamma_i - \psi_i\right]^T \Sigma_i^{-1} \left[\gamma_i - \psi_i\right]\right).
\end{aligned}
$$

To simplify notation, define $\tilde{Y}_i = Y_i - X_i\beta$. This allows us to write

$$\begin{aligned}
&\left[\tilde{Y}_i - Z_i\gamma_i\right]^T \left(\sigma^2 I_n\right)^{-1} \left[\tilde{Y}_i - Z_i\gamma_i\right] \\
&= \left[\tilde{Y}_i - Z_i\hat{\gamma}_i + Z_i\hat{\gamma}_i - Z_i\gamma_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i + Z_i\hat{\gamma}_i - Z_i\gamma_i\right] \\
&= \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right] + \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right] \\
&\qquad + \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right] + \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right] \\
&= \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right] + \left[\hat{\gamma}_i - \gamma_i\right]^T \left(\frac{1}{\sigma^2} Z_i^T Z_i\right) \left[\hat{\gamma}_i - \gamma_i\right].
\end{aligned}$$

Then apply the "complete-the-square" formula below, which is proven in the appendix, to combine the quadratic terms for γ_i in the exponent. This results in a Normal kernel for γ_i and a second term that is free of γ_i, making it possible for us to easily marginalize.

Proposition 3.1. *For conformable vectors X, μ_1, and μ_2; and for conformable symmetric matrices A_1 and A_2;*

$$\begin{aligned}
&(X-\mu_1)^T A_1 (X-\mu_1) + (X-\mu_2)^T A_2 (X-\mu_2) \\
&\quad = (X-\mu^*)^T (A_1+A_2)(X-\mu^*) + (\mu_1-\mu_2)^T A_1 (A_1+A_2)^{-1} A_2 (\mu_1-\mu_2),
\end{aligned}$$

where $\mu^* = (A_1+A_2)^{-1}(A_1\mu_1 + A_2\mu_2)$.

In our context for the linear mixed model, we substitute γ_i for X above. We take $\mu_1 = \psi_i$ and $A_1 = \Sigma_i^{-1}$ for the first quadratic portion. For the second, we choose $\mu_2 = \hat{\gamma}_i = (Z_i^T Z_i)^{-1} Z_i^T (Y_i - X_i\beta)$ and $A_2 = \sigma^{-2} Z_i^T Z_i$. Using the complete-the-square formula, we have

$$\begin{aligned}
&(\gamma_i - \psi_i)^T \Sigma_i^{-1} (\gamma_i - \psi_i) + \frac{1}{\sigma^2} (\gamma_i - \hat{\gamma}_i)^T Z_i^T Z_i (\gamma_i - \hat{\gamma}_i) \\
&\quad = (\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right) (\gamma_i - \gamma_i^*) \\
&\qquad + \frac{1}{\sigma^2} (\psi_i - \hat{\gamma}_i)^T \Sigma_i^{-1} \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} Z_i^T Z_i (\psi_i - \hat{\gamma}_i),
\end{aligned}$$

where

$$\gamma_i^* = \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} \left(\Sigma_i \psi_i + \frac{1}{\sigma^2} Z_i^T (Y_i - X_i \beta)\right).$$

We use this new expression to rewrite the joint density for Y_i and γ_i.

$$\begin{aligned}
f(Y_i, \gamma_i \mid \theta) &= (2\pi)^{-(n+q)/2} \sigma^{-n} |\Sigma_i|^{-1/2} \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)^T \left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}(\psi_i - \hat{\gamma}_i)^T \Sigma_i^{-1} \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} Z_i^T Z_i (\psi_i - \hat{\gamma}_i)\right) \\
&\quad \times \exp\left(-\frac{1}{2}(\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)(\gamma_i - \gamma_i^*)\right)
\end{aligned}$$

Note that γ only appears in the final term, which has the form of a Normal kernel. This allows us to rewrite the joint density as follows:

$$\begin{aligned}
f(Y_i, \gamma_i \mid \theta) &= (2\pi\sigma^2)^{-n/2} |\Sigma_i|^{-1/2} \left|\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right|^{-1/2} \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)^T \left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}(\psi_i - \hat{\gamma}_i)^T \Sigma_i^{-1} \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} Z_i^T Z_i (\psi_i - \hat{\gamma}_i)\right) \\
&\quad \times (2\pi)^{-q/2} \left|\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right|^{1/2} \exp\left(-\frac{1}{2}(\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)(\gamma_i - \gamma_i^*)\right) \\
&= f(Y_i \mid \theta) \times f(\gamma_i \mid Y_i, \theta)
\end{aligned}$$

Since observations on different clusters are assumed to be conditionally independent, the marginal for the entire data set is just $\prod_{i=1}^k f(Y_i \mid \theta)$, and we thus obtain a numerical approx-

imation to DIC_m using the marginal density:

$$
\begin{aligned}
f(Y \mid \theta) = (2\pi\sigma^2)^{-kn/2}|\Sigma|^{-1/2} \left|\Sigma^{-1} + \frac{1}{\sigma^2}\mathbf{Z}^T\mathbf{Z}\right|^{-1/2} \\
\times \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\right)^T\left(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\right)\right) \\
\times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\psi} - \hat{\boldsymbol{\gamma}})^T\Sigma^{-1}\left(\Sigma^{-1} + \frac{1}{\sigma^2}Z^TZ\right)^{-1}Z^TZ(\boldsymbol{\psi} - \hat{\boldsymbol{\gamma}})\right),
\end{aligned}
$$

where $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta)$.

We remind the reader that, given a sample from the joint posterior, $P(\theta, \gamma \mid y)$, the iterates for θ are from the marginal posterior $P(\theta \mid y)$, say $\{\theta^{(s)} : s = 1, 2, ...B\}$. Then

$$
\begin{aligned}
p_{Dm} &\doteq -\frac{2}{B}\sum_{s=1}^{B}\log f\left(\mathbf{Y} \mid \theta^{(s)}\right) + 2\log f\left(\mathbf{Y} \mid \hat{\theta}\right) \\
DIC_m &\doteq -\frac{4}{B}\sum_{s=1}^{B}\log f\left(\mathbf{Y} \mid \theta^{(s)}\right) + 2\log f\left(\mathbf{Y} \mid \hat{\theta}\right)
\end{aligned}
\tag{3.11}
$$

The MCMC sample can be obtained using OpenBUGS, JAGS, STAN or any other available sampling software.

In the next chapter, we will consider a special case of the generalized linear mixed model (GLMM), where closed-form marginalization is not possible, but where the expression we have here derived points the way toward a new method for approximate marginalization.