Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1

A Conservation Law for Posterior Predictive Variance

Dean Dustin[*], Sanjay Chaudhuri[†], Bertrand Clarke[†]

Abstract. We use the law of total variance to generate multiple expressions for the posterior predictive variance in Bayesian hierarchical models. These expressions are sums of terms involving conditional expectations and conditional variances. Since the posterior predictive variance is fixed given the hierarchical model, it represents a constant quantity that is conserved over the various expressions for it. The terms in the expressions can be assessed in absolute or relative terms to understand the main contributors to the length of prediction intervals. Also, sometimes these terms can be interpreted in the context of the hierarchical model. We show several examples, closed form and computational, to illustrate this approach to predictive model assessment.

MSC2020 subject classifications: Primary 62F15; secondary 62J10.

Keywords: prediction interval, posterior predictive variance, law of total variance, Bayes model averaging.

1 The Setting and Intuition

Consider a generic Bayesian hierarchical model (BHM) for a response $Y = y$ given $V = (V_1, \ldots, V_k, \ldots, V_K)^T$ taking values $v = (v_1, \ldots, v_K)^T$ for some $K \in \mathbb{N}$:

$$\begin{array}{ccl} V_1 & \sim & w(v_1) \\ V_2 & \sim & w(v_2|v_1) \\ \vdots & \vdots & \vdots \\ V_K & \sim & w(v_K|v_1, \ldots, v_{K-1}) \\ Y & \sim & p(y|v), \end{array} \tag{1.1}$$

where the w's represent prior densities for the V_k's as indicated by their arguments and $p(\cdot|v)$ is the likelihood. All densities are with respect to Lebesgue measure when the random variable is continuous and with respect to counting measure when the random variable is discrete.. For discrete random variables we regard the density as being taken with respect to counting measure. We denote n outcomes of Y by $Y^n = (Y_1, \ldots, Y_n)^T$ with outcomes $y^n = (y_1, \ldots, y_n)^T$.

It is common practice to adopt an estimation perspective. That is, choose a parameter, here one of the V_k's, and obtain credibility sets for it from the posterior $w(v_k|y^n)$. If the credibility set for a given V_k is sufficiently small as determined by hypothesis

[*] First Citizens' Bank, Raleigh, NC, USA deandust55@gmail.com

[†] Department of Statistics, University of Nebraska-Lincoln, NE, USA, 68583-0963 schaudhuri2@unl.edu, bclarke3@unl.edu

testing, say, then it may make sense to drop the k-th level of the hierarchy. However, it is unclear in the abstract how to compare the length of a credibility set for one V_k to the length of a credibility set for $V_{k'}$ for $k \neq k'$. Aside from asymptotics usually based on the Fisher information, there is no common scale on which the variances of different V_k's can be compared. The reason is that the size of $\mathrm{Var}(V_k|y^n)$ is unrelated to the size of $\mathrm{Var}(V_{k'}|y^n)$. Nothing necessarily ties the K marginal posteriors $w(v_k|y^n)$ together with a common scale pre-asymptotically. Indeed, when estimating a value of v_k, it is not in general clear how the sizes of other $v_{k'}$'s affect it. Moreover, while we might try to derive a likelihood for a posterior variance so as to do a hypothesis test, this is difficult to do and would be hard to interpret.

An alternative analysis of hierarchical models follows from a predictive perspective. Instead of looking at posterior variances, we look at terms that sum to the posterior predictive variance and compare their relative importance. This way all variances are on the same scale. Without further discussion, we assume that posterior means and variances in general are the right quantities to study. This is true under squared error loss; other choices of loss function would yield different, but analogous, reasoning.

Given y^n, we assign the posterior predictive density to future values Y_{n+1}, that is

$$Y \sim p(y_{n+1}|y^n) = \int p(y_{n+1}|v) w(v|y^n) \mathrm{d}v, \tag{1.2}$$

where $w(v|y^n)$ is the posterior density and $\mathrm{d}v$ is summation or integration as appropriate. At this point the posterior predictive variance within the context of the model (1.1) is fixed. Denote it $\mathrm{Var}(Y_{n+1}|y^n)$. When a random variable in the top K levels of the hierarchy is visible we say it is explicit Otherwise we say it is implicit. Thus, $\mathrm{Var}(Y_{n+1}|y^n)$ depends implicitly on the top K levels of (1.1).

Recall the standard probability theory result called the Law of Total Variance (LTV). Generically, for random variables W and Z on the same probability space it is

$$\mathrm{Var}(W) = E[\mathrm{Var}(W|Z)] + \mathrm{Var}[E(W|Z)]. \tag{1.3}$$

By reinterpreting (1.3) in the posterior context we have

$$\mathrm{Var}(W|y^n) = \mathrm{Var}_{W|y^n}(W|y^n) = E_{Z|y^n}[\mathrm{Var}(W|Z, y^n)] + \mathrm{Var}_{Z|y^n}[E(W|Z, y^n)], \tag{1.4}$$

assuming that W and Z are functions on the same probability space as used to write (1.1). In (1.4) the densities used for the expectations on the right, usually suppressed, are indicated and both sides are functions of y^n.

The predictive approach takes W to be a future value Y_{n+1}, rather than any of the v_k's. For generality, we will also take y^n to be the pre-$n+1$ data, and hence condition on $\mathcal{D} = \mathcal{D}_n$. In contrast to y^n, $\mathcal{D}$ may include values of explanatory variables for each time step. Unless stated otherwise, we assume the data is independent from time step to time step. Now we have

$$\mathrm{Var}_{Y_{n+1}|\mathcal{D}}(Y_{n+1}|\mathcal{D}) = E_{Z|\mathcal{D}}[\mathrm{Var}(Y_{n+1}|Z, \mathcal{D})] + \mathrm{Var}_{Z|\mathcal{D}}[E(Y_{n+1}|Z, \mathcal{D})]. \tag{1.5}$$

Independent of the choice of Z, the left hand side of (1.5) is a constant depending only on the hierarchy (1.1) and $\mathcal{D}$. That is, (1.5) is a conservation law for the posterior predictive variance over choices of conditioning. We can choose Z to be any function of a subset of the entries of V. In particular, if $Z = V_1$, we get

$$\mathrm{Var}_{Y_{n+1}|\mathcal{D}}(Y_{n+1}|\mathcal{D}) = E_{V_1|\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_1,\mathcal{D})] + \mathrm{Var}_{V_1|\mathcal{D}}[E(Y_{n+1}|V_1,\mathcal{D})]. \qquad (1.6)$$

More is true. The LTV can be applied iteratively to either term in (1.5). Indeed, it is seen that $\mathrm{Var}(Y_{n+1}|Z,\mathcal{D})$, the first term on the right in (1.5), is of the same form as the left hand side of (1.5) – simply replace $\mathcal{D}$ by $(Z,\mathcal{D})$. If we take $Z = V_1$, condition on $V_1 = v_1$, and use another instance of the LTV, this time with $Z = V_2$, we get

$$\mathrm{Var}(Y_{n+1}|v_1,\mathcal{D}) = E_{V_2|v_1,\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_1,V_2,\mathcal{D})] + \mathrm{Var}_{V_2|v_1,\mathcal{D}}[E(Y_{n+1}|V_1,V_2,\mathcal{D})]. \quad (1.7)$$

Using (1.7) in (1.5) we get, with some simplification of notation,

$$\begin{aligned}\mathrm{Var}(Y_{n+1}|\mathcal{D}) \quad = \quad & E_{V_1|\mathcal{D}}E_{V_2|V_1,\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_1,V_2,\mathcal{D})] \\ & +E_{V_1|\mathcal{D}}\mathrm{Var}_{V_2|V_1,\mathcal{D}}[E(Y_{n+1}|V_1,V_2,\mathcal{D})] \\ & +\mathrm{Var}_{V_1|\mathcal{D}}[E(Y_{n+1}|V_1,\mathcal{D})]. \qquad (1.8)\end{aligned}$$

Now, (1.6) and (1.8) are two expressions for the same $\mathrm{Var}(Y_{n+1}|\mathcal{D})$. They are generic in that the role of V_1 and V_2 can be played by any two functions of entries of V. That is, the posterior predictive variance admits a very large number of two term and three term generic expressions.

This procedure can be iterated in multiple ways to include any other V_k, thereby generating even more expressions for $\mathrm{Var}(Y_{n+1}|\mathcal{D})$. Indeed, every time an expression of the form $\mathrm{Var}(Y_{n+1}|W,\mathcal{D})$ for any suitable random variable W occurs from using the LTV, the LTV can be applied again provided a further suitable conditioning variable Z can be found. That is, the conservation law for posterior predictive variance in (1.6) extends to a far larger class of sums of terms involving conditional expectations and variances than (1.5) initially suggests.

Here have used the LTV only on the first term, the ‘E var’ term. The LTV can be used on the ‘Var E’ terms as well. However, we want to retain the last ‘Var E’ term in (1.8) because when it is small it may be a good reason to drop V_1 from the BHM.

We call the collection of expressions for the posterior predictive variance in a fixed hierarchical model its LTV-scope. Thus, the posterior predictive variance is invariant or conserved over its scope. We regard the introduction of an extra level in a hierarchical model as creating a new model and hence a new LTV-scope. The point of this work is not only to look within the LTV-scope of one hierarchical model but to compare LTV-scopes across models. Expressions in the scope of a hierarchical model also admit an interpretation in terms of analysis of variance and associated frequentist testing, see [5], but we do not discuss this here. Fixing a hierarchical model and looking at the scope of its posterior predictive variance lets us choose which decomposition has the interpretation we want to use to decide which components of the BHM are more important than others, in relative or absolute terms. Otherwise put, we can examine and compare multiple decompositions of the posterior variance for the same hierarchical model and then compare decompositions across hierarchical models because it is fair to compare posterior predictive variances across models.

Expressions like (1.8) may be useful in a practical sense as well because predictive intervals (PI's) for Y_{n+1} can be derived from the distribution of

$$\frac{Y_{n+1} - E(Y_{n+1}|\mathcal{D}_n)}{\sqrt{Var(Y_{n+1} - E(Y_{n+1}|\mathcal{D}_n)|\mathcal{D}_n)}} = \frac{Y_{n+1} - E(Y_{n+1}|\mathcal{D}_n)}{\sqrt{\mathrm{Var}(Y_{n+1}|\mathcal{D}_n)}}. \tag{1.9}$$

The denominator on the right in (1.9) is the posterior predictive variance and controls the length of the PI. Our expressions for it allow us to identify the relative sizes of their terms. That is, because the posterior predictive variance ties multiple sources of variability together within a hierarchical model, we can look at relative contributions of terms to the PI. For instance, it is meaningful to compare the sizes of terms such as

$$\frac{E_{V_1|\mathcal{D}}E_{V_2|V_1,\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_1,V_2,\mathcal{D})]}{\mathrm{Var}(Y_{n+1}|\mathcal{D})} \quad \text{and} \quad \frac{E_{V_1|\mathcal{D}}\mathrm{Var}_{V_2|V_1,\mathcal{D}}[E(Y_{n+1}|V_1,V_2,\mathcal{D})]}{\mathrm{Var}(Y_{n+1}|\mathcal{D})}.$$

A relative assessment of their contributions to the posterior predictive variance allows us to identify the biggest contributions to the length of a PI. Terms that do not contribute much, relatively, can be omitted thereby identifying which terms are driving the width of PI's. We see an instance of this in an example in Sec. 5.

This decomposition is similar to [7] who expanded the posterior variance $\mathrm{Var}(\Theta|y^n)$. However, ours is predictive, on a common scale, and hence directly useful in expressions for PI's from, say, (1.9). Moreover, in [7], the terms were forced into a single 'standard error' interpretation rather than treated as distinct patterns of expectations and variances that could be interpreted in the context of quantifying the variability in the levels of the hierarchy. In short, we get a complete uncertainty quantification.

Another way these decompositions may be useful is in terms of reducing the number of levels in the hierarchy. Consider the last term in (1.8). There are two basic ways we can get $\mathrm{Var}_{V_1|\mathcal{D}_n}(E(Y_{n+1}|V_1,\mathcal{D}_n)) = 0$. First, the distribution of V_1 concentrates at a single value $V_1 = v_1$. Second, the models i.e., values of V_1 that get non-zero weights, give the same predictions given $\mathcal{D}$. That is,

$$E(Y_{n+1}; V_1 = v_1, \mathcal{D}) = E(Y_{n+1}; V_1 = v_2, \mathcal{D}) \tag{1.10}$$

for any v_1 and v_2 getting positive weight. Identifying these sets is essentially intractable. However, by carefully selecting the models $V = v$ to ensure they are different and having a large enough n the chance of satisfying (1.10) for two values of V_1 will be vanishingly small. Thus, on pragmatic grounds, with some foresight, if the last term on the right is chosen so it explicitly depends only on a single component of V and that term drops out i.e., is close to zero, we can simply set V_1 to be a constant meaning that level of modeling drops out. In a three term case we would be left with only the first two terms on the right hand side that depend on V_2 in which V_1 was a constant. The resulting expression reduces to (1.6) but with V_2 in place of V_1.

The rest of this paper studies expressions such as (1.6) and (1.8). In Sec. 2, we present parametric examples of instances of our conservation of posterior predictive variance law. In Sec. 3 we treat Bayesian model averages as BHM's and see what our decompositions look like. In Sec. 4 we present some generic results on expressions in the LTV-scope of a posterior predictive variance. In Sec. 5, we revisit two examples from [4] and show how the terms in our decomposition behave for a two way ANOVA. In a concluding section, Sec. 6, we discuss the methodological implications of our representations for the posterior predictive variance. Details that are necessary but ancillary to our main points are relegated to Appendices.

2 Parametric Examples

The point of this section is to present the parametric case. In particular, we see two normal examples that are amenable to our LTV iterative procedure.

For the remainder of this paper we note that our reasoning only requires a generic hierarchical model. There is no constraint on the levels in the hierarchy except that the whole inferential structure satisfies the containment principle of Bayesian statistics, i.e., the entire model is contained in one explicit probability space. Even with this constraint, the range of choices for (K, V) is vast and two plausible hierarchical models may have very different behaviors. In addition, as will be seen in Sec. 5, conditioning variables need not have any correlate in reality; they may be aspects of modeling more commonly thought to be part of the likelihood.

2.1 Two Level Hierarchical Models

The simplest hierarchical model has two levels i.e., has $K = 1$:

$$\begin{aligned} \Theta &\sim w(\theta) \\ Y &\sim p(y|\theta), \end{aligned} \tag{2.1}$$

where w is the density of a real parameter $\Theta = \theta$ and $p(\cdot|\theta)$ is the conditional density of $Y = y$, both with respect to Lebesgue measure. The posterior density is

$$w(\theta|y^n) \propto w(\theta)p(y^n|\theta)$$

with normalizing constant

$$m(y^n) = \int w(\theta)p(y^n|\theta)\mathrm{d}\theta. \tag{2.2}$$

The posterior predictive density is now

$$p(y_{n+1}|y^n) = \int p(y_{n+1}|\theta)w(\theta|y^n)\mathrm{d}\theta$$

with mean

$$E(Y_{n+1}|y^n) = \int y_{n+1}p(y_{n+1}|y^n)\mathrm{d}y_{n+1}$$

and

$$\mathrm{Var}(Y_{n+1}|y^n) = \int (y_{n+1} - E(Y_{n+1}|y^n))^2 p(y_{n+1}|y^n)\mathrm{d}y_{n+1}.$$

So, the LTV gives the posterior predictive variance as

$$\mathrm{Var}(Y_{n+1}|y^n) = E_\Theta(\mathrm{Var}(Y_{n+1}|\Theta, y^n)|y^n) + \mathrm{Var}_\Theta(E(Y_{n+1}|\Theta, y^n)|y^n). \tag{2.3}$$

The first term on the right is the variability of the high posterior probability predictive distributions. The second term on the right is an assessment of how important the model used for prediction is. This interpretation is, in fact, independent of the fact that Θ is a real parameter. Only two-term examples i.e., one usage of the LTV, admit a concise interpretation in general. We will see that with two or more usages of the LTV and so $K \geq 2$, we get three or more terms and the interpretation is much more complex and depends delicately on the choice of conditioning variables.

In some cases, (2.3) can be worked out explicitly. Let $Y_i \sim N(\theta, \sigma)$ be independent and identically distributed (IID) for $i = 1, \ldots, n$ where $\theta \sim N(\theta_0, \tau^2)$ and θ_0, σ and τ are known. It is easy to see that

$$p(y_{n+1}|\theta, y^n) = \frac{p(y^{n+1}, \theta))}{p(y^n, \theta)} = p(y_{n+1}|\theta),$$

where σ has been suppressed in the notation. So, it is also easy to see that

$$E(Y_{n+1}|\theta, y^n) = \theta \quad \text{and} \quad \mathrm{Var}(Y_{n+1}|\theta, y^n) = \sigma^2.$$

Since $\mathrm{Var}(Y_{n+1}|\theta, y^n)$ is a constant, its expectation under the posterior for θ is unchnaged. Thus, the first term on the right in (2.3) is

$$E_{\Theta|y^n}\mathrm{Var}(Y_{n+1}|\Theta, y^n) = \sigma^2.$$

For the second term on the right in (2.3) recall the posterior for θ given y^n is

$$w(\theta|y^n) = \frac{1}{\sqrt{2\pi\tau_n^2}} e^{-(1/2\tau_n^2)(\theta-\theta_n)^2}$$

where

$$\theta_n = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \frac{n}{\sigma^2}\left(\bar{y} + \frac{\theta_0}{\tau^2}\right) \quad \text{and} \quad \tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1}.$$

Now,

$$\mathrm{Var}_{\Theta|y^n}(E(Y_{n+1}|\Theta, y^n)) = \mathrm{Var}_{\Theta|y^n}(\Theta) = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1},$$

and (2.3) is

$$\mathrm{Var}(Y_{n+1}|y^n) = \sigma^2 + \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1} = \sigma^2 + \mathcal{O}(1/n),$$

in which the 'E-Var' term dominates asymptotically.

Another specific example that can be evaluated in closed form to obtain similar results sets $Y_i \sim N(\mu, \sigma^2)$ to be IID for $i = 1, \ldots, n$ with the usual priors $\mu \sim N(0, \sigma^2)$ and $\sigma^2 \sim InvGamma(\alpha, \beta)$ for $\alpha > 2$ and $\beta > 0$.

For contrast, note that a different LTV problem – using the Beta-Binomial – gives that the 'Var-E' term dominates; see [3], p. 168.

As early as the mid-90's, David Draper noted that the first term on the right was the main assessment of variability that authors consider but that often this was insufficient. This observation is consistent with the normal example above and holds more generally. Indeed, the first term on the right in (2.3) is

$$\begin{aligned} & \int \int \left((y_{n+1} - E(Y_{n+1}|\theta, y^n))^2 p(y_{n+1}|\theta, y^n)\right) \mathrm{d}y_{n+1} w(\theta|y^n) \mathrm{d}\theta \\ = & \int \int \left((y_{n+1} - E(Y_{n+1}|\theta))^2 p(y_{n+1}|\theta)\right) \mathrm{d}y_{n+1} w(\theta|y^n) \mathrm{d}\theta \\ = & \int \mathrm{Var}_\theta(Y_{n+1}) w(\theta|y^n) \mathrm{d}\theta \end{aligned} \tag{2.4}$$

and the second term on the right in (2.3) is

$$\begin{aligned} & \mathrm{Var}_\Theta\left(E(Y_{n+1}|\Theta)|y^n)\right) \\ = & \int \left(E(Y_{n+1}|\theta) - \int E(Y_{n+1}|\theta) w(\theta|y^n) \mathrm{d}\theta\right)^2 w(\theta|y^n) \mathrm{d}\theta \\ = & \int E^2(Y_{n+1}|\theta) w(\theta|y^n) \mathrm{d}\theta - \left(\int E(Y_{n+1}|\theta) w(\theta|y^n) \mathrm{d}\theta\right)^2 . \end{aligned} \tag{2.5}$$

When the posterior concentrates at a true value θ_0, in distribution, L^1 or a.e., as $n \to \infty$, (2.4) converges to $\mathrm{Var}_{\theta_0}(Y_{n+1})$ and (2.5) converges to zero in the same mode. So, the first term asymptotically dominates. This reasoning holds anytime the posterior concentrates as it typically does in $\mathcal{M}$-closed problems; more generally, see [1]. However, this says little about the relative sizes of the two terms in finite samples.

In the general case, the inner expressions on the right in (2.3) are $\mathrm{Var}(Y_{n+1}|\theta, y^n)$ and $E(Y_{n+1}|\theta, y^n)$ and they have different meanings. In particular, the first term is small when $\mathrm{Var}(Y_{n+1}|\theta, y^n)$ is small over the typical region of θ under the posterior and the second term is small when $E(Y_{n+1}|\theta, y^n)$, as a function of θ, changes little, again over the typical region of θ. Loosely, the difference is whether the variance is small or the mean changes little.

If we are sure that the mean changes little, i.e., $E(Y_{n+1}|\theta, y^n)$ is nearly constant over the range of θ's most likely under the posterior, then

$$\mathrm{Var}(Y_{n+1}|y^n) \approx E_\Theta(\mathrm{Var}(Y_{n+1}|\Theta, y^n)|y^n).$$

However, if we are sure that for y^n the variance is small, i.e., for the θ's most likely under the posterior we have that $\mathrm{Var}(Y_{n+1}|\theta, y^n)$ is small, then

$$\mathrm{Var}(Y_{n+1}|y^n) \approx \mathrm{Var}_\Theta(E(Y_{n+1}|\Theta, y^n)|y^n).$$

Another way to interpret (2.3) is as follows. When the 'E-Var' term is large, relative to the 'Var-E' term, there is more variability in the predictive distributions from the high posterior probability models than there is variability across models so the model doesn't matter very much; all the commonly occurring models (high posterior probability) are good. When the 'Var-E' term is large relative to the 'E-Var' term it means that the specific model used for prediction is much more important than the variability within models used for prediction.

2.2 Three Term Normal Case

Although three term expansions are often difficult to work out explicitly, for the case of the normal with unknown mean and variance we can extend the derivations from the two term case in Subsec. 2.1.

Let $Y_i \sim N(\mu, \lambda^2)$ be IID for $i = 1, \ldots, n$ and use the conjugate priors $\mu \sim N(\mu_0, 1/\kappa_0\lambda^2)$ with $\lambda^2 \sim \mathsf{Gamma}(\alpha_0, \beta_0)$. Now we have two three-term expansions depending on whether we condition on μ first or λ first. Conditioning on μ first we get

$$\begin{aligned} \mathrm{Var}(Y_{n+1}|y^n) &= E_{\lambda^2|y^n} E_{\mu|y^n,\lambda^2} \mathrm{Var}(Y_{n+1}|y^n, \mu, \lambda^2) && (2.6)\\ &\quad + E_{\lambda^2|y^n} \mathrm{Var}_{\mu|y^n,\lambda^2} E(Y_{n+1}|y^n, \mu, \lambda^2) && (2.7)\\ &\quad + \mathrm{Var}_{\lambda^2|y^n} E_{\mu|y^n,\lambda^2} E(Y_{n+1}|y^n, \mu, \lambda^2). && (2.8) \end{aligned}$$

It is easy to see that (2.8) is zero. Indeed,

$$\mathrm{Var}_{\lambda^2|y^n} E_{\mu|y^n,\lambda^2} E(Y_{n+1}|y^n, \mu, \lambda^2) = \mathrm{Var}_{\lambda^2|y^n} E_{\mu|y^n,\lambda^2}(\mu) = \mathrm{Var}_{\lambda^2|y^n}\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}\right) = 0.$$

For (2.6) and (2.7) we use the fact that, by conjugacy, there is an α_n and β_n so that $\lambda^2|y^n \sim \mathsf{Gamma}(\alpha_m, \beta_n)$. This gives that

$$E_{\lambda^2|y^n}\left(\frac{1}{\lambda^2}\right) = \frac{\beta_n}{\alpha_n - 1}.$$

Now, dropping the conditioning on y^n in the variance on the right of (2.6) it is

$$E_{\lambda^2|y^n} E_{\mu|y^n,\lambda^2} \mathrm{Var}(Y_{n+1}|\mu, \lambda^2) = E_{\lambda^2|y^n} E_{\mu|y^n,\lambda^2}\left(\frac{1}{\lambda^2}\right) = \frac{\beta_n}{\alpha_n - 1}. \tag{2.9}$$

Likewise, we can show that for $\kappa_n = n + \kappa_0$, (2.7) is

$$E_{\lambda^2|y^n} \mathrm{Var}_{\mu|y^n,\lambda^2} E(Y_{n+1}|\mu, \lambda^2) = E_{\lambda^2|y^n} \mathrm{Var}_{\mu|y^n,\lambda^2}(\mu) = E_{\lambda^2|y^n}\left(\frac{1}{\lambda^2\kappa_n}\right) = \frac{\beta_n}{\kappa_n(\alpha_n - 1)}.$$

Thus, we have that

$$\mathrm{Var}(Y_{n+1}|y^n) = \left(\frac{\kappa_n + 1}{\kappa_n}\right)\frac{\beta_n}{\alpha_n - 1}. \tag{2.10}$$

If we condition on λ first and then μ we find that

$$\begin{aligned}\mathrm{Var}(Y_{n+1}|y^n) &= E_{\lambda^2|y^n}E_{\mu|y^n,\lambda^2}\mathrm{Var}(Y_{n+1}|y^n,\mu,\lambda^2) && (2.11)\\ &+E_{\lambda^2|y^n}\mathrm{Var}_{\mu|y^n,\lambda^2}E(Y_{n+1}|y^n,\mu,\lambda^2) && (2.12)\\ &+\mathrm{Var}_{\lambda^2|y^n}E_{\mu|y^n,\lambda^2}E(Y_{n+1}|y^n,\mu,\lambda^2). && (2.13)\end{aligned}$$

Parallel to (2.6), it is easy to see that (2.12) is zero. By Fubini, (2.11) is the same as (2.6) as given by (2.9). Finally, since $\mathrm{Var}(Y_{n+1}|y^n)$ is constant independent of the condition, we can solve for (2.13). If desired, we can calculate $\mathrm{Var}(Y_{n+1}|y^n)$ directly and hence verify (2.10). We give one version of this in Appendix A.

Comparing the two orders of conditioning we see that in both the EEVar terms are the same. In the first decomposition, the VarEE term is zero whereas in the second decomposition the EVarE term is zero. Finally, in the first, the EVarE term has the κ_n while in the second the VarEE term has the κ_n. In particular, this shows that it is not a priori clear which terms will dominate in three-term expansions.

3 Bayes Model Averages as a BHM

Bayesian model averages (BMA's) can be seen as either a two level BHM or as a 3 level BHM. We see both here and relate them to the use of the LTV and posterior predictive decompositions.

3.1 Bayesian Model Averages I

One step up from (2.1) we can consider a BMA. Let $j = 1,\ldots,J$ index a collection of models $\mathcal{M} = \{M_1,\ldots,M_J\}$. Assume each M_j consists of a likelihood $p(y|\theta_j)$ and a prior $w(\theta_j,j) = w(\theta_j|j)w(j)$ where the across models prior $w(j)$ is discrete. Writing J for j as a random variabel as well as for the number of models will cause no confusion because the context will indicate which is meant. Now, we can represent this as a two level hierarhical model

$$\begin{aligned}(J,\theta_J) &\sim w(\theta_j,j)\\ Y &\sim p(y|\theta_j). && (3.1)\end{aligned}$$

Now, the L^2 BMA predictor is

$$E(Y_{n+1}|y^n) = \sum_{j=1}^{J} E(Y_{n+1}|y^n,M_j)W(M_j|y^n). \quad (3.2)$$

In (3.2), the two conditioning random variables, namely J and θ_j are treated explicitly and implicitly, respectively. In this case, it is not hard to see that one usage of the LTV recovers the usual formula for the posterior variance. Indeed, using the expression for posterior variance from p. 383 of [8], we find that (3.2) is

$$\mathrm{Var}(Y_{n+1}|y^n) = \sum_{j=1}^{J}\mathrm{Var}(Y_{n+1}|y^n,M_j)W(M_j|y^n)$$

$$
\begin{aligned}
& + \sum_{j=1}^{J} E(Y_{n+1}|y^n, M_j)^2 W(M_j|y^n) - E(Y_{n+1}|y^n)^2 \\
= \quad & E(\text{Var}(Y_{n+1}|M_J, y^n)|y^n) + \text{Var}(E(Y_{n+1}|M_J, y^n)). \qquad (3.3)
\end{aligned}
$$

So (3.3) is the result of using the LTV and conditioning on M_k. We have treated θ_j implicitly by integrated over it before conditioning on the M_j's. Reversing this i.e., integrating over j and using the LTV with Θ_k's would have been mathematically well-defined but statistically inappropriate for BMA. However, we shall see that different treatments of conditioning variables, when they make sense, typically give different terms for the same $\text{Var}(Y_{n+1}|y^n)$.

Let us interpret (3.3) similarly to how we interpreted (2.3) but using the M_k's, not the θ_k's. When the first term on the right 'E-Var' is large, we see most variability is in the predictive distributions from the high posterior probability models rather than from the variability across models. The second term on the right being small means that it doesn't matter very much which model you use for prediction. On the other hand, if Var-E is large, model selection is important but the smallness of the E Var term means the high posterior probability models are good.

For the sake of completeness, let us record another two term expression in the scope of (3.1) but conditioning on (J, Θ_J) as a two dimensional random variable:

$$\text{Var}(Y_{n+1}|y^n) = \text{Var}_{\Theta_K, K}(E(Y_{n+1}|\Theta_K, K, y^n)) + E_{\Theta_K, K}(\text{Var}(Y_{n+1}|\Theta_K, K, y^n)). \qquad (3.4)$$

This is different from (3.3) where we mixed out over the Θ_j's before examining the variability in M_J. That is, in (3.3), Θ_j's are implicit whereas in (3.4) they are explicit.

3.2 Bayesian Model Averages II: Three Level Hierarchical Model

Now write (3.1) as an equivalent three level hierarchical model:

$$
\begin{aligned}
J &\sim w(j) \\
\theta_j | J = j &\sim w(\theta_j | j) \\
Y &\sim p(y|\theta_j). \qquad (3.5)
\end{aligned}
$$

If we apply the law of total variance first to bring M_j into $\text{Var}(Y_{n+1}|y^n)$ we get (3.3). If we then use the LTV again in the first term on the right in (3.3) to bring in θ_j, we get

$$
\begin{aligned}
\text{Var}(Y_{n+1}|y^n, M_j) \quad = \quad & E_{\Theta_j|y^n, M_j} \text{Var}_{Y_{n+1}|y^n, M_j, \theta_j}(Y_{n+1}|y^n, M_j, \Theta_j = \theta_j,) \\
& + \text{Var}_{\Theta_j|y^n, M_j} E_{Y_{n+1}|y^n, M_j, \theta_j}(Y_{n+1}|y^n, M_j, \Theta_j = \theta_j). \qquad (3.6)
\end{aligned}
$$

Using (3.6) in (3.3) gives

$$
\begin{aligned}
\text{Var}(Y_{n+1}|y^n) \quad = \quad & E_J E_{\Theta_J|y^n, M_J} \text{Var}_{Y_{n+1}|y^n, M_J \Theta_j}(Y_{n+1}|y^n, M_J, \Theta_J) \\
& + E_J \text{Var}_{\Theta_J|y^n, M_J} E_{Y_{n+1}|y^n, M_J, \Theta_J}(Y_{n+1}|y^n, M_J, \Theta_J) \\
& + \text{Var}_J(E(Y_{n+1}|M_J, y^n)), \qquad (3.7)
\end{aligned}
$$

an instance of (1.8).

In (3.7) we conditioned first on M_J and then on Θ_J because the Θ_j's are naturally nested in the M_j's. There is nothing to prevent us from setting up a mathematical structure in which we can condition on Θ_j first and M_j second but that is not the natural way to think about this situation. (In the three term normal example, both orders of conditioning made sense.) In general, the order of conditioning affects which terms appear but the value of $\mathrm{Var}(Y_{n+1}|y^n)$ on the left is fixed once the hierarchy is fixed. In particular, the two models (3.1) and (3.5) are the same so the posterior variances on the left in (3.3) and (3.7) are equal pointwise in y^n. Hence the the expressions on the right are equal albeit different and are in the LTV-scope of the model. We can choose whichever sums of term in the scope of a given model we want depending on the variabilities of modeling quantities that concern us most.

It is seen from (3.7) that a three level hierarchy can lead to a three term expression for the posterior predictive variance because we have used the law of total variance twice, one for each level of the hierarchy above the likelihood. In general, each usage of the LTV generates one extra term.

We can also apply the law of total variance to the second term in (2.3), i.e., the last term on the right in (3.7). However, that will bring in the conditional expectation of a conditional variance of a conditional expectation which will not simplify. Such terms while mathematically correct are very difficult to handle. Moreover, the terms in (3.3) treat the Θ as latent and so depend on its distribution even though it is not explicitly indicated. For this reason, here, we only apply the LTV to the variances that occur in the leading term, i.e., the one of the form 'E Var', not any that have a 'Var E'. That is, while the full scope of a posterior predictive variance contains many terms from using the LTV in all possible ways, we focus on the subset of the scope where each term has exactly one variance operation that moves from left to right with appropriate conditioning. We call this this the Cochran Scope or C-Scope for short and henceforth limit our attention to sums of that form.

In this treatment of posterior variance the relative size of the terms is a tradeoff among the size of model list, the proximity of the parametric models on the list to each other, the across-models prior weights on models on the list, and the within-model priors. It's no longer purely a probabilistic model. We have to choose which terms we want to control in our model selection.

4 Generic Decompositions for the Posterior Predictive Variance: C-scope case

Recall the generic hierarchical model (1.1). Limiting attention to V_1, the LTV can be applied to give

$$\mathrm{Var}(Y_{n+1}|\mathcal{D}_n) = E(\mathrm{Var}(Y_{n+1}|V_1,\mathcal{D}_n)) + \mathrm{Var}(E(Y_{n+1}|V_1,\mathcal{D}_n)). \tag{4.1}$$

In (4.1), $\mathrm{Var}(Y_{n+1}|\mathcal{D}_n)$ looks the same as in other expressions such as (2.3) and (3.7) but in fact it depends on the full hierarchy in (1.1) because the posterior predictive variance ties all levels of the hierarchy together. That is levels 2 through K in (1.1) affect the posterior predictive variance on the left – and the terms on the right – even though they are suppressed in the notation.

We can now apply the law of total variance iteratively to itself, i.e.., to the first term – 'E Var' – in (4.1) by introducing conditioning on V_2. We can do the same in the new 'E E Var' term with V_3 and so on, generating one new term for each V_k at each iteration. Overall, this gives us $K+1$ terms involving means and variances. The expression for $K=2$ is now

$$\begin{aligned} Var(Y_{n+1}|\mathcal{D}_n) &= E_{V_1}E_{V_2}Var(Y_{n+1}|\mathcal{D}_n,V_1,V_2) + E_{V_1}Var_{V_2}E(Y_{n+1}|\mathcal{D}_n,V_1,V_2) \\ &\qquad + Var_{V_1}E(Y_{n+1}|\mathcal{D}_n,V_1) \end{aligned} \tag{4.2}$$

The left hand is a constant (given $\mathcal{D}_n$) independent of the order of conditioning on the right although different orders of conditioning will give different terms and of course, different hierarchical models will have different posterior variances.

4.1 General Structure

To address the general case and thereby quantify the uncertainty of the subjective choices we must make, recall $V=(V_1,\ldots,V_K)$, where V_k represents the values of the k-th potential choice that must be made to specify a predictor. Analogous to the terminology in ANOVA, we call V_k a *factor* in the prediction scheme, and we define the m_k levels of V_k to be $v_{k1},\ldots,v_{km_k}$. Thus, we take V to be discrete having probability mass function $W(v)=W(V_1=v_1\ldots,V_K=v_K)$. Effectively we are assuming that any continuous parameters are at the first level of the hierarchy above the likleihood and have been integrated out as in the BMA example in Subsec. 3.1. The V_k's are not in general independent under the prior W. Our model list is

$$\mathcal{V}^K=\{v_{11},\ldots,v_{1m_1}\}\times\ldots\times\{v_{K1},\ldots,v_{Km_K}\}.$$

We assume the $m_1\cdots m_K$ models in $\mathcal{V}^K$ are distinct and if they have a hierarchical structure (separate from the prior) we are ignoring it.

Our first result gives a decomposition of the posterior predictive variance by conditioning on the V_k's successively in the first term of (1.6). We only expand the first term because as was seen in Subsec. 2.1, the last term often goes to zero with increasing n. The general K case is in Clause (i) of Prop. 4.1. However, the order of conditioning will give different terms on the right. In practice, the ordering is chosen so that the terms most important to the analyst can be readily assessed. Clause (ii) of Prop. 4.1 is a variant on Clause (i) from collapsing all the levels in the hierarchy above the likelihood into a single conditioning variable.

Proposition 4.1. *We have the following two expressions for the posterior predictive variance when the factors correspond to a model list.*
Clause (i): For $K=1$ in (1.1) we have

$$Var(Y_{n+1}|\mathcal{D}_n)=E(\,Var(Y_{n+1}|V,\mathcal{D}_n))+Var(E(Y_{n+1}|V,\mathcal{D}_n)). \tag{4.3}$$

and, for $K \geq 2$ in (1.1), the posterior predictive variance of Y_{n+1} as function of the factors defining the predictive scheme is

$$\begin{aligned}Var(Y_{n+1}|\mathcal{D}_n)(\mathcal{V}^K) = {} & E_{(V_1,\dots,V_k)}Var(Y_{n+1}|\mathcal{D}_n, V_1,\dots,V_K) \\ & + \sum_{k=2}^{K} E_{(V_1,\dots,V_{k-1})}Var_{V_k}E(Y_{n+1}|\mathcal{D}_n, V_1,\dots,V_k) \\ & + Var_{V_1}E(Y_{n+1}|\mathcal{D}_n, V_1). \end{aligned} \tag{4.4}$$

Clause (ii): For any K, the posterior predictive variance $Var(Y_{n+1}|\mathcal{D}_n)(\mathcal{V}^K)$ can be condensed into a two term decomposition:

$$\begin{aligned}Var(Y_{n+1}|\mathcal{D}_n)(\mathcal{V}^K) = {} & E_{(V_1,\dots,V_K)}Var(Y_{n+1}|\mathcal{D}_n, V_1,\dots,V_K) \\ & + Var_{(V_1,\dots,V_K)}E(Y_{n+1}|\mathcal{D}_n, V_1,\dots,V_K). \end{aligned} \tag{4.5}$$

Remark: Clause i) is a generalization of (13) in [4]; Clause ii) is a formal statement of (12) in [4].

Proof. The proof of *Clause i)* is a straightforward iterated application of the law of total variance and *Clause ii)* follows from the law of total variance simply treating V as a single long vector rather than as the string of its components. □

A natural question is: how large is the C-scope of a BHM? As suggested by the last result, we have to account for the order of conditioning and which of the K variables are used explicitly in each conditioning step (variance or expectation). Let u be the number of usages of the LTV and let $M \leq K$ be the total number of the V_k's used in the sequence of conditionings. To obtain the general expression for the cardinality of the C-scope, let $S(M,u)$ be the Stirling number of the second kind. That is, for fixed M and u, $S(M,u)$ is the number of ways to form non-void, disjoint, and exhaustive collections of u sets from M distinct objects. As with Prop. 4.1, we only expand the leading $E - Var$ term. We have the following formula.

Proposition 4.2. *We have the following expressions for the $\#(C - \text{scope})$. Fix K, M, u where K is from (1.1), M is the number of manifest V_k's, and u is the number of usages of the LTV. Then $u \leq M \leq K$ and the number of variance decomposition after u uses of the LTV for a fixed number M of manifest variables from $(V_1,\dots,V_K)$, is*

$$u!\binom{K}{M}S(M,u). \tag{4.6}$$

Consequently, for fixed K, the total number of variance decompositions is

$$C - \text{scope} = \sum_{u=1}^{K}\sum_{M=u}^{K} u!\binom{K}{M}S(M,u). \tag{4.7}$$

Remark: For $K = 2$, the C-scope from the RHS of (4.7) is five. We can list these possibilities as follows. Consider (1.8) or equivalently (4.4) with $K = 2$. Then $M = 1, 2$. Here are the cases. For $M = 1$ and $u = 1$ there are two possibilities: Condition on V_1 alone or V_2 alone i.e., manifest, so that V_2 or respectively V_1 is latent. For $M = 2$ and $u = 1$, there is one possibility, condition on (V_1, V_2). For $M = 2$ and $u = 2$ there are two possibilities, condition on V_1 and then V_2 or condition on V_2 an then V_1. In all five cases, the BHM is fixed so $\mathrm{Var}(Y_{n+1}|y^n)$ is the same even though the decompositions are different. So, we have five different ways to model the posterior variance in the same hierarchical model. While they are equivalent mathematically and come from the same BHM they are not equivalent statistically.

Proof. Start by observing that if we are going to use u instances of the LTV, then we must have M disjoint nonvoid subsets of $(V_1, \ldots, V_K)$, i.e., not counting permutations, there are $S(M, u)$ possible choices. Since we can permute these sets any way we want, we get a factgor of $u!$. Since we can do this for any choice of M manifest variables we get a factor of 'K choose M', thereby giving (4.6). Summing over all the possible values of $M$and u gives (4.7). □

Using stacking – or any other model averaging procedure – in place of the BMA leads to results analogous to Props. 4.1 and 4.2; see [5].

4.2 Choosing V_K

Our work here parallels testing whether a factor in ANOVA should be retained. Indeed, if data collectors think they know pre-experimentally which are V_k's are important to retain the methods here will not help. However, this is usually not the case and more generally when a BHM is not physically motivated, e.g., the V_k's are mathematical aspects of the likelihood, it will often not be clear which V_k's (or values of V_k's) are important to retain.

An example may help. One choice of V with $K = 2$, that can often be used to winnow down a model list is the following. Consider trying to assess the importance of sets of variables. Suppose we have a list of models $\mathcal{M} = \{m_1, \ldots, m_q\}$ and a set of explanatory variables $\mathcal{X} = \{X_1, \ldots, X_p\}$. If $q = 2$, for instance, m_1 may be a linear model and m_2 may be a non-linear model. Write $\mathcal{P}(\mathcal{X}) = \{\{X\}_1, \ldots, \{X\}_{2^p}\}$ to mean the power set of $\mathcal{X}$. Now we can consider each model with each subset of explanatory variables as inputs to the modeling. Here, V_1 corresponds to the uncertainty in the predictive problem due to the models and V_2 corresponds to the selection of variables we use in a models. We use a version of this in Subsec. 5.1. The idea is that the experimenter has little information about which variables should be included and would prefer using a linear model if possible for ease of interpretation.

Now, we can use the decomposition in Subsec. 3.1 or 3.2. In addition, using a Bayes model average we write the posterior predictive density

$$p(Y_{n+1}|\mathcal{D}_n) = \sum_{i=1}^{q} p(m_i|\mathcal{D}_n) \sum_{j=1}^{2^p} p(\{X\}_j|\mathcal{D}_n, m_i) p(Y_{n+1}|\mathcal{D}_n, \{X\}_j, m_i), \tag{4.8}$$

generically denoting densities as p. Now, we can calculate the posterior probability for each set of explanatory variables from

$$p(\{X\}_j|\mathcal{D}_n) = \sum_{i=1}^{q} p(m_i|\mathcal{D}_n)p(\{X\}_j|\mathcal{D}_n, m_i).$$

This posterior probability is a measure of "variable set importance". A similar expression gives a measure of importance for an individual model. Thus, when a level in the BHM represents a mathematical quantity that we does not have a clear physical correlate, and hence cannot be included or excluded based on physical modeling, we can use our decompositions to assess whether a factors in V is worth including or can be collapsed to a single value.

We can represent any conditioning quantity as $V = (V_1, \ldots, V_K)^T$. As in Subsec. 2.1 for $K = 1$, V_1 might simply be a parameter. As in Subsec. 3.2, for $K = 2$, V_1 might be a model and V_2 might correspond to a parameter. Or, as in Sec. 5, V_1 may correspond to the choice of link function in a GLM while V_2 may correspond to selections of explanatory variables (as above). Thus, we must choose a K and we can regard each V_k as an aspect of a modeling strategy. For instance, if $K = 2$, V_1 may be a 'scenario' and V_2 may be a 'model' in the sense of [4], a parallel we develop in Sec. 5.1. We will write as if the V_k's are discrete modeling choices remembering that the law of total variance applies for continuous random variables as well.

Note that the sort of hierarchical modeling we advocate here can seem artificial in the sense that we can use a strictly mathematical approach, simplifying the model down to the quantities that seem to matter predictively, and then using the resulting model to form PI's. Once good prediction has been achieved the quest for a more realistic model (that will often not perform as well predictively) can begin and compared with the formal model our approach yields. Consequently, we advocate the generation of multiple BHM's, using different mathematical features. In Sec. 5, we use link functions in a GLM as a conditioning variable. In [5], we used a unidimensional V to represent the choice of a shrinkage method in penalized linear regression.

Typically, one of the most important levels in a BHM is the selection over models. We can enlarge a model list simply by including more plausible models. However, this may lead to problems such as dilution; see [6]. So, we want to assess the effect of a model list on the variance of predictions. Consider a model list $\mathcal{M}$ and suppose we don't believe it adequately captures the uncertainty (including mis-specification) of the the predictive problem. We can expand the list by including other competing models and this can be done by adding more models to it or by embedding the models on the list in various 'scenarios' as is done in [4]. Once a new model list $\mathcal{M}'$ is constructed, if it contains different models with positive posterior probability the posterior predictive distribution $p(Y_{n+1}|\mathcal{D}_n)$ resulting from $\mathcal{M}'$ will be differ from $p(Y_{n+1}|\mathcal{D}_n)$ resulting from $\mathcal{M}$. Hence, we can use the decompositions here to help decide which of model $\mathcal{M}$ and $\mathcal{M}'$ is more reasonable and hope that both simplify to the same predictor.

In reality, the relative sizes of terms in the various decompositions of the form (4.4) depend heavily on the choice of K, V, and the likelihood. Fortunately, in practice, we usually only have one largest V_K that we most want to consider even though different orderings of the V_k's will affect which terms appear in the decomposition. Once K and V have been chosen, posterior predictive variance decompositions using the V_k's can be generated and examined for which terms are important. The selection of V in general is an aspect of model list selection that is beyond our present scope.

5 Decompositions For Uncertainty Quantification

Here we redo and extend some two examples developed from [4] and a further example drawn from Bayesian Two-Way ANOVA.

5.1 Revisiting Draper (1995)

We can apply our techniques to two examples given in [4] and one further example that his second example motivates. The first example involves predicting the price of oil; the second example involves predicting the chance of failure of O-rings in a space shuttle at a new temperature. Our third example for this subsection is an extension of the latter data type with a more difficult variable selection problem. Draper's main point was that when making predictions, we need to consider the uncertainty of the 'structural' choices we make or we can be led to bad decisions. Here, we have formalized Draper's concept of structural choices in our conditioning variable V. One danger in poor structural choices is that a PI may be found that is unrealistically small leading to over-confidence.

Oil Prices

In the oil prices example in [4] there are two 'structural' components to the modeling namely, 12 economic scenarios with 10 economic models nested inside them. These components represent 120 models and hence introduce model uncertainty that must be quantified to generate good PI's.

In Draper's analysis each model was used given the parameters of each scenario. This corresponds to $K = 2$ and a three term posterior predictive variance decomposition. Let s_i denote scenario i and m_{ij} be model j within scenario i. Write $s_i \in S$ and $m_{ij} \in M_i \subset M$ where M_i is the set of models for scenario i and M is the union of the M_i's. Now,

$$\begin{aligned} Var(Y_{n+1}|\mathcal{D}_n)(S,M) &= E_S E_M Var(Y_{n+1}|\mathcal{D}_n, S, M) && (5.1)\\ &+ E_S Var_M(E(Y_{n+1}|\mathcal{D}_n, S, M)) && (5.2)\\ &+ Var_S(E_M(E(Y_{n+1}|\mathcal{D}_n, S))) && (5.3)\\ &= 178 + 363 + 354 = 895, && (5.4)\end{aligned}$$

which is exactly equation (13) in [4]. In our notation, the E_M in Draper's last term is suppressed in Clause (i) of Prop. 4.1. We cannot recompute this example because neither the data nor the details on the scenarios or models are available to us. Nevertheless, we have the following interpretations. The proportion of the PPV attributable to the predictions within models and scenarios, i.e., term (5.1) , is about 20% (178/895). The proportion of the PPV attributable to the between-models within scenarios variance, i.e., term (5.2), is about 40% (363/895). And, the proportion of the PPV attributable to the between-scenarios variance, i.e., term (5.3), is also about 40% (354/895). (See Table 1 in [5] for the general definition of terms following the usage in [4] p. 58-9.) Thus unless one is wiling to ignore 20% of the predictive variability, all the three terms in (5.4) must be used when forming PI's.

Challenger Disaster

Making the decision to launch the space shuttle at an ambient temperature at which the various components had not been tested ended up being catastrophic – and could have been avoided had a proper uncertainty analysis had been done. Statistically, the error of the decision makers was to choose a single model from a model list rather than incorporating all sources of predictive uncertainty into their analysis. The goal of this example originally was to show that a correct analysis of the various sources of uncertainty would have led to a credibility interval for $p_{t=31}$ the probability of an O-ring failure (at 31°) of $(.33, 1]$. Thus, using any reasonable value of $\hat{p}_{t=31}$ would have led to a PI with far too high a probability of failure for a launch to be safe. Our goal in re-analyzing Draper's example based on BHM's and the LTV is to identify which sources of uncertainty can be neglected.

We have 23 observations of the number of damaged O-rings ranging from zero to six (because each shuttle had six O-rings). Each observation also has a temperature t and a 'leak-check' pressure s. Following Draper's analysis we also use t^2 as an explanatory variable. Thus we have 24 vectors, each of length four.

We assume the number of damaged O-rings follows a $Binomial(6, p)$ distribution where p is a function of the explanatory variables via one of three link functions, logit, $c \log \log$, and probit. Thus, we have structural uncertainty in the choice of variables and in the choice of link function. In our notation, we set $V_1 = \{L, C, P\}$ for the choice of link function, logit, $c \log \log$, and probit respectively. Also let $V_2 = \{t, t^2, s, \text{no effect}\}$ where no effect means an intercept-only model. The 24 models are listed in Table 1.

Table 1: **List of models for the Challenger disaster data:** This table lists all 24 models under consideration broken down by their structural choices – link functions and explanatory variables.

$\mathcal{V}^{(2)}$	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}	m_{12}	m_{13}
V_1	L	L	L	L	L	L	L	L	C	C	C	C	C
V_2	t	t^2	s	t, t^2	t, s	t^2, s	t, t^2, s	no effect	t	t^2	s	t, t^2	t, s
$\mathcal{V}^{(2)}$	m_{14}	m_{15}	m_{16}	m_{17}	m_{18}	m_{19}	m_{20}	m_{21}	m_{22}	m_{23}	m_{24}		
V_1	C	C	C	P	P	P	P	P	P	P	P		
V_2	t^2, s	t, t^2, s	no effect	t	t^2	s	t, t^2	t, s	t^2, s	t, t^2, s	no effect		

In fact, Draper did not consider all of these models. Essentially he put zero prior probability on all models except for m_1, m_4, m_5, m_7, m_8, and m_{15}. Accordingly, he only considered the set

$$\mathcal{M} = \{m_1, m_4, m_5, m_7, m_8, m_{15}\}$$

with a uniform prior. Draper then gave a table of posterior quantities for the structural choices, and a posterior predictive variance decomposition for within-structure and between-structure variances as

$$Var(p_{t=31}|\mathcal{D}_{23}) = Var_{within} + Var_{between} = 0.0338 + 0.0135 = 0.0473. \tag{5.5}$$

That is, Draper used a two term decomposition based on Clause ii) of Prop. 4.1. Draper's conclusion was that .0135/.0473 ≈ 28.5% so the uncertainty represented by the second term in (5.5) could not be neglected.

Here we extend Draper's analysis and confirm that structural uncertainty should not have been ignored. For our implementation, we use the full set of 24 models but do not employ the same approximations. Then, we use the BMA package in R to get the posterior distributions of the parameters of the models and the posterior weights for V_2. We also use the rjmcmc package to get the posterior weights for V_1. The resulting posterior distributions are qualitatively similar to Draper's approximate posteriors.

Considering all sources of uncertainty yields a posterior predictive variance decomposition of

$$\begin{aligned} Var(p_{t=31}|\mathcal{D}_{23}) &= E_{V_1}E_{V_2}Var(p_{t=31}|\mathcal{D}_{23},V_1,V_2) + E_{V_1}Var_{V_2}E(p_{t=31}|\mathcal{D}_{23},V_1,V_2) \\ &\quad + Var_{V_1}E(p_{t=31}|\mathcal{D}_{23},V_1) \\ &= .054 + .099 + .003 = .155. \end{aligned} \tag{5.6}$$

This is almost three times the variance as obtained by Draper. We confirm his intuition that structural uncertainty was much greater than assumed when making the decision to launch the shuttle. Moreover, Draper commented that other analyses could lead to larger posterior variances. So, (5.6) is consistent with his intuition.

Looking at the numbers in (5.6) we can see the last is an order of magnitude smaller than the other two. Thus, we conclude that the terms representing the between-models within-link functions variance and the between-predictions within-models and links variance are terms that must be retained and the third term can be taken as zero. A frequentist testing approach confirms this; see [5]. So, we would be led to consider a new hierarchical model that did not include V_1 and therefore had a two term decomposition using only V_2 giving a new value of $\mathrm{Var}(Y_{n+1}|\mathcal{D}_n)$. In effect, we would compare this decomposition with the first two terms on the right in (5.6) to see which expression for the posterior predictive variance is more convincing.

So, we drop the V_2 level in the BHM and form a new hierarchical model by setting V_1 to be logit. Now, we are back to a two term decomposition – with a new value of $\mathrm{Var}(Y_{n+1}|\mathcal{D}_n)$. Thus we have

$$\begin{aligned} Var(p_{t=31}|\mathcal{D}_{23}) &= E_{V_2}Var(p_{t=31}|\mathcal{D}_{23},V_2) \\ &\quad + Var_{V_2}E(p_{t=31}|\mathcal{D}_{23},V_2) \\ &= .088 + .064 = .152. \end{aligned}$$

Now, both terms look important so we can't drop V_2. Note that .152 in (5.7) is close to .155 in (5.6) and the values of the two larger terms, while similar, indicate a reverse importance, i.e., $.054 < .088$ and $.099 > .064$.

We remark that Draper's formulation is predictive only in the sense that the variability in p_t determines how we would predict a future Y. For a purely predictive formulation we would use a three term decomposition like that in **Oil Prices**:

$$\begin{aligned} Var(Y_{n+1}|\mathcal{D}_n) &= E_{V_1}E_{V_2}Var(Y_{n+1}|\mathcal{D}_n, V_1, V_2) + E_{V_1}Var_{V_2}E(Y_{n+1}|\mathcal{D}_n, V_1, V_2) \\ &+ +Var_{V_1}E(Y_{n+1}|\mathcal{D}_n, V_1) \end{aligned} \tag{5.7}$$

but our 'Y_{n+1}" here would be the number of successes in 30 trials, a random variable, as opposed to a probability such as $p_{t=31}$. We did not do this here because we wanted to compare directly with Draper's work.

5.2 Bayesian Two-Way ANOVA

Consider the two-way ANOVA model defined by

$$Y_{ij} = \tau_i + \beta_j + \epsilon_{ij}, \tag{5.8}$$

where $i = 1, \ldots, T$, $j = 1, \ldots, B$, and we have the following distributional properties:

$$\begin{aligned} \tau_i &\sim N((\tau_0, \sigma_\tau^2) \\ \beta_j &\sim N(\beta_0, \sigma_\beta^2) \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \end{aligned} \tag{5.9}$$

with

$$\tau_i \perp\!\!\!\perp \tau_j, \quad \beta_i \perp\!\!\!\perp \beta_j$$

for $i \neq j$ and for all i, j

$$\tau_i \perp\!\!\!\perp \beta_j, \quad \tau_i, \beta_j \perp\!\!\!\perp \epsilon_{ij}.$$

Essentially, for each time step n we have a $T \times B$ matrix of random variables $(Y_{ijh})_{i=1,\ldots,T; j=1,\ldots,B}$ that we can write as $Y_{n;ij}$. An obvious simplification of this is to take $B = 1$. Here we regard these matrices as a sequence of two-way ANOVA's with one observation per cell. Our reasoning can be extended to multiple observations per cell but this becomes very complicated.

After n time steps we have a sequence of n $T \times B$ matrices that we denote $\mathbf{y}^n$. Thus our predictive problem is to use the first n matrices to obtain an expression for the $T \times B$ conditional covariance matrix for the $n+1$ random variable's Y_{ij} given $\mathbf{y}^n$, i.e.,

$$\begin{aligned} \text{Var}(Y_{ij;n+1}|\mathbf{y}^n) &= E_{\tau|\mathbf{y}^n}E_{\beta|\mathbf{y}^n,\tau}\text{Var}(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) \\ &\quad +E_{\tau|\mathbf{y}^n}\text{Var}_{\beta|\mathbf{y}^n,\tau}E(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) \\ &\quad +\text{Var}_{\tau|\mathbf{y}^n}E_{\beta|\mathbf{y}^n,\tau}E(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau). \end{aligned} \tag{5.10}$$

We have the following expressions for the three terms in (5.10):

$$\text{Term } 1 = \sigma_\epsilon^2$$

$$
\begin{aligned}
\text{Term 2} &= \left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)^{-1} \\
\text{Term 3} &= \frac{1}{\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)^2} \left(\left(\frac{T+1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right)^2 \cdot \frac{1}{a} \left(1 - \frac{Bb}{a + bBT} \right) + \frac{(T-1)}{\sigma_\epsilon^4} \frac{1}{a} \left(1 - \frac{Bb}{a + bBT} \right) \right. \\
&\quad \left. + \frac{2}{\sigma_\epsilon^2} \left(\frac{T+1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right) (T-1) \frac{-Bb}{a(a + BbT)} \right),
\end{aligned}
$$

where $a = \frac{B}{\sigma_\epsilon^2} + \frac{1}{\sigma_\tau^2}$ and $b = -\left(\sigma_\epsilon^4 (\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2})\right)^{-1}$; derivations are given in Appendix B.

The key issue is the relative sizes of these three terms. Term 1 is easy to visualize because it's a constant. However, Terms 2 and 3 are difficult to visualize. So, we have generated some graphs as an effort to determine how they behave as a function of their inputs, namely T, B, σ_ϵ, σ_τ, and σ_β. The top row in Fig. 1 shows the the three terms as functions of T, setting $B = 2$, and $\sigma_\epsilon = \sigma_\tau = \sigma_\beta = 1$.

In the top left panel, past about $T = 8$, we see that Term 1 dominates as $T \to \infty$ and the other two terms decrease to zero. This is confirmed by the top right panel and we see that around $T = 20$, Term 2 can be omitted at a threshold of about 5%. In the bottom left panel, we see that Term 3 dominates as $\sigma_\beta \to \infty$ while Term 3 goes to zero and Term 1 stays constant. This is corroborated in the bottom left panel where we see that both Terms 1 and 2 can eventually be dropped at a value of σ_β to the right of the graph. This shows that which terms dominates depends delicately on the exact scenario in which the PPV is computed.

6 Discussion

The main contribution of this paper is to provide a decomposition of the posterior predictive variance (PPV) for a Bayesian hierarchical model. An immediate benefit from this is that we have a conservation law over decompositions for the PPV. This is important for two reasons. First, the posterior predictive variance controls the width of prediction intervals so we want to know what aspects of variance are contributing most to it. Second, we want to identify what levels of a BHM can be collapsed to a single value. This is analogous to testing for whether a factor can be dropped in a frequentist multi-way ANOVA.

Our decompositions start with a fixed BHM and hence a fixed PPV that can be expressed in multiple decompositions depending on the how use of the LTV is iterated. The various decompositions depend on the ordering of the conditioning variables from the levels of the BHM. We focus on what we call the C-scope of a BHM – the collection of decompositions of the posterior predictive variance that arise from using the law of total variance only on terms in which an expectation of a variance appears. In Prop. 4.2 we give an explicit expression for the cardinality of the C-scope.

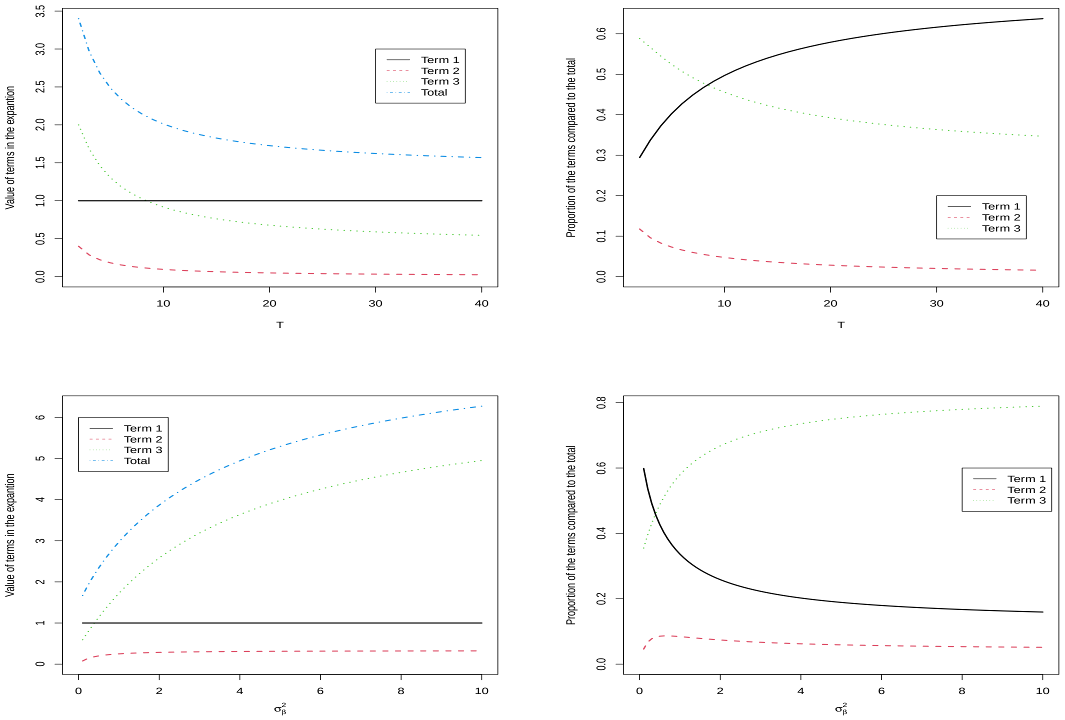

Figure 1: Top: graphs of the three terms for $\tau = 2, \ldots, 40$. Left: actual values of the terms and their total. Right: proportions of each term relative to the total. Bottom: graphs of the three terms for σ_β ranging over $[.1, 10]$. Left: actual values of the terms and their total. Right: proportions of each term relative to the total.

The main modeling implication of our work is that we can more readily use BHM's where we might have used Bayesian nonparametrics. Indeed, we can represent any feature of a statistical model and a random variable with a prior. These features may or may not have any physical correlate: in Subsec. 5.1 we use variable selection as a level in a BHM and this is part of physical modeling. We also take the link function in a GLM as a feature and this is not necessarily an aspect of modeling. Elsewhere, see [5], we used selection if a shrinkage method as a feature of modeling and this does not really have a physical correlate.

One effect of using variance is that the metric properties of the model list become important as well as its probabilistic properties. Thus, as a matter of model list design we want to choose a BHM so that its PPV is neither too small nor too large relative to the data so that using multiple decompositions to prune out levels will be effective. Indeed, we may want to construct a BHM so that the higher the level the less it is thought to matter and then order the uses of the LTV so that we start by conditioning on the level we most think we can eliminate. It is a sort of folk-theorem that the higher the level the less important is and our procedure can assess this. Even though we can construct examples of arbitrary many levels in which the top level does matter, the intuition holds and extensions of our work may be able to provide a formal way to decide if upper levels in a BHM should be retained.

One drawback of our procedure is that it we do not have a formal way to assess the relative contributions of terms in the decomposition. We have relied on essentially a user specified threshold for whether a term is large enough to retain. This is so because in general we do not have a likelihood for these terms and therefore cannot do Bayes testing directly. On the other hand, there are ways around this e.g., pseudo-Bayes posteriors in which a likelihood is formed from an empirical risk. We have not investigated this possibility, but it is promising as it is in the spirit of the mathematical modeling we advocate here, namely being willing to use mathematical quantities without physical motivation as a way to produce predictive analyses.

We conclude by observing that the treatment we have given for variance can, in principle, be extended to higher level moments even though it looks hard. For instance, [2] gives a way to calculate cumulants of a distribution that can be a posterior quantity. He gives a formula similar to our Prop. 4.1 and gives examples using this result for sums of variables and mixture distributions. The order of the cumulants is arbitrary but lower orders would likely be easier to use. In addition, we could have used the Shannon mutual information in place of the variance and invoked its chain rule. We have not chosen these because the first seems quite hard and the second is not as readily applicable to data.

Appendix A: Calculations for the Three Term Normal

Our task is to derive a expression for $\mathrm{Var}(Y_{n+1}|y^n)$ directly. Using the definitions in Subsec. 2.2, we have two parameters μ and λ as well as three hyperparameters κ_0, α_0, and β_0. For simplicity, write $\gamma = \lambda^2$. The conditional density of y^n given μ and λ^2 is

$$p(y^n|\mu,\gamma) = \gamma^{n/2} e^{-\gamma/2\sum_{i=1}^n (y_i-\mu)^2} \sqrt{\gamma\kappa_0} e^{-\kappa_0\gamma/2\sum_{i=1}^n (\mu-\mu_0)^2} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \gamma^{\alpha_0-1} e^{-\beta_0\gamma}. \tag{A.1}$$

We have that

$$\begin{aligned} p(y_{n+1}|y^n) &= \int\int p(y_{n+1}|y^n,\mu,\gamma)p(\mu,\gamma|y^n)\mathrm{d}\mu\mathrm{d}\gamma \\ &= \int\int p(y_{n+1}|y^n,\mu,\gamma)p(\mu|y^n,\gamma)p(\gamma|y^n)\mathrm{d}\mu\mathrm{d}\gamma. \end{aligned} \tag{A.2}$$

We want to identify the three densities in the integrand. We know the first.

For the second, with some foresight, let

$$\begin{aligned} \mu_n &= \frac{n\bar{y} + \kappa_0 \mu_0}{\kappa_n} \\ \gamma_n &= \gamma(n + \kappa_0) = \gamma\kappa_n \\ \kappa_n &= n + \kappa_0 \\ \alpha_n &= \alpha_0 + (n/2) \\ \beta_n &= \beta_0 + \frac{1}{2}. \end{aligned} \tag{A.3}$$

Step 1: We begin by seeing that $p(\mu|y^n, \gamma) \sim N(\mu_n, 1/\gamma_n)$. The squared terms in the exponent in (A.1) are

$$\begin{aligned} & -\frac{\gamma}{2}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{\kappa_0\gamma}{2}(\mu - \mu_0)^2 \\ = & -\frac{\gamma}{2}\left[\sum_{i=1}^{n} y_i^2 + n\mu^2 - 2n\bar{y}\mu + \kappa_0\mu^2 + \kappa_0\mu_0^2 - 2\kappa_0\mu\mu_0\right] \\ = & -\frac{\gamma}{2}\left[\mu^2(n + \kappa_0) - 2\mu(n\bar{y} + \kappa_0\mu_0) + \sum_{i=1}^{n} y_i^2 + \kappa_0\mu_0^2\right] \end{aligned} \tag{A.4}$$

Completing the square in μ means (A.4) becomes

$$\begin{aligned} & -\frac{\gamma}{2}\left[\mu^2(n + \kappa_0) - 2\mu\sqrt{n + \kappa_0}\frac{(n\bar{y} + \kappa_0\mu_0)}{\sqrt{n + \kappa_0}} + \frac{(n\bar{y} + \kappa_0\mu_0)^2}{n + \kappa_0}\right] \\ & \quad -\frac{\gamma}{2}\left[\sum_{i=1}^{n} y_i^2 + \kappa_0\mu_0^2 - \frac{(n\bar{y} + \kappa_0\mu_0)^2}{n + \kappa_0}\right] \\ & = -\frac{\gamma(n + \kappa_0)}{2}\left[\mu - \frac{n\bar{y} + \kappa_0\mu_0}{n + \kappa_0}\right]^2 - \frac{\gamma}{2}\left[\sum_{i=1}^{n} y_i^2 + \kappa_0\mu_0^2 - \frac{(n\bar{y} + \kappa_0\mu_0)^2}{n + \kappa_0}\right]. \end{aligned} \tag{A.5}$$

Note that the 'extra' $\sqrt{\gamma}$ in (A.1) is absorbed in the normal density. This completes step 1.

Step 2: Next, we see that $p(\gamma|y^n) \sim \mathsf{Gamma}(\alpha_n, \beta_n)$. By exponentiating the second term in (A.5) and multiplying it by the 'active' factors in (A.1) we get that the rest of the likelihood is proportional to

$$\gamma^{\alpha_0 + (n/2) - 1}\gamma^{n/2}e^{-\gamma(\beta_0 + (1/2)[\sum_{i=1}^{n} y_i^2 + \kappa_0\mu_0^2 - \kappa_n\mu_n^2]}. \tag{A.6}$$

Upon normalization this gives Step 2.

We comment that in principle, we now have the right hand side of (A.2). However, finding $\mathrm{Var}(Y_{n+1}|y^n)$ directly is a lot of work (probably involving t-distributions). So, we use a two term expansion. For this we derive the following.

Step 3: Obtain the conditional posterior

$$p(y_{n+1}|y^n, \gamma) \sim N\left(\mu_n, \frac{\kappa_n + 1}{\kappa_n\gamma}\right). \tag{A.7}$$

To do this, first note

$$p(y_{n+1},\gamma|y^n) = p(y_{n+1}|y^n,\gamma)p(\gamma|y^n). \tag{A.8}$$

Since we have $p(\gamma|y^n)$ it is enough to find the right hand side of (A.8). To do this recall that by definition

$$p(y_{n+1},\gamma|y^n) = p(\gamma|y^n)\int p(y_{n+1}|\mu,\gamma)p(\mu|y^n,\gamma)\mathrm{d}\mu \tag{A.9}$$

The integrand in (A.9) (in μ) is

$$\begin{aligned}&\propto \sqrt{\gamma}e^{-(\gamma/2)(y_{n+1}-\mu)^2}\times\sqrt{\gamma_n}e^{-(\gamma/2)(\mu-\mu_n)^2}\\&= \gamma\sqrt{\kappa_n}e^{-(\gamma/2)[(y_{n+1}-\mu)^2+\kappa_n(\mu-\mu_n)^2]}.\end{aligned} \tag{A.10}$$

By some notational gymnastics, completing the square in (A.10) gives that

$$\begin{aligned}&(y_{n+1}-\mu)^2+\kappa_n(\mu-\mu_n)^2\\&= (1+\kappa_n)\left(\mu-\frac{y_{n+1}+\kappa_n\mu_n}{1+\kappa_n}\right)^2+y_{n+1}^2+\kappa_n\mu_n^2-\frac{(y_{n+1}+\kappa_n\mu_n)^2}{1+\kappa_n}.\end{aligned} \tag{A.11}$$

Using (A.11) in (A.10) gives that the integrand in (A.9) is

$$\begin{aligned}&\propto \frac{\sqrt{\kappa_n}}{\sqrt{1+\kappa_n}}\sqrt{\gamma(1+\kappa_n)}e^{-(\gamma/2)\left[(1+\kappa_n)\left(\mu-\frac{y_{n+1}+\kappa_n\mu_n}{1+\kappa_n}\right)^2\right]}\\&\times\sqrt{\gamma}e^{-(\gamma/2)\left[y_{n+1}^2+\kappa_n\mu_n^2-\frac{(y_{n+1}+\kappa_n\mu_n)^2}{1+\kappa_n}\right]}.\end{aligned} \tag{A.12}$$

The first factor can be integrated over μ and the exponent in the second factor is

$$\begin{aligned}&y_{n+1}^2+\kappa_n\mu_n^2-\frac{(y_{n+1}+\kappa_n\mu_n)^2}{1+\kappa_n}\\&= \frac{1}{1+\kappa_n}\left[\kappa_n\left(y_{n+1}^2+\mu_n^2-2y_{n+1}\mu_n.\right)\right]\\&= \frac{\kappa_n}{1+\kappa_n}\left(y_{n+1}-\mu_n\right)^2.\end{aligned} \tag{A.13}$$

Now we see that the integral in (A.9) gives (A.7), completing Step 3.

To complete the derivation of the posterior variance, write

$$\begin{aligned}\mathrm{Var}(Y_{n+1}|y^n) &= E\left[\mathrm{Var}(Y_{n+1}|y^n,\gamma)\right]+\mathrm{Var}\left[E(Y_{n+1}|y^n,\gamma)\right]\\&= \frac{1+\kappa_n}{\kappa_n}E\left[\frac{1+\kappa_n}{\kappa_n}\frac{1}{\gamma}\right]+\mathrm{Var}(\mu_n)\\&= \frac{1+\kappa_n}{\kappa_n}\frac{\beta_n}{\alpha_n-1},\end{aligned} \tag{A.14}$$

since μ does not depend on γ.

Appendix B: Calculations for the Two Way ANOVA

Here we give the details for working out the three term variance decomposition for a two-way random effects ANOVA from Subsec. 5.2.

Step 1: Decompose the log-likelihood. For any i, j we have that

$$\begin{aligned}
\ln p(y_{ij}, \tau_i \beta_j) &= \ln p(y_{ij}|\tau_i, \beta_j) + \ln p(\tau_i) + . \ln p(\beta_j) \\
&= \left[-\frac{1}{2\sigma_\epsilon^2} \sum_{i,j} (y_{ij} - \tau_i - \beta_j)^2 - \frac{1}{2\sigma_\tau^2} \sum_{i,j} (\tau_i - \tau_0)^2 - \frac{1}{2\sigma_\beta^2} \sum_{i,j} (\beta_j - \beta_0)^2 \right] \\
&+ \quad ExtraTerms
\end{aligned} \tag{B.1}$$

Apart from the $-1/2$ factor, the part of expression (B.1) in square brackets is

$$\begin{aligned}
& \frac{1}{\sigma_\epsilon^2} \sum_i \sum_j \left(y_{ij} + \tau_i^2 + \beta_j^2 - 2y_{ij}\tau_i - 2y_{ij}\beta_j - 2\tau_i\beta_j \right) \\
& \quad + \frac{1}{\sigma_\tau^2} \sum_i \left(\tau_j^2 + \tau_0 - 2\tau_i\tau_0 \right) \\
& \quad + \frac{1}{\sigma_\beta^2} \sum_j \left(\beta_j^2 + \beta_0^2 - 2\beta_j\beta_0 \right) \\
= & \sum_j \left(\beta_j^2 \left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right) - 2\beta_j \left(\frac{y_{+j} + \tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2} \right) \right) \\
& + \sum_i \left(\tau_i^2 \left(\frac{B}{\sigma_\epsilon^2} + \frac{1}{\sigma_\tau^2} \right) - 2\tau_i \left(\frac{y_{i+}}{\sigma_\epsilon^2} + \frac{\tau_0}{\sigma_\tau^2} \right) \right) \\
& + \left(\frac{1}{\sigma_\epsilon^2} \sum_{ij} y_{ij}^2 + \frac{B\beta_0^2}{\sigma_\beta^2} + \frac{T\tau_0^2}{\sigma_\tau^2} \right) \\
\equiv & \sum_j (T1)_j + \sum_i (T2)_i + T3.
\end{aligned} \tag{B.2}$$

Step 2: Use (B.2) to obtain

$$p(\beta_j|\mathbf{y}, \tau) \sim N \left(\frac{\frac{y_{+j} + \tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2}}{\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}}, \left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right)^{-1} \right)$$

where $\mathbf{y}$ is the matrix of y_{ij}'s, τ is the vector of τ_i's, and the subscript + indicates a sum over the appropriate index.

To see this, set up a completing the square in β_j. That is, write

$$(T1)_j = \beta_j^2 \left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right) - \frac{2\beta_j \left(\frac{y_{+j} + \tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2} \right)}{\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right)^{1/2}} \times \left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right)^{1/2}$$

$$\pm \frac{\left(\frac{y_{+j}+\tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2}\right)^2}{\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)^2} \tag{B.3}$$

$$= \left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right) \times \left(\beta_j - \frac{\frac{y_{+j}+\tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2}}{\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}}\right)^2 - ET_{1,j} \tag{B.4}$$

where $ET_{1,j}$ is the positive version of the last term in (B.3). From (B.4) we get Step 2.

Step 3: Verify that the rest of the 'active terms' in the exponent

$$\sum_i (T2)_i + \sum_j (ET)_{1,j} \tag{B.5}$$

generate a quadratic form for an appropriate matrix and vector space. With some foresight, let

$$a = \frac{B}{\sigma_\epsilon^2} + \frac{1}{\sigma_\tau^2} \quad \text{and} \quad b = -\frac{1}{\sigma_\epsilon^4(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2})}.$$

Also, write

$$v_i = \left(\frac{y_{i+}}{\sigma_\epsilon^2} + \frac{\tau_0}{\sigma_\tau^2}\right)\left(\frac{\frac{y_{++}}{\sigma_\beta^2} + \frac{B\beta_0}{\sigma_\beta^2}}{\sigma_\epsilon^2}\right)$$

and $\mathbf{v} = (v_1, \ldots, v_T)^T$). Now, the active terms from (B.5) equal

$$\sum_i \left(\tau_i^2\left(\frac{B}{\sigma_\epsilon^2} + \frac{1}{\sigma_\tau^2}\right) - 2\tau_i\left(\frac{y_{i+}}{\sigma_\epsilon^2} + \frac{\tau_0}{\sigma_\tau^2}\right)\right) - \frac{1}{\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)}\left[\sum_j \left(\frac{y_{+j} + \tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2}\right)^2\right]. \tag{B.6}$$

The expression in square brackets in (B.6) can be re-expressed as

$$\left[\frac{B(\sum_i \tau_i^2 + \sum_{k \neq i} \tau_k \tau_i)}{\sigma_\epsilon^4} + 2\sum_i \frac{\tau_i(\frac{y_{++}}{\sigma_\epsilon^2} + \frac{B\beta_0}{\sigma_\beta^2}\sigma_\beta^2)}{\sigma_\epsilon^2} + \sum_j (OT)_j\right] \tag{B.7}$$

where $(OT)_j$ represents the 'other terms' in the expansion of the expression in square brackets that do not involve τ. Using (B.7) in (B.6) gives

$$\sum_i \quad \left[\tau_i^2\left(\left(\frac{B}{\sigma_\epsilon^2} + \frac{1}{\sigma_\tau^2}\right) - \frac{\frac{B}{\sigma_\epsilon^2}}{\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}}\right) - 2\tau_i\left(\frac{y_{i+}}{\sigma_\epsilon^2} + \frac{\tau_0}{\sigma_\tau^2}\right)\left(\frac{\frac{y_{++}}{\sigma_\beta^2} + \frac{B\beta_0}{\sigma_\beta^2}}{\sigma_\epsilon^2}\right)\right.$$
$$\left. - \quad \frac{2B}{\sigma_\epsilon^4\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)}\tau_i \sum_{k \neq i} \tau_k\right]$$

$$
\begin{aligned}
&= \sum_i \left(\tau_i^2(a+Bb) - 2\tau_i v_i + 2Bb\tau_i \sum_{k\neq i} \tau_k\right) \\
&= \tau^T \begin{pmatrix} a+Bb & Bb & \dots & \dots & Bb \\ Bb & a+Bb & \dots & \dots & Bb \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Bb & Bb & \dots & a+Bb & Bb \\ Bb & Bb & \dots & Bb & a+Bb \end{pmatrix} \tau - 2\tau^T V \\
&\equiv \tau^T A + 2\tau^T \\
&= \tau^T A\tau - 2\tau^T \mu_\tau \qquad\qquad \text{(B.8)}
\end{aligned}
$$

where the matrix $\mu_\tau = -\mathbf{v}$ and A is of the form

$$A = aI_T + Bb\mathbf{1}\mathbf{1}^T. \tag{B.9}$$

Step 4: Derive the posterior variances and covariances for the τ_i's. From the Sherman-Morrison formula we have that

$$A^{-1} = \frac{1}{a}\left(I_T - \frac{Bb\mathbf{1}\mathbf{1}^T}{a+BbT}\right). \tag{B.10}$$

Continuing from (B.8) and again completing the square, this part of the exponent in the likelihood (see (B.1)) is

$$
\begin{aligned}
&= -\frac{1}{2}\left(\tau^T(A^{-1})^{-1}\tau - 2\tau\mu_\tau\right) \\
&= -\frac{1}{2}(\tau-\mu_\tau)\sigma^{-1}(\tau-\mu_\tau) + LowerOrderTerms, \qquad\qquad \text{(B.11)}
\end{aligned}
$$

for some $n\times n$ matrix Σ. Writing a_{ij} for the elements of A and $\sigma_{ij}^{(-1)}$ for the elements in Σ we see that for any i and j that

$$a_{ij}\tau_i\tau_j = \sigma_{ij}^{(-1)}\tau_i\tau_j.$$

Hence, $A = \Sigma^{-1}$ and $\Sigma = A^{-1}$ and both are symmetric and positive definite. Now, from the Sherman-Morrison formula we see that

$$\mathrm{Var}(\tau_i|\mathbf{y}) = \frac{1}{a}\left(1 - \frac{Bb}{a+BbT}\right) \tag{B.12}$$

and

$$\mathrm{Cov}(\tau_i,\tau_j|\mathbf{y}) = -\frac{Bb}{a(a+BbT)}. \tag{B.13}$$

As a check, we observe that

$$a + BbT = C\left[\left(\frac{B}{\sigma_\epsilon^2} + \frac{1}{\sigma_\tau^2}\right)\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right) - \frac{BT}{\sigma_\epsilon^4}\right]$$

for a suitable $C > 0$ and it is easy to see that the right hand side is strictly positive. So, (B.12) and (B.13) are well defined. Since $b < 0$ both are positive as well.

Step 5:. Now we can derive an expression for the posterior covariance of $\mathrm{Var}(Y_{ij;n+1}|\mathbf{y}^n)$. By two uses of the LTV we have

$$\begin{aligned}\mathrm{Var}(Y_{ij;n+1}|\mathbf{y}^n) &= E_{\tau|\mathbf{y}^n} E_{\beta|\mathbf{y}^n,\tau} \mathrm{Var}(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) \\ &\quad + E_{\tau|\mathbf{y}^n} \mathrm{Var}_{\beta|\mathbf{y}^n,\tau} E(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) \\ &\quad + \mathrm{Var}_{\tau|\mathbf{y}^n} E_{\beta|\mathbf{y}^n,\tau} E(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau). \end{aligned} \tag{B.14}$$

The first term is

$$E_{\tau|\mathbf{y}^n} E_{\beta|\mathbf{y}^n,\tau} \mathrm{Var}(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) = E_{\tau|\mathbf{y}^n} E_{\beta|\mathbf{y}^n,\tau} \left(\sigma_\epsilon^2\right) = \sigma_\epsilon^2.$$

The second term is

$$E_{\tau|\mathbf{y}^n} \mathrm{Var}_{\beta|\mathbf{y}^n,\tau} E(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) = E_{\tau|\mathbf{y}^n} \mathrm{Var}_{\beta|\mathbf{y}^n,\tau} \left(\tau_i + \beta_j\right) = \frac{1}{\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}}, \tag{B.15}$$

using the fact that i) τ_i and β_j are independent, ii) $\mathrm{Var}_{\beta|\mathbf{y}^n,\tau}(\tau_i) = 0$, and iii) the result from Step 2.

The third term in (B.14) is

$$\begin{aligned} & \mathrm{Var}_{\tau|\mathbf{y}^n} E_{\beta|\mathbf{y}^n,\tau} E(Y_{ij;n+1}|\mathbf{y}^n, \beta, \tau) \\ = \; & \mathrm{Var}_{\tau|\mathbf{y}^n} E_{\beta|\mathbf{y}^n,\tau} \left(\tau_i + \beta_j\right) \\ = \; & \mathrm{Var}_{\tau|\mathbf{y}^n} \left(\tau_i + \frac{\frac{y_{+j} + \tau_+}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2}}{\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}} \right) \\ = \; & \frac{1}{\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)^2} \mathrm{Var}_{\tau|\mathbf{y}^n} \left(\tau_i \left(\frac{T+1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right) + \left(\frac{y_{+j}}{\sigma_\epsilon^2} + \frac{\beta_0}{\sigma_\beta^2} \right) + \sum_{j \neq i}^{T} \frac{\tau_j}{\sigma_\epsilon^2} \right) \\ = \; & \frac{1}{\left(\frac{T}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)^2} \left(\left(\frac{T+1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right)^2 \cdot \frac{1}{a} \left(1 - \frac{Bb}{a + bBT} \right) + \frac{T-1}{\sigma_\epsilon^4} \frac{1}{a} \left(1 - \frac{Bb}{a + bBT} \right) \right. \\ & \left. + \frac{2}{\sigma_\epsilon^2} \left(\frac{T+1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right) (T-1) \frac{-Bb}{a(a + BbT)} \right). \end{aligned} \tag{B.16}$$

In the last term in (B.16), we have recognized $2\mathrm{Cov}(\tau_i, \tau_j|\mathbf{y}^n)$ and that the number of τ_j's not equal to a given τ_i is $T - 1$.

Appendix C: Computation

We now describe our procedure to approximately compute the three terms in the expansion described above.

Maybe argue: What we did is technically wrong, but the results are accurate because of continuity arguments???

For clarity, we denote L to be the link taking values in $\{1, 2, \dots, l\}$ and M to be the model taking values in $\{1, 2, \dots, m\}$. We estimate the terms in the expansion:

$$\begin{aligned} Var[Y_{n+1}|\mathcal{D}_n] &= E_L E_M Var[Y_{n+1}|\mathcal{D}_n, L, M] + E_L Var_M E[Y_{n+1}|\mathcal{D}_n, L, M] \\ &\quad + Var_L E[Y_{n+1}|\mathcal{D}_n, L] \end{aligned}$$

As a first step, we perform a Bayesian Model Averaging (BMA) for every link $L_i \in \{1, 2, \dots, l\}$. This is done in order to perform a model selection given the data $\mathcal{D}_n$.

Note that, we are not conditioning on the regerssion coefficients in the chosen models. Thus, in order to sample from the predictive distribution of Y_{n+1} given model $M_j \in \{1, 2, \dots, m\}$ and $L_i \in \{1, 2, \dots, l\}$, we follow the following procedure.

Let β_{ji} denote the regression coefficient for model M_j and link L_i. Clearly,

$$Pr\left[Y_{n+1}, \beta_{ji}|\mathcal{D}_n, M_j, L_i\right] = \int Pr\left[Y_{n+1}, |\mathcal{D}_n, \beta_{ji}, M_j, L_i\right] Pr\left[\beta_{ji}|\mathcal{D}_n, M_j, L_i\right] d\beta_{ji}.$$

The BMA provides the marginal posterior mean and the variance of the regression coeffiecients in the averaged model. We first simulate β_{ji} from a Gaussian density with expectation as the corresponding marginal posterior mean and the corresponding marginal posterior variance.

Next, the predicted observations of Y_{n+1} is simulated from the distribution $p[Y_{n+1}|\mathcal{D}_n, M_j, L_i, \beta_{ji}]$. Clearly, the vector (Y_{n+1}, β_{ji}) is an observation from the distribution $p[(Y_{n+1}, \beta_{ji})|\mathcal{D}_n, M_j, L_i]$. Now, discarding the β_{ji} observations we can obtaine observations from the required predictive distribution of $p[Y_{n+1}|\mathcal{D}_n, M_j, L_i]$.

From R such samples $Y_{n+1}^{(1)}$, $Y_{n+1}^{(2)}$, ..., $Y_{n+1}^{(R)}$, from $p[Y_{n+1}|\mathcal{D}_n, M_j, L_i]$, we can easily estimate the right-most expectation and the variance in the first two terms.

We define:

$$\begin{aligned} e_{ji} &= \hat{E}[Y_{n+1}|\mathcal{D}_n, M_j, L_i] = \frac{1}{R}\sum_{r=1}^{R} Y_{n+1}^{(r)}. \\ v_{ji} &= \hat{Var}[Y_{n+1}|\mathcal{D}_n, M_j, L_i] = \frac{1}{R}\sum_{r=1}^{R}\left(Y_{n+1}^{(r)} - \frac{1}{R}\sum_{k=1}^{R} Y_{n+1}^{(k)}\right)^2. \end{aligned}$$

Now suppose $p_M^{(j)} = \hat{pr}[M_j|\mathcal{D}_n, L_i]$ is the posterior probability of the model M_j given the data and the link L_i. Then we can estimate:

$$ev_i = \hat{E}_M \hat{Var}[Y_{n+1}|\mathcal{D}_n, M_j, L_i] = \sum_{j=1}^{m} p_M^{(j)} v_{ji}$$

$$ee_i = \sum_{j=1}^{m} p_M^{(j)} e_{ji}$$

$$ve_i = \hat{Var}_M \hat{E}[Y_{n+1}|\mathcal{D}_n, M_j, L_i] = \sum_{j=1}^{m} p_M^{(j)} (e_{ji} - ee_i)^2$$

In order to compute the outermost expectation and variance over the links we need to estimate the posterior probability of each link given the data.

Using Bayes rule the posterior is given by:

$$\begin{aligned} Pr[L_i|\mathcal{D}_n] &\propto Pr[\mathcal{D}_n|L_i]Pr[L_i] \\ &\propto \sum_{j=1}^{m} \int_{\beta_{ji}} Pr[\mathcal{D}_n|L_i, M_j, \beta_{ji}]Pr[\beta_{ji}|M_j, L_i]Pr[M_j|L_i]Pr[L_i]d\beta_{ji} \\ &\propto \sum_{j=1}^{m} Pr[M_j|L_i]Pr[L_i] \int_{\beta_{ji}} Pr[\mathcal{D}_n|L_i, M_j, \beta_{ji}]Pr[\beta_{ji}|M_j, L_i]d\beta_{ji} \end{aligned}$$

Note that, the distributions $Pr[\beta_{ji}|M_j, L_i]$, $Pr[M_j|L_i]$ and $Pr[L_i]$ are all priors and do not depend on the data. We make the following choices:

$$\begin{aligned} Pr[L_i] &= 1/l \;\; \text{for all } i \\ Pr[M_j|L_i] &= 1/m \;\; \text{for all } i \text{ and } j \\ Pr[\beta_{ji}|M_j, L_i] &= \delta_{\hat{\beta}_{ji}} \text{ where } \hat{\beta}_{ji} \text{ is the mle of the model } M_j \text{ and link } L_i \end{aligned}$$

With the above choices and noting that both model and link are finite, discrete random variables, we can compute the posterior probability as:

$$\begin{aligned} p_L^{(i)} = \hat{Pr}[L_i|\mathcal{D}_n] &= \frac{\sum_{j=1}^{m} Pr[M_j|L_i]Pr[L_i] \int_{\beta_{ji}} Pr[\mathcal{D}_n|L_i, M_j, \beta_{ji}]Pr[\beta_{ji}|M_j, L_i]d\beta_{ji}}{\sum_{i=1}^{l} \sum_{j=1}^{m} Pr[M_j|L_i]Pr[L_i] \int_{\beta_{ji}} Pr[\mathcal{D}_n|L_i, M_j, \beta_{ji}]Pr[\beta_{ji}|M_j, L_i]d\beta_{ji}} \\ &= \frac{\sum_{j=1}^{m} \frac{1}{m}\frac{1}{l} Pr[\mathcal{D}_n|L_i, M_j, \hat{\beta}_{ji}]}{\sum_{i=1}^{l} \sum_{j=1}^{m} \frac{1}{m}\frac{1}{l} Pr[\mathcal{D}_n|L_i, M_j, \hat{\beta}_{ji}]} \\ &= \frac{\sum_{j=1}^{m} Pr[\mathcal{D}_n|L_i, M_j, \hat{\beta}_{ji}]}{\sum_{i=1}^{l} \sum_{j=1}^{m} Pr[\mathcal{D}_n|L_i, M_j, \hat{\beta}_{ji}]}. \end{aligned}$$

With the above estimate of the posterior link probabilities the terms in the estimates can be finally estimated as follows:

$$\widehat{First-Term} = \hat{E}_L \hat{E}_M \hat{Var}[Y_{n+1}|\mathcal{D}_n, L, M] = \sum_{i=1}^{l} p_L^{(i)} ev_i$$

$$\widehat{Second-Term} = \hat{E}_L \hat{Var}_M \hat{E}[Y_{n+1}|\mathcal{D}_n, L, M] = \sum_{i=1}^{l} p_L^{(i)} ve_i$$

$$\widehat{Third - Term} = \hat{Var}_L \hat{E}[Y_{n+1}|\mathcal{D}_n, L] = \sum_{i=1}^{l} p_L^{(i)} \left(ee_i - \sum_{k=1}^{l} p_L^{(k)} ee_k \right)^2 .$$

Acknowledgments

Dustin acknowledges funding from the University of Nebraska Program of Excellence in Computational Science. All authors acknowledge computational support from the Holland Computing Center at the University of Nebraska. The third author thanks Sujit Ghosh for helpful discussions about how to use the decompositions proposed here.

References

[1] Berk, R. (1966). "Limiting Behavior of Posterior Distributions when the Model is Incorrect." *An. Math. Stat.*, 37: 51–58.

[2] Brillinger, D. (1969). "The calculation of cumulants by conditioning." *Ann. Inst. Math. Stat.*, 21: 215–218.

[3] Casella, G. and Berger, R. (2002). *Statsitical Inference 2nd Edition.* Duxbury, Australia.

[4] Draper, D. (1995). "Assessment and Propagation of Model Uncertainty." *J. R. S. S. B*, 57(1): 45–97.

[5] Dustin, D., Ghosh, S., and Clarke, B. (2025). "Testing for the Important Components of Posterior Predictive Variance." *Stat. Anal. and Data Mining*, 18.

[6] George, E. (2010). "Dilution priors: Compensating for model space redundancy." In *IMS Collections Vol. 6*, 158–165. Inst. Math. Statist.

[7] Gustafson, P. and Clarke, B. (2004). "Decomposing Posterior Variance." *J. Stat. Planning and Inference*, 119: 311–327.

[8] Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). "Bayesian Model Averaging: A Tutorial." *Statist. Sci.*, 14(4): 382–417.