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

# Uncertainty Quantification Via the Posterior Predictive Variance

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

**Abstract.** We use the law of total variance to generate multiple expressions for the posterior predictive variance. These expressions are sums of terms involving conditional expectations and conditional variances and provide a quantification of the sources of predictive uncertainty. Since the posterior predictive variance is fixed given the model, it represents a constant quantity that is conserved over these expressions. The terms in the expressions can be assessed in absolute or relative sense to understand the main contributors to the length of prediction intervals. 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

Intro

Everyone uses prediction intervals (PI's) but few examine their structure or more precisely how they should be interpreted in the context of a model with multiple components. Often PI's seem overconfident (too narrow) or useless (too wide). More often than not, PI's are an afterthought to modeling rather than the focus: most sample size selection procedures, for instance, focus on estimation or testing, not prediction.

Both frequentist and Bayesian practitioners routinely report PIs. It is common for frequentists to estimate a model and then use it, perhaps even without adjustment, to give PI's; see Shen et al. (2004) and more recently Bachoc et al. (2019), Tian (2020), and Liang et al. (2025). By contrast, it is common for Bayesians to give a PI simply by simulation from the posterior predictive distribution and report the posterior predictive variance (PPV) itself as a scalar summary without making the relationship between the width of the PI and the PPV explicitly. Here, instead, we treat the PPV as the quantity of interest. The goal of our discussion is therefore to pull together many existing ideas about the PPV in a relatively complete and organized way so they can be seen as a coherent body of material.

Analogous to classical components-of-variance models, e.g., split plot designs, we use the Law of Total Variance (LTV) to expand the PPV into interpretable contributions. We implicitly adopt a Bayesian standpoint not because we accept it (although we do) but because of its Containment Principle: all relevant distributions exist and are 'contained'

[*]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

in a single measure space. Nonetheless, readers who prefer non-Bayesian formulations may regard our parameters simply as random variables.

Looking at a variance brings in the metric properties of a distribution. The difference between a PPV and simulating a PI is that the former relies on posterior normality and the squared error distances between random variables and their means – essentially regarding the mean as a constant random variable whereas the latter only uses probabilities. To see this, consider a generic two-level example. Suppose

$$\begin{aligned} Z &\sim w(z) \\ Y^n = (Y_1, \ldots, Y_n) &\sim p(y|z) \end{aligned} \tag{1.1}$$

where $Y = Y^n$ is independent and identical distributed (IID) data and $Z$ is a conditioning variable e.g., a unidimensional parameter. Denoting $n$ outcomes by $y^n = (y_1, \ldots, y_n)^T$, the conditional distribution $(Y_{n+1}|y^n)$ will give PI's. Alternatively, the PPV is $\mathsf{Var}(Y_{n+1}|y^n)$, and the LTV gives

$$\mathsf{Var}(Y_{n+1}|y^n) = \mathsf{E}\mathsf{Var}_{Z|y^n}(Y_{n+1}|y^n, Z) + \mathsf{Var}_{Z|y^n}\mathsf{E}(Y_{n+1}|y^n, Z) \tag{1.2}$$

which can be interpreted. Loosely, the first term on the right in (1.1) is the variability from the likelihood and the second term on the right is the variability from $w$. Viewing predictive uncertainty via this expansion clarifies why PI's have the width they do. In this case, if the conditional distributions $p(\cdot|z)$ concentrate tightly in the space of densities, the first term may be negligible, implying the hierarchy effectively has lower dimension than it appears. On the other hand, if the conditional means $\mathsf{E}(Y_{n+1}|y^n, Z)$ vary little across $Z$, the second term may be small. We can inadvertently (or artificially) make a term small by choosing the densities in a parametric family to be close to each other but not identical, essentially a metric property.

These ideas extend naturally to hierarchical models of arbitrary depth. Consider a hierarchical model (HM) for a response $Y = y$ given $\mathcal{Z} = (Z_1, \ldots, Z_k, \ldots, Z_K)^T$ taking values $z = (z_1, \ldots, z_K)^T$ for some $K \in \mathbb{N}$:

$$\begin{aligned} Z_1 &\sim w(z_1) \\ Z_2 &\sim w(z_2|z_1) \\ \vdots \;&\;\vdots\;\; \vdots \\ Z_K &\sim w(z_K|z_1, \ldots, z_{K-1}) \\ Y^n &\sim p(y|z), \end{aligned} \tag{1.3}$$

where the $w$'s represent distributions for the $Z_k$'s as indicated by their arguments and $p(\cdot|z)$ is the likelihood. Again, $n$ IID copies of $Y$ are denoted by $Y^n = (Y_1, \ldots, Y_n)^T$ with outcomes $y^n = (y_1, \ldots, y_n)^T$.

Sequential application of the LTV to (1.3) produces $K+1$ variance components, each interpretable in terms of contributions from different levels of the hierarchy. Because (1.3) satisfies the Containment Principle, it is straightforward to see how assumptions on conditional distributions affect the expansion of the PPV. By tracking the relative sizes of these components, we obtain a common scale for all sources of uncertainty, allowing coherent comparison of contributions from different levels. In particular, we will track how one term being zero in one sequence of uses of the LTV can imply how a related term is zero in another sequence of uses of the LTV.

This paper provides a conceptual framework for understanding predictive uncertainty via LTV expansions of the PPV. This framework applies broadly, offering insight into the components that determine PI length and the interpretation of predictive statements in practice. Our goal is not to introduce new mathematical theory or computational methods, but to organize and clarify existing ideas. We focus on showing how these expansions help interpret PI's and the structure of predictive inference more generally.

À propos of this, it is well-recognized that prediction requires calibration and often re-calibration; see Qian:etal:2025a Qian et al. (2025b) and Qian:etal:2025b Qian et al. (2025a) and the references therein for recent contributions in regression and classification respectively. We dodge this question here because it is not immediately germane to our analysis of the PPV.

This paper proceeds as follows. In Sec. 2 we present a series of examples that illustrate many of the properties of expansions of the PPV for models like (1.3). In Sec. 3, we develop properties of the use of LTV expansions for two and three-term cases. In Sec. 4, we discuss the use of more general expansions for uncertainty quantification. In Sec. 5 we present some computational work comparing how the terms in a three-term expansion behaves as functions of its inputs along with a data-driven example of this. In Sec. 6, we discuss the implications and uses of these expansions for UQ. Details of derivations are relegated to Appendices A and B.

# 2 Two-term and Three-term Expansions of Posterior Predictive Variance

sec:2

Let $\mathcal{D}$ denote the available data, which includes $y^n$, and any other covariate that might be available. Given $\mathcal{D}$, the posterior predictive density to future values $Y_{n+1}$, that is

$$Y \sim p(y_{n+1}|\mathcal{D}) = \int p(y_{n+1}|v)w(v|\mathcal{D})\mathrm{d}v, \tag{2.1}$$

where $w(v|\mathcal{D})$ is the posterior density. At this point, the PPV within the context of the model (1.3) is fixed. Denote it by $\mathrm{Var}(Y_{n+1}|y^n)$. When a random variable in the top $K$ levels of the hierarchy are 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 (1.3).

## 2.1 Two-term Expansions

sec:twoTerm

Let $V_1 \in \mathcal{Z}$. From the Law of Total Variance (LTV) it immediately follows that:

$$\mathrm{Var}_{Y_{n+1}|\mathcal{D}}(Y_{n+1}|\mathcal{D}) = E_{V_1|\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_1,\mathcal{D})] + \mathrm{Var}_{V_1|\mathcal{D}}[E(Y_{n+1}|V_1,\mathcal{D})]. \tag{2.2}$$

postpredvarV1

We define the above expansion as a two-term expansion of the PPV conditional on $V_1$ and $\mathcal{D}$. The expansion easily extends to subsets $V_1 \subseteq \mathcal{Z}$ of size larger than one.

We consider a few common examples of two-term expansions from the parametric Bayesian Hierarchical models.

**Example 2.1.** *Consider using the Law of Total Variance (LTV) on the posterior predictive variance (PPV) from a normal likelihood with a conjugate prior, i.e.,*

$$\mathsf{Var}(Y_{n+1}|y^n) = \mathsf{E}_{\mu|y^n}\mathsf{Var}(Y_{n+1}|y^n,\mu) + \mathsf{Var}_{\mu|y^n}\mathsf{E}(Y_{n+1}|y^n,\mu) \tag{2.3}$$

*where $\mu \sim N(\mu_0, \tau_0^2)$ and the $Y_i$'s are independently and identically distributed (IID) as $\mathsf{N}(\mu, \sigma_0^2)$, where $\mu_0$, $\sigma_0$ and $\tau_0$ are known. Here, $y^n = (y_1, \ldots, y_n)^T$ is an outcome of the Vector $Y^n = (Y_1, \ldots, Y_n)^T$. It is easy to see that $p(y_{n+1}|y^n,\mu) = p(y_{n+1}|\mu)$, where $\sigma_0$ has been suppressed in the notation. So, it is also easy to see that*

$$\mathsf{E}(Y_{n+1}|y^n,\mu) = \mu \quad \textit{and} \quad \mathsf{Var}(Y_{n+1}|y^n,\mu) = \sigma_0^2.$$

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

$$\mathsf{E}_{\mu|y^n}\mathsf{Var}(Y_{n+1}|\mu, y^n) = \sigma_0^2.$$

*For the second term on the right in (2.3) recall the posterior for $\mu$ given $y^n$ is $\mathsf{N}(\mu_n, \tau_n^2)$ where*

$$\mu_n = \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1} \frac{n}{\sigma_0^2}\left(\bar{y} + \frac{\mu_0}{\tau_0^2}\right) \quad \textit{and} \quad \tau_n^2 = \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1}.$$

*Now,*

$$\mathsf{Var}_{\mu|y^n}(\mathsf{E}(Y_{n+1}|\mu, y^n)) = \mathsf{Var}_{\mu|y^n}(\mu) = \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1}.$$

*So, (2.3) is*

$$\mathsf{Var}(Y_{n+1}|y^n) = \sigma_0^2 + \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1} = \sigma_0^2 + \mathcal{O}(1/n),$$

*in which the 'E-Var' term dominates asymptotically. The first term is the intrinsic variance of $Y_{n+1}$ and the second term is the extra variability due to not knowing $\mu$.* □

exm:expF **Example 2.2.** *It is easy to generalize (2.3) to a one dimensional exponential family for $Y$, say $h(y)e^{\eta T(y) - \phi(\theta)}$ on some domain, equipped with its conjugate prior. Indeed, some routine calculations show*

$$\mathsf{Var}(Y_{n+1}|y^n) = \mathsf{E}_{\theta|y^n}\phi''(\theta) + \mathsf{Var}_{\theta|y^n}\phi'(\theta). \tag{2.4}$$

*Again, the first term represents intrinsic variability in $Y_{n+1}$ and the second term represents the extra uncertainty from not knowing $\theta$.* □

**Example 2.3.** *This applies, for instance, to the Beta-Binomial problem. Let* $Y_i \sim Bin(m_i, p)$ *where* $\sum_{i=1}^n m_i = M$*, we set* $S = \sum_{i=1}^n Y_i$*, and* $p \sim \mathsf{Beta}(\alpha, \beta)$*. Conjugacy gives that the posterior for* $p$ *is* $(p|Y^n) \sim \mathsf{Beta}(\alpha_n, \beta_n)$ *where* $\alpha_n = \alpha + S$ *and* $\beta_n = \beta + M - S$*. Now, the posterior predictive distribution for* $(Y_{n+1}|Y^n)$ *is, by direct calculation,* $\mathsf{Beta-Binomial}(m_{n+1}, \alpha_n, \beta_n)$*, with variance*

$$\mathsf{Var}(Y_{n+1}|Y^n) = m_{n+1}\frac{\alpha_n\beta_n}{(\alpha_n+\beta_n)^2}\frac{(\alpha_n+\beta_n+m_{n+1})}{(\alpha_n+\beta_n+1)}. \tag{2.5}$$

*We get the same result from evaluating the two terms in (2.4). For* $\theta = \log(p/(1-p))$*, the canonical form is* $P(Y_{n+1} = y_{n+1}|\theta) = C(m_{n+1}, y_{n+1})e^{y\theta - m_{n+1}\log(1+\exp(\theta))}$*, so* $\phi(\theta) = m_{n+1}\log(1+e^\theta)$*. The general form of the conjugate prior is* $w(\theta|\alpha,\beta) \propto e^{\alpha\theta - \beta\phi(\theta)}$ *giving the posterior* $w(\theta|y^n) \propto e^{(\alpha+s)\theta - (\beta+M)\log(1+\exp(\theta))}$*. Since* $\phi'(\theta) = m_{n+1}p$ *and* $\phi''(\theta) = m_{n+1}p(1-p)$*, we get*

$$\mathsf{E}_{\theta|y^n}\phi''(\theta) = \mathsf{E}(m_{n+1}p(1-p)|y^n) = m_{n+1}\frac{\alpha_n\beta_n}{\alpha_n+\beta_n+1}$$

*and*

$$\mathsf{Var}_{\theta|y^n}\phi'(\theta) = \mathsf{Var}(m_{n+1}p|y^n) = m_{n+1}^2\frac{\alpha_n\beta_n}{(\alpha_n+\beta_n)^2(\alpha_n+\beta_n+1)}$$

*which, upon summing and re-arranging, gives (2.5).* □

**Example 2.4.** *In canonical exponential family form, the* $\mathsf{Poisson}(\lambda)$ *is*

$$P(Y_i = y_i|\theta) = (1/y!)e^{y\theta - \exp(\theta)},$$

*so* $\phi(\theta) = e^\theta = \lambda$*. The general form of the conjugate prior in* $\theta$ *is* $w(\theta) \propto e^{\alpha\theta - \beta\exp\theta}$ *giving the posterior* $w(\theta|y^n) \propto e^{(\alpha+S)\theta - (\beta+n)\exp(\theta)}$*. By direct calculation, for* $s = \sum y_i$*,*

$$\mathsf{Var}(Y_{n+1}|y^n) = \frac{\alpha+s}{\beta+n}\left(1+\frac{1}{\beta+n}\right). \tag{2.6}$$

*Since* $\phi'(\theta) = \phi''(\theta) = e^\theta$*, we get that the PPV is the sum of*

$$\mathsf{E}(\phi''(\theta)|y^n) = \frac{\alpha+s}{\beta+n} \quad \text{and} \quad \mathsf{Var}(\phi'(\theta)|y^n) = \frac{\alpha+s}{(\beta+n)^2}$$

*that gives (2.6). The Poisson(*$\lambda$*) distribution with a Gamma prior (giving a negative binomial posterior) can also be worked out explicitly.* □

sec:threeTerm

## 2.2 Three-term Expansions

As the three foregoing examples indicate, our focus here is to develop and study expansions of the PPV. As will be seen, our goal is to find ways to reduce the number of terms in multi-term expansions to eliminate unnecessary conditioning variables.

To see that this is possible, let $\{V_1, V_2\} \subseteq \mathcal{Z}$, and extend (2.3) by a second use of the LTV in the first term,

$$\mathrm{Var}(Y_{n+1}|v_1,\mathcal{D}) = E_{V_2|v_1,\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_1,V_2,\mathcal{D})] + \mathrm{Var}_{V_2|v_1,\mathcal{D}}[E(Y_{n+1}|V_1,V_2,\mathcal{D})]. \tag{2.7}$$

LTV2termgen

By substituting the above expression in (2.2), one gets the *three-term* expansion of the PPV as:

$$\begin{aligned}\mathsf{Var}(Y_{n+1}|\mathcal{D}) &= \mathsf{E}_{V_1|\mathcal{D}}\mathsf{E}_{V_2|V_1,\mathcal{D}}[\mathsf{Var}(Y_{n+1}|V_1,V_2,\mathcal{D})] &\text{(2.8a)}\\ &+ \mathsf{E}_{V_1|\mathcal{D}}\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1,V_2,\mathcal{D})] &\text{(2.8b)}\\ &+ \mathsf{Var}_{V_1|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1,\mathcal{D})]. &\text{(2.8c)}\end{aligned}$$

eq:1a eq:1b eq:1c

Note that we could have used a two-term expansion for the PPV simply by using only one of the two conditioning variables. The order of conditioning, however, matters. The three-term expansion defined in (2.8) depends on the sequence in which we condition on $V_1$ and $V_2$. That is, by conditioning on $V_2$ first and then on $V_1$, we can alternatively write (2.8) as:

LTV2termgen2

$$\begin{aligned}\mathrm{Var}(Y_{n+1}|\mathcal{D}) &= E_{V_2|\mathcal{D}}E_{V_1|V_2,\mathcal{D}}[\mathrm{Var}(Y_{n+1}|V_2,V_1,\mathcal{D})] &\text{(2.9a)}\\ &+ E_{V_2|\mathcal{D}}\mathrm{Var}_{V_1|V_2,\mathcal{D}}[E(Y_{n+1}|V_2,V_1,\mathcal{D})] &\text{(2.9b)}\\ &+ \mathrm{Var}_{V_2|\mathcal{D}}[E(Y_{n+1}|V_2,\mathcal{D})]. &\text{(2.9c)}\end{aligned}$$

eq:2a eq:2b eq:2c

Terms in the RHS of equations (2.8) are not equal to the corresponding terms in (2.9). Furthermore, the fact that any of these terms is zero does not reduce the three-term expansion to a valid two-term expansion as defined in (2.2).

We consider some illustrative examples of three-term expansions below:

exm:NNG

**Example 2.5.** *Again, consider a normal likelihood, but this time with a normal prior on the mean and a Gamma prior on the variance. More precisely, let* $Y_i \sim \mathsf{N}(\mu,\lambda^2)$ *be IID for* $i = 1,\dots,n$ *and use the conjugate priors* $\mu \sim \mathsf{N}(\mu_0, 1/(\kappa_0\lambda^2))$ *with* $\lambda^2 \sim \mathsf{Gamma}(\alpha_0,\beta_0)$*. Now we have two three-term expansions depending on whether we condition on* $\mu$ *first (and then* $\lambda$*) or* $\lambda$ *first (and then* $\mu$*).*

*Conditioning on* $\mu$ *first, i.e., setting* $V_1 = \lambda^2$ *and* $V_2 = \mu$ *in* (2.8) *we get*

$$\begin{aligned}\mathsf{Var}(Y_{n+1}|y^n) &= \mathsf{E}_{\lambda^2|y^n}\mathsf{E}_{\mu|y^n,\lambda^2}\mathsf{Var}(Y_{n+1}|y^n,\mu,\lambda^2) &\text{(2.10a)}\\ &+ \mathsf{E}_{\lambda^2|y^n}\mathsf{Var}_{\mu|y^n,\lambda^2}\mathsf{E}(Y_{n+1}|y^n,\mu,\lambda^2) &\text{(2.10b)}\\ &+ \mathsf{Var}_{\lambda^2|y^n}\mathsf{E}_{\mu|y^n,\lambda^2}\mathsf{E}(Y_{n+1}|y^n,\mu,\lambda^2), &\text{(2.10c)}\end{aligned}$$

EEVar EVarE VarEE

*in which it is easy to see that* (2.10c) *is zero:*

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

*For (2.10a) and (2.10b), we use the fact that, by conjugacy, there is an $\alpha_n$ and $\beta_n$ so that $\lambda^2|y^n \sim \mathsf{Gamma}(\alpha_n, \beta_n)$. This gives that*

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

*Now, dropping the conditioning on $y^n$ in the variance on the right of (2.10a) it is*

$$\mathsf{E}_{\lambda^2|y^n}\mathsf{E}_{\mu|y^n,\lambda^2}\mathsf{Var}(Y_{n+1}|\mu,\lambda^2) = \mathsf{E}_{\lambda^2|y^n}\mathsf{E}_{\mu|y^n,\lambda^2}\left(\frac{1}{\lambda^2}\right) = \frac{\beta_n}{\alpha_n - 1}. \tag{2.11}$$

*Likewise, we can show that for $\kappa_n = n + \kappa_0$, (2.10b) is*

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

*Thus, we have that*

$$\mathsf{Var}(Y_{n+1}|y^n) = \left(\frac{\kappa_n + 1}{\kappa_n}\right)\frac{\beta_n}{\alpha_n - 1}. \tag{2.12}$$

*If we condition on $\lambda$ first and then $\mu$ we find that*

$$\begin{aligned}\mathsf{Var}(Y_{n+1}|y^n) \quad = \quad & \mathsf{E}_{\mu|y^n}\mathsf{E}_{\lambda^2|y^n,\mu}\mathsf{Var}(Y_{n+1}|y^n,\mu,\lambda^2) & (2.13)\\ & +\mathsf{E}_{\mu|y^n}\mathsf{Var}_{\lambda^2|y^n,\mu}\mathsf{E}(Y_{n+1}|y^n,\mu,\lambda^2) & (2.14)\\ & +\mathsf{Var}_{\mu|y^n}\mathsf{E}(Y_{n+1}|y^n,\mu). & (2.15)\end{aligned}$$

*Parallel to (2.10c), it is easy to see that (2.14) is zero. By Fubini, (2.13) is the same as (2.10a) as given by (2.11). Finally, since Var$(Y_{n+1}|y^n)$ is constant independent of the condition, we can solve for (2.15). If desired, we can calculate Var$(Y_{n+1}|y^n)$ directly and hence verify (2.12). 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 expansion, the VarEE term is zero, whereas in the second expansion, the EVarE term is zero. Finally, in the first, the EVarE term has the $\kappa_n$ while in the second, the VarEE term has the $\kappa_n$. In particular, this shows that it is not a priori clear which terms will dominate in three-term expansions.

We see that the interpretation of the two-term expansion (intrinsic variability plus extra parameter uncertainty) generalizes to the three term case. The first term continues to represent the intrinsic variability, but the second and third terms summarize the extra variability due to parameter uncertainty – the second term for the first conditioning parameter and the third term for the second conditioning parameter.

exm:BPG **Example 2.6.** *As a second three-term example, suppose the $Y_i$'s are IID $\mathsf{Bin}(N_i, p)$ where $p$ is fixed and the $N_i$'s are drawn IID from a $\mathsf{Poisson}(\lambda)$ and $\lambda \sim \mathsf{Gamma}(a, b)$. It is not hard to verify that*

$$\mathsf{Var}(Y_{n+1}|y^n) = p\frac{\sum_{i=1}^n y_i + a}{b + pn}\left(1 + \frac{p}{b + pn}\right) \tag{2.16}$$

*and, as in the normal case, there are two 3-term expansions depending on the order of conditioning on the $N_i$'s and $\lambda$. The more natural ordering conditions on $N$ first:*

$$\mathsf{E}_{\lambda|y^n}\mathsf{E}_{N|y^n,\lambda}\mathsf{Var}(Y_{n+1}|y^n,\lambda,N) \tag{2.17}$$
$$+\mathsf{E}_{\lambda|y^n}\mathsf{Var}_{N|y^n,\lambda}\mathsf{E}(Y_{n+1}|y^n,\lambda,N) \tag{2.18}$$
$$+\mathsf{Var}_{\lambda|y^n}\mathsf{E}(Y_{n+1}|y^n,\lambda). \tag{2.19}$$

*Let $s = \sum y_i$. The corresponding terms are*

$$p(1-p)\frac{s+a}{b+pn} + p^2\frac{s+a}{b+pn} + p^2\frac{s+a}{(b+pn)^2}$$

*that clearly sum to (2.16). In order, the terms represent binomial, Poisson, and Gamma variability, and all terms are functions of $s$ with the last term beinf smaller order in $n$ than the first two. If the model is believed, one can plot curves for the four terms as a function of $s$ to see which, if any, are small enough to be ignored.*

*If we condition in the reverse order, the $N$'s can be integrated out to give a* Poisson-Gamma *model, i.e., it reduces to a two-term expansion because the term parallel to (2.18) is zero. We have that the PPV is*

$$\mathsf{E}_{N|y^n}\mathsf{E}_{\lambda|y^n,N}\mathsf{Var}(Y_{n+1}|y^n,\lambda,N) + \mathsf{Var}_{N|y^n}\mathsf{E}(Y_{n+1}|y^n,N)$$
$$= \left(p(1-p)\frac{s+a}{b+pn}\right) + \left(p^2\frac{s+a}{b+pn} + p^2\frac{s+a}{(b+pn)^2}\right).$$

If we redo the calculations with a single fixed $N$ chosen at the beginning according to a $\mathsf{Gamma}(a,b)$, we get the same PPV and terms. The reason is that the terms only depend on $s$ and this updates the Poisson rate $p\lambda$ the same way in both cases.

**Example 2.7.** *Here, we see that the Bayes model average (BMA) can be seen as a two or three-term hierarchical model. 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 variable, 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 hierarchical model*

$$\begin{aligned}(J,\theta_J) &\sim w(\theta_j,j)\\ Y &\sim p(y|\theta_j).\end{aligned} \tag{2.20}$$

*Now, the $L^2$ BMA predictor is*

$$E(Y_{n+1}|y^n) = \sum_{j=1}^{J} E(Y_{n+1}|y^n,M_j)W(M_j|y^n). \tag{2.21}$$

*In (2.21), the two conditioning random variables, namely $J$ and $\theta_j$ 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 PPV. Indeed, using the expression for posterior variance from p. 383 of Hoeting et al. (1999), we find that (2.21) is*

$$\begin{aligned} Var(Y_{n+1}|y^n) &= \sum_{j=1}^{J} Var(Y_{n+1}|y^n, M_j) W(M_j|y^n) \\ &\quad + \sum_{j=1}^{J} E(Y_{n+1}|y^n, M_j)^2 W(M_j|y^n) - E(Y_{n+1}|y^n)^2 \\ &= E(\, Var(Y_{n+1}|M_J, y^n)|y^n) + Var(E(Y_{n+1}|M_J, y^n)). \end{aligned} \tag{2.22}$$

*So (2.22) is the result of using the LTV and conditioning on $M_k$. We have treated $\theta_j$ implicitly by integrating over it, before conditioning on the $M_j$'s. Reversing this, i.e., integrating over $j$ and using the LTV with $\Theta_k$'s would have been mathematically well-defined but statistically inappropriate for BMA.*

*When the first term on the right 'E-Var' in (2.22) 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 all the high posterior probability models are good.*

*Now write (2.20) as an equivalent three-level hierarchical model:*

$$\begin{aligned} J &\sim w(j) \\ \theta_j|J = j &\sim w(\theta_j|j) \\ Y &\sim p(y|\theta_j). \end{aligned} \tag{2.23}$$

*Now, using (2.8) with the choice of $V_1$ and $V_2$ that is more natural gives*

$$\begin{aligned} Var(Y_{n+1}|y^n) &= E_J E_{\Theta_J|y^n, M_J} Var_{Y_{n+1}|y^n, M_J \Theta_j}(Y_{n+1}|y^n, M_J, \Theta_J) \\ &\quad + E_J Var_{\Theta_J|y^n, M_J} E_{Y_{n+1}|y^n, M_J, \Theta_J}(Y_{n+1}|y^n, M_J, \Theta_J) \\ &\quad + Var_J(E(Y_{n+1}|M_J, y^n)). \end{aligned} \tag{2.24}$$

*In this treatment of PPV, the relative size of the terms is a tradeoff among the size of the 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.*

Here, we limit ourselves to iteratively using the LTV in the 'E-Var' term. There is nothing to stop us from applying the LTV in the 'Var E' term as well; however, such terms are very difficult to handle. That is, while the full 'LTV-scope' of a PPV contains many terms from using the LTV in all possible ways, we focus on the subset of the LTV-scope where each term has exactly one variance operation that moves from left to right with appropriate conditioning. We call this the Cochran Scope or C-Scope for short. This terminology recognized the analogy between a set of LTV expansions and the sums of squares decompositions that arise in frequentist ANOVA; see Dustin et al. (2025). Henceforth, we limit our attention to the C-scope's of a HM's.


Figure 1: Schematic diagram of a Hidden Markov model. With the definitions of $V_1$, $V_2$ and $\mathcal{D}$, the conditional independence relationships, $V_1 \perp\!\!\!\perp V_2 \mid \mathcal{D}$ and $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_2, \mathcal{D})$ hold. MAKE NOTATION HERE CONSISTENT WITH TEXT and (1.3). fig:hmm

Although the examples of HM's we have seen so far are 'vertical' in the sense that each $Z_k$ in (1.3) sits 'above' $Z_{k+1}$, this is not necessary for our expansions. Indeed, consider the following 'horizontal' example diagrammed in Fig. 1.

exm:hmm **Example 2.8.** *Consider a Hidden Markov model where we observe $X_1, \dots, X_n$ assumed to be generated from the hidden outcomes of a Markov process $Y_1, \dots, Y_n$ respectively. The problem is to predict $Y_{n+1}$ using the earlier $Y_i$'s as our 'data' $\mathcal{D}$ even though they are unobserved. Since the $Z_k$'s in (1.3) are simply random variables, not necessarily parameters, we can use them in a three-term PPV expansion analogous to (2.8a)-(2.8c) or (2.9a)-(2.9c) for $\mathsf{Var}(Y_{n+1}|y^n)$.*

*If we use $X_1, \dots, X_n$ as $Z_1$ and $X_{n+1}$ as $Z_2$, then by construction, each $X_i$ depends only on its $Y_i$ for $i = 1, \dots, n+1$. Moreover, we have $(Z_1 \perp\!\!\!\perp Z_2|\mathcal{D})$ and $(Y_{n+1} \perp\!\!\!\perp Z_2|\mathcal{D}, Z_1)$. Looking ahead, using Theorem 3.4 it immediately follows that term (2.8b) in (2.8) and term (2.9b) in (2.8) are zero, and the three-term expansions reduce to a two-term expansion involving $V_2$.*

*Furthermore, suppose we have predicted values $\hat{Y}_i$ of $Y_i$ that are functions of the $X_1$, $X_2$, $\dots$, $X_i$. Let $\mathcal{D} = \{\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n\}$. Even then, the conditions $V_2 \perp\!\!\!\perp V_1 \mid \mathcal{D}$ and $Y_{n+1} \perp\!\!\!\perp Z_1 \mid (Z_2, \mathcal{D})$ hold, and Theorem 3.4 applies. That is, from a prediction standpoint, we are setting $\widehat{\mathsf{Var}}(Y_{n+1}) = \mathsf{Var}[Y_{n+1}|\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n]$.*

The above procedure can be iterated in multiple ways to include many other $Z_k$'s, each usage of the LTV generating an extra term in the PPV. Prop. 4.1 counts how many there are as a function of $K$. The PPV is constant over all these expansions and so represents a 'conservation of variance law'.

sec:quant

# 3 UQ for Two and Three-term Expansions

In this section, we look at the behaviour of the individual terms in two- and three-term expansions. Our goal is to identify conditional independence assumptions – that we call structural – under which some of the terms in the RHS of equations (2.2) and (2.8) are small, perhaps zero. One implicit goal is to see if a level the HM can be dropped, as a consequence of dropping terms in the expansion. For simplicity, we assume that both $V_1$

and $V_2$ are univariate. It will be seen that these results generalize. Indeed, three term expansions exhibit all possible generic characteristics of expansions for $K$ level HM's because of the '$\mathsf{E}^{\mathsf{m}} - \mathsf{Var}_{\mathsf{q}} - \mathsf{E}^{\ell}$' (where $m + q + \ell = K$) structure of the terms.

By an application of Fubini's Theorem, it is easy to see that the leading terms in the RHS of (2.8) and (2.9) are equal: for any $V_1$, $V_2$:

$$\begin{aligned} \mathsf{E}_{V_1|\mathcal{D}}\mathsf{E}_{V_2|V_1,\mathcal{D}}[\mathsf{Var}(Y_{n+1}|V_1,V_2,\mathcal{D})] &= \mathsf{E}_{V_2|\mathcal{D}}\mathsf{E}_{V_1|V_2,\mathcal{D}}[\mathsf{Var}(Y_{n+1}|V_2,V_1,\mathcal{D})] \\ &= \mathsf{E}_{V_1,V_2|\mathcal{D}}[\mathsf{Var}(Y_{n+1}|V_2,V_1,\mathcal{D})]. \end{aligned} \tag{3.1}$$

Indeed, in any valid modeling setting, this term will be strictly positive and typically no other term will be asymptotically larger as $n$ increases.

`condlindpdce`

## 3.1 UQ under Structural Assumptions

Here we give a sufficient condition for the reduction of three-term expansions two term expansions. Then, we also show that, at least in the normal case, the higher the level in the hierarchical model, the larger the conditional variances are.

`asmp:condInd`

**Assumption 1.** *$Y_{n+1}$ and $\mathcal{D}$ are conditionally independent of $V_2$ given $V_1$, i.e.,*

$$(Y_{n+1}, \mathcal{D}) \perp\!\!\!\perp V_2 \mid V_1. \tag{3.2}$$ `condInd`

We call this structural because it must hold for all values of $V_1$ and $V_2$. That is, the condition (3.2) should be satisfied in any HM with $V_1$ being a parameter and $V_2$ being a hyperparameter; see Fig. 2.

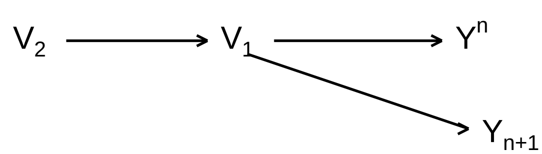



Figure 2: Diagram for a three level HM. $V_1$ is a parameter and $V_1$ is a hyperparameter. Further hyper-hyper-parameters such as $V_3$ would extend the diagram to the left. Note that we think of this as a vertical model because it is hierarchical but we have displayed it horizontally for convenience. `fig:horizontal`

`thm:struct`

**Theorem 3.1.** *Suppose Assumption 1 holds. Then, the term (2.8b) in the three-term expansion (2.8) is zero. Furthermore, the expansion (2.8) reduces to a two-term expansion conditional on $V_1$ and $\mathcal{D}$.*

*Proof.* Assumption 1 is equivalent to $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$ and $\mathcal{D} \perp\!\!\!\perp V_2 \mid V_1$ taken together. From the first we get

$$\mathsf{E}(Y_{n+1}|V_1,V_2,\mathcal{D}) = \mathsf{E}(Y_{n+1}|V_1,\mathcal{D}) \quad \text{and} \quad \mathsf{Var}(Y_{n+1}|V_1,V_2,\mathcal{D}) = \mathsf{Var}(Y_{n+1}|V_1,\mathcal{D}).$$

Since $\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})$ is independent of $V_2$,

$$\mathsf{E}_{V_1}\mathsf{Var}_{V_2}E(Y_{n+1}|\mathcal{D}_n, V_1, V_2) = \mathsf{E}_{V_1|\mathcal{D}}\mathsf{Var}_{V_2|V_1,\mathcal{D}}E(Y_{n+1}|V_1, \mathcal{D}) = 0.$$

Hence,

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

□

In the context of Fig. 2, Theorem 3.1 shows that conditioning only on the parameter nearest the data is enough. That is, given the parameters, the hyperparameters don't matter. In fact, the proof of Theorem 3.1 shows that it is enough for $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$ to hold so Assumption 1 is sufficient, but not necessary. On the other hand, Assumption 1 holds in HM, but not, for instance, in a hidden Markov model.

Now, in the special case of normality, we can show explicitly that as you go up a hierarchy satisfying Assumption 1 the conditional variances increase. This means that as the unknown quantity gets further and further from the data, the data say less and less about it. We have the following.

thm:var **Theorem 3.2.** *Suppose the conditional independence relation in (3.2) holds. Assume that the variables are jointly distributed as a multivariate Gaussian density. Then for any choice of the parameters:*

1. $Var(Y_{n+1} \mid V_2, \mathcal{D}) \geq Var(Y_{n+1} \mid V_1, \mathcal{D})$.
2. $Var_{V_1|\mathcal{D}}[E(Y_{n+1}|V_1, )] \geq Var_{V_2|\mathcal{D}}[E(Y_{n+1}|V_2, \mathcal{D})]$

*Proof.* Suppose $\Sigma$ is the variance-covariance matrix of all variables and we treat $\mathcal{D}$ as a multivariate component. We start with the first clause.

Using the formula of conditional covariance for multivariate normals, we get:

$$\sigma_{Y_{n+1}V_2|V_1} = \sigma_{Y_{n+1}V_2} - \frac{\sigma_{Y_{n+1}V_1}\sigma_{V_1V_2}}{\sigma_{V_2V_2}}$$

. Now if $Y_{n+1} \perp\!\!\!\perp V_2 \mid V_1$, $\sigma_{Y_{n+1}V_2|V_1} = 0$, so we get

$$\sigma_{Y_{n+1}V_2} = \frac{\sigma_{Y_{n+1}V_1}\sigma_{V_1V_2}}{\sigma_{V_2V_2}}. \tag{3.3}$$ eq:cond

Next, note two identities:

$$\begin{aligned}\sigma_{Y_{n+1}Y_{n+1}|V_2\mathcal{D}} &= \sigma_{Y_{n+1}Y_{n+1}|\mathcal{D}} - \frac{\sigma^2_{Y_{n+1}V_2|\mathcal{D}}}{\sigma_{V_2V_2|\mathcal{D}}} \\ \sigma_{Y_{n+1}Y_{n+1}|V_1\mathcal{D}} &= \sigma_{Y_{n+1}Y_{n+1}|\mathcal{D}} - \frac{\sigma^2_{Y_{n+1}V_1|\mathcal{D}}}{\sigma_{V_1V_1|\mathcal{D}}}.\end{aligned}$$

Now, it is enough to show that

$$\sigma_{Y_{n+1}Y_{n+1}|V_2\mathcal{D}} - \sigma_{Y_{n+1}Y_{n+1}|V_1\mathcal{D}} = \frac{\sigma^2_{Y_{n+1}V_1|\mathcal{D}}}{\sigma_{V_1V_1|\mathcal{D}}} - \frac{\sigma^2_{Y_{n+1}V_2|\mathcal{D}}}{\sigma_{V_2V_2|\mathcal{D}}} \geq 0.$$

We have that

$$\sigma_{Y_{n+1}V_2|\mathcal{D}} = \sigma_{Y_{n+1}V_2} - \Sigma_{Y_{n+1}\mathcal{D}}\Sigma^{-1}_{\mathcal{D}\mathcal{D}}\Sigma_{\mathcal{D}V_2}.$$

So, from (3.2) we get $Y_{n+1} \perp\!\!\!\perp V_2 \mid V_1$ and $\mathcal{D} \perp\!\!\!\perp V_2 \mid V_1$. Similar to the argument above in (3.3) we get $\Sigma_{\mathcal{D}V_2} = \Sigma_{\mathcal{D}V_1}\sigma_{V_1V_2}/\sigma_{V_1V_1}$. Now, by substitution we get:

$$\begin{aligned}
\sigma_{Y_{n+1}V_2|\mathcal{D}} &= \frac{\sigma_{Y_{n+1}V_1}\sigma_{V_1V_2}}{\sigma_{V_1V_1}} - \Sigma_{Y_{n+1}\mathcal{D}}\Sigma^{-1}_{\mathcal{D}\mathcal{D}}\Sigma_{\mathcal{D}V_1}\frac{\sigma_{V_1V_2}}{\sigma_{V_1V_1}} \\
&= \frac{\sigma_{V_1V_2}}{\sigma_{V_1V_1}}\left\{\sigma_{Y_{n+1}V_1} - \Sigma_{Y_{n+1}\mathcal{D}}\Sigma^{-1}_{\mathcal{D}\mathcal{D}}\Sigma_{\mathcal{D}V_1}\right\} \\
&= \frac{\sigma_{V_1V_2}}{\sigma_{V_1V_1}}\sigma_{Y_{n+1}V_1|\mathcal{D}}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sigma_{V_2V_2|\mathcal{D}} &= \sigma_{V_2V_2} - \Sigma_{V_2\mathcal{D}}\Sigma^{-1}_{\mathcal{D}\mathcal{D}}\Sigma_{\mathcal{D}V_2} \\
&= \sigma_{V_2V_2} - \frac{\sigma^2_{V_1V_2}}{\sigma^2_{V_1V_1}}\Sigma_{V_1\mathcal{D}}\Sigma^{-1}_{\mathcal{D}\mathcal{D}}\Sigma_{\mathcal{D}V_1} \\
&= \sigma_{V_2V_2} - \frac{\sigma^2_{V_1V_2}}{\sigma^2_{V_1V_1}}\left(\sigma_{V_1V_1} - \sigma_{V_1V_1|\mathcal{D}}\right) \\
&= \sigma_{V_2V_2} - \frac{\sigma^2_{V_1V_2}}{\sigma_{V_1V_1}} + \frac{\sigma^2_{V_1V_2}}{\sigma^2_{V_1V_1}}\sigma_{V_1V_1|\mathcal{D}}.
\end{aligned}$$

Re-arranging, we see that

$$\sigma_{V_2V_2|\mathcal{D}} - \frac{\sigma^2_{V_1V_2}}{\sigma^2_{V_1V_1}}\sigma_{V_1V_1|\mathcal{D}} = \sigma_{V_2V_2} - \frac{\sigma^2_{V_1V_2}}{\sigma_{V_1V_1}} = \sigma_{V_2V_2|V_1} \geq 0.$$

So, we get the first clause:

$$\begin{aligned}
\frac{\sigma^2_{Y_{n+1}V_1|\mathcal{D}}}{\sigma_{V_1V_1|\mathcal{D}}} - \frac{\sigma^2_{Y_{n+1}V_2|\mathcal{D}}}{\sigma_{V_2V_2|\mathcal{D}}} &= \frac{\sigma^2_{Y_{n+1}V_1|\mathcal{D}}}{\sigma_{V_1V_1|\mathcal{D}}} - \frac{\sigma^2_{Y_{n+1}V_1|\mathcal{D}}\sigma^2_{V_1V_2}}{\sigma^2_{V_1V_1}\sigma_{V_2V_2|\mathcal{D}}} \\
&= \frac{\sigma^2_{Y_{n+1}V_1|\mathcal{D}}}{\sigma_{V_1V_1|\mathcal{D}}\sigma_{V_2V_2|\mathcal{D}}}\left(\sigma_{V_2V_2|\mathcal{D}} - \frac{\sigma^2_{V_1V_2}}{\sigma^2_{V_1V_1}}\sigma_{V_1V_1|\mathcal{D}}\right) \geq 0.
\end{aligned}$$

Clause 2 follows by combining Clause 1 with the observation that the totals of the terms in the two two-term expansions are equal. □

We illustrate Theorem 3.2 a slightly trivial Bayesian hierarchical model.

exm:nnn **Example 3.1.** *Suppose $Y_i$ for $i = 1, ..., n$ are IID $\mathsf{N}(\mu, \sigma_0^2)$ where $\sigma_0$ is known. Assume $\mu$ is distributed as $\mathsf{N}(\nu, \tau_0^2)$ and $\nu$ is distributed as $\mathsf{N}(a, b^2)$ where both $a$ and $b$ are known. Let $\eta_n = (n/\sigma_0^2 + 1/(\tau_0^2 + b^2))^{-1}$. Then, the posterior predictive distribution $(Y_{n+1}|y^n)$ is $\mathsf{N}(\eta_n(n\bar{y}/\sigma_0^2 + a/(\tau_0^2 + b^2)), \sigma_0^2 + \eta_n)$ with*

$$\mathsf{Var}(Y_{n+1}|y^n) = \sigma_0^2 + (n/\sigma_0^2 + 1/(\tau_0^2 + b^2))^{-1}. \tag{3.4}$$

*The conditional distribution $(\mu|y^n, \nu)$ is*

$$\mathsf{N}((n/\sigma_0^2 + 1/\tau_0^2)^{-1}(n\bar{Y}/\sigma_0^2 + \nu/\tau_0^2), (n/\sigma_0^2 + 1/\tau_0^2)^{-1}).$$

*The conditional distribution $(\nu|y^n)$ is*

$$\mathsf{N}((1/b^2 + 1/(\tau_0^2 + \sigma_0^2/n))^{-1}(a/b^2 + \bar{y}/(\tau_0^{+}\sigma_0^2/n)), (1/b^2 + 1/(\tau_0^2 + \sigma_0^2/n))^{-1}).$$

*Now, it is easy to see that*

$$\begin{aligned}
\mathsf{E}_{\nu|y^n}\mathsf{E}_{\mu|\nu,y^n}\mathsf{Var}(Y_{n+1}|y^n, \mu, \nu) &= \sigma_0^2, && (3.5)\\
\mathsf{E}_{\nu|y^n}\mathsf{Var}_{\mu|\nu,y^n}\mathsf{E}(Y_{n+1}|y^n, \mu, \nu), &= (n/\sigma_0^2 + 1/\tau_0^2)^{-1} && (3.6)\\
\mathsf{Var}_{\nu|y^n}\mathsf{E}_{\mu|\nu,y^n}\mathsf{E}(Y_{n+1}|y^n, \mu, \nu) &= \frac{(n/\sigma_0^2 + 1/\tau_0^2)^{-2}}{\tau_0^4}(1/b^2 + 1/(\tau_0^2 + \sigma_0^2/n))^{-1}, && (3.7)
\end{aligned}$$

*and that these three terms sum to (3.4). As expected,*

$$\mathsf{Var}(Y_{n+1}|y^n, \mu) = \sigma_0^2 < \sigma_0^2 + \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1} = \mathsf{Var}(Y_{n+1}|y^n, \nu).$$

□

As in Theorem 3.1, Assumption 1 is sufficient but not necessary for Theorem 3.2. As is evident from the proof, the required conditions are $Y_{n+1} \perp\!\!\!\perp V_2 \mid V_1$ and $\mathcal{D} \perp\!\!\!\perp V_2 \mid V_1$. It is well-known that (3.2) implies these two conditions, but the converse does not hold.

Theorem 3.2 also confirms the fact that a hyperparameter has less information about the data and the predicted value than a parameter does. This is similar to the data processing inequality, see Cover:Thomas:2006 Cover and Thomas (2006). In addition, if the joint density is Gaussian, then $\rho^2_{y_{n+1}V_2|\mathcal{D}} \leq \rho^2_{y_{n+1}V_1|\mathcal{D}}$, where $\rho$ is the partial correlation of its subscripts.

## 3.2 UQ under posterior independence

Empirically, often a term in (2.8) being zero coincides with a term in (2.9) being zero as well. We give results showing when this happens. Consider the following.

asmp:postcond **Assumption 2.** *$V_1$ is conditionally independent of $V_2$ given the data $\mathcal{D}$, i.e.,*

$$V_1 \perp\!\!\!\perp V_2 \mid \mathcal{D}. \tag{3.8}$$

eq:postcond

While constricting, Assumption 2 is satisfied by a large number of parametric families. An obvious example is when two experiments are combined, e.g., $Y_i \sim \mathsf{Poisson}(\lambda_i)$ and $\lambda_i \sim \mathsf{Gamma}(a_i, b_i)$ for $i = 1, 2$. This easily extends to many outcomes of $Y$. Another class of examples is exponential families whose sufficient statistics split additively across parameters equipped with conjugate priors. A more interesting example is the following.

exm:poissontime **Example 3.2.** *Let $Y_1 \sim \mathsf{Poisson}(t\lambda_1)$ and $Y_2 \sim \mathsf{Poisson}(t\lambda_2)$, where $\lambda_i \sim \mathsf{Gamma}(a_i, b)$, with the $Y_i$'s and the $\lambda_i$'s independent. The likelihood factorises into separate parts for $\lambda_1$ and $\lambda_2$ though both factors have $t$. The priors are independent so Assumption 2 holds.*

We have the following implications when a term in a three term expansion is zero.

thm:crsExp **Theorem 3.3.** *Suppose Assumption 2 holds. Then we have:*

1. *If term* (2.8b) *in* (2.8) *is zero, then term* (2.9c) *in* (2.9) *is zero. That is,*

$$\mathsf{E}_{V_1|\mathcal{D}}\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] = 0 \implies \mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2, \mathcal{D})] = 0. \quad (3.9)$$ eq:first

2. *If term* (2.9b) *in* (2.9) *is zero, then term* (2.8c) *in* (2.8) *is zero. That is,*

$$\mathsf{E}_{V_2|\mathcal{D}}\mathsf{Var}_{V_1|V_2,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2, V_1, \mathcal{D})] = 0 \implies \mathsf{Var}_{V_1|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, \mathcal{D})] = 0. \quad (3.10)$$ eq:second

*Proof.* We will only prove Clause I. The proof of Clause II is similar.

Since $V_1 \perp\!\!\!\perp V_2 \mid \mathcal{D}$, we have

$$\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] = \mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] = 0.$$

The LHS of (3.9) is zero, so it follows that

$$\begin{aligned} &\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] = 0 \\ \Longrightarrow &\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D}) \text{ is a constant in terms of } V_2, \text{ for } V_1 \text{ and } \mathcal{D}. \\ \Longrightarrow &\frac{\partial}{\partial V_2}\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D}) = 0 \ \forall \ V_1,\ V_2 \text{ and } \mathcal{D}. \end{aligned}$$

Now, using the Leibniz rule for all $V_1$, $V_2$, and $\mathcal{D}$:

$$\begin{aligned} \frac{\partial}{\partial V_2}\mathsf{E}_{V_1|\mathcal{D}}\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D}) &= \frac{\partial}{\partial V_2}\int \mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D}) f_{V_1|\mathcal{D}} dV_1 \\ = \int \frac{\partial}{\partial V_2}\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D}) f_{V_1|\mathcal{D}} dV_1 = 0. \end{aligned}$$

That is, $\mathsf{E}_{V_1|\mathcal{D}}\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D}) = \mathsf{E}[Y_{n+1}|V_2, \mathcal{D}]$ is a constant in terms of $V_2$, for all $V_1$ and $\mathcal{D}$. This implies that $\mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2, \mathcal{D})] = 0$. □

Under Assumption 2, to satisfy the condition on the left of (3.9), it is enough for $\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})$ to be is independent of $V_2$ for all $V_1$ and $\mathcal{D}$. The conditional independence $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$ is not required. In fact, $\mathsf{Var}[Y_{n+1}|V_1, V_2, \mathcal{D}]$ may still depend on $V_2$. The analogous statements hold for the condition on the left of (3.10).

In general, if both (3.9) and (3.10) hold, we do not get any meaningful extra reduction. Indeed, suppose

$$\mathsf{E}_{V_1|\mathcal{D}}\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1,V_2,\mathcal{D})] = 0$$

and

$$\mathsf{E}_{V_2|\mathcal{D}}\mathsf{Var}_{V_1|V_2,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2,V_1,\mathcal{D})] = 0.$$

Then in the three term expansion starting with $V_1$, see (1.2), the middle term is zero from the first. From the second we get

$$\mathsf{Var}_{V_1|V_2,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2,V_1,\mathcal{D})] = 0,$$

which gives that $\mathsf{E}(Y_{n+1}|V_2,V_1,\mathcal{D})$ is free of $V_1$, i.e., $\mathsf{E}(Y_{n+1}|V_2,V_1,\mathcal{D}) = \mathsf{E}(Y_{n+1}|V_2,\mathcal{D})$. However, using this in (2.8c) gives

$$\mathsf{Var}_{V_1|\mathcal{D}}\mathsf{E}_{V_2|V_1,\mathcal{D}}\mathsf{E}(Y_{n+1}|V_2,\mathcal{D})$$

which need not be zero because $\mathsf{E}_{V_2|V_1,\mathcal{D}}\mathsf{E}(Y_{n+1}|V_2,\mathcal{D})$ may depend on $V_1$ even though $Y_{n+1}$ does not.

Assumption 2 is sufficient but not necessary for Theorem 3.3 to hold. The next two examples show that i) without Assumption 2 Theorem 3.3 need not hold, and ii) the conclusions of Theorem 3.3 can hold even when Assumption 2 does not.

exm:norm1 **Example 3.3.** *Let $Y_1$, $Y_2$, …, $Y_n$ be IID $\mathsf{N}\left(\mu, 1/\lambda^2\right)$ with $\mu \sim \mathsf{N}\left(\mu_0, 1/\lambda_0^2\right)$ with known $\mu_0$ and $\lambda_0$ and $\lambda^2 \sim \mathsf{Gamma}\left(\alpha_0,\beta_0\right)$ with known $\alpha_0$ and $\beta_0$. Set $V_1 = \mu$, $V_2 = \lambda^2$, and $\mathcal{D} = \{Y_1, Y_2, \dots, Y_n\}$. Then we can show that*

$$\mu|\lambda^2,\mathcal{D} \sim \mathsf{N}\left((\lambda^2\sum_{i=1}^n Y_i + \lambda_0^2\mu_0)(n\lambda^2+\lambda_0^2)^{-1}, (n\lambda^2+\lambda_0^2)^{-1}\right).$$

*Since this distribution depends on $\lambda^2$, $\mu \not\perp \lambda^2|\mathcal{D}$ and Assumption 2 does not hold.*

*We see that Theorem 3.3 does not hold either. Even though $\mathsf{E}[Y_{n+1}|\mu,\lambda^2,\mathcal{D}] = \mu$, is free of $\lambda^2$, i.e.,*

$$\mathsf{E}_{\mu|\mathcal{D}}\mathsf{Var}_{\lambda^2|\mu,\mathcal{D}}[\mathsf{E}(Y_{n+1}|\mu,\lambda^2,\mathcal{D})] = \mathsf{E}_{\mu|\mathcal{D}}\mathsf{Var}_{\lambda^2|\mu,\mathcal{D}}[\mu] = 0$$

*we also have that*

$$\mathsf{Var}_{\lambda^2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|\lambda^2,\mathcal{D})] = \mathsf{Var}_{\lambda^2|\mathcal{D}}\mathsf{E}_{\mu|\mathcal{D},\lambda}[\mathsf{E}[Y_{n+1}|\mu,\lambda^2,\mathcal{D}]] \tag{3.11}$$

$$=\mathsf{Var}_{\lambda^2|\mathcal{D}}\mathsf{E}_{\mu|\mathcal{D},\lambda}[\mu] = \mathsf{Var}_{\lambda^2|\mathcal{D}}\left[\frac{\lambda^2\sum_{i=1}^n Y_i + \lambda_0^2\mu_0}{n\lambda^2+\lambda_0^2}\right] > 0. \tag{3.12}$$

*The last inequality holds because $\mathsf{E}\mu|\mathcal{D},\lambda^2$ is not free of $\lambda^2$.* □

exm:norm2 **Example 3.4.** *Suppose $\mu \sim \mathsf{N}\left(\mu_0, 1/(\kappa_0\lambda^2)\right)$ is used in Example 3.3 so that Assumption 2 still does not hold. Now, $\mu|\lambda^2,\mathcal{D} \sim \mathsf{N}\left((\sum_{i=1}^n Y_i + \kappa_0\mu_0)(n+\kappa_0)^{-1}, (n+\kappa_0)^{-1}\lambda^{-2}\right)$ and*

*$\mu$ remains dependent on $\lambda^2$ (given $\mathcal{D}$). We also see that $E[Y_{n+1}|\mu, \lambda^2, \mathcal{D}] = \mu$ remains free of $\lambda^2$ and this gives*

$$\mathsf{E}_{\mu|\mathcal{D}}\mathsf{Var}_{\lambda^2|\mu,\mathcal{D}}[\mathsf{E}(Y_{n+1}|\mu, \lambda^2, \mathcal{D})] = \mathsf{E}_{\mu|\mathcal{D}}\mathsf{Var}_{\lambda^2|\mu,\mathcal{D}}[\mu] = 0.$$

*In addition, unlike Example 3.3, we find*

$$\mathsf{Var}_{\lambda^2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|\lambda^2, \mathcal{D})] = \mathsf{Var}_{\lambda^2|\mathcal{D}}\mathsf{E}_{\mu|\mathcal{D},\lambda}[\mu] = \mathsf{Var}_{\lambda^2|\mathcal{D}}\left[\frac{\sum_{i=1}^n Y_i + \kappa_0\mu_0}{n + \kappa_0}\right] = 0.$$

*That is, the conclusions of Theorem 3.3 are satisfied.* □

Even though Assumption 2 is not necessary, from Examples 3.3 and 3.4 it is evident that the necessary condition could be difficult to specify, might depend on the specific parametrisation, and be quite hard to verify in practice. Indeed, to make term (2.8c) zero, without $V_1 \perp\!\!\!\perp V_2 \mid \mathcal{D}$, we effectively need $\mathsf{E}[Y_{n+1}|V_1, V_2, \mathcal{D}] = \mu(V_1, \mathcal{D})$ and $\mathsf{E}_{V_1|V_2,\mathcal{D}}[\mu(V_1, \mathcal{D})]$ to be free of $V_2$. Example 3.3 clearly shows that, depending on the parametrisation, such conditions may not hold. Moreover such conditions may be hard to verify and interpret whereas the sufficient Assumption 2 can be relatively easily verified and interpreted, e.g., in Example 2.8.

It is easy to see that under Assumption (2), neither (3.9) nor (3.10) leads to an interpretable two-term expansion of the predictive variance. For instance, if we have $\mathsf{E}_{V_1|\mathcal{D}}\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] = 0$ then we get

$$\mathsf{Var}(Y_{n+1}|\mathcal{D}) = \mathsf{E}_{V_1,V_2|\mathcal{D}}[\mathsf{Var}(Y_{n+1}|V_2, V_1, \mathcal{D})] + \mathsf{Var}_{V_1|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, \mathcal{D})],$$

in which $\mathsf{Var}(Y_{n+1}|V_2, V_1, \mathcal{D})$ is not in general independent of $V_2$, similarly if $V_1$ and $V_2$ are interchanged.

The converses do not hold either without extra assumptions. We have the following result, which is symmetric in $V_1$ and $V_2$.

thm:rev

**Theorem 3.4.** *Suppose Assumption 2. Then:*

1. *If, additionally, for all $V_1$, $V_2$, $\mathcal{D}$, we have that $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$, then*

$$\mathsf{E}_{V_1|\mathcal{D}}\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] = 0 \quad \textit{and} \quad \mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2, \mathcal{D})] = 0.$$

2. *In this case, the three-term expansion using the LTV on $V_1$ first and $V_2$ second reduces to a two term expansion.*

*Proof.* For Clause 1, if $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$ for all $V_1$, $V_2$, $\mathcal{D}$, then it trivially follows that $\mathsf{E}[Y_{n+1}|V_1, V_2, \mathcal{D}] = \mathsf{E}[Y_{n+1}|V_1, \mathcal{D}]$, which is independent of $V_2$. That is, we have both

$$\mathsf{Var}_{V_2|V_1,\mathcal{D}}[\mathsf{E}[Y_{n+1}|V_1, V_2, \mathcal{D}]] = \mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_1, \mathcal{D})] = 0$$

and

$$\begin{aligned}\mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}(Y_{n+1}|V_2, \mathcal{D})] &= \mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}_{V_1|V_2,\mathcal{D}}\mathsf{E}(Y_{n+1}|V_1, V_2, \mathcal{D})] \\ &= \mathsf{Var}_{V_2|\mathcal{D}}[\mathsf{E}_{V_1|\mathcal{D}}\mathsf{E}(Y_{n+1}|V_1, \mathcal{D})] = 0.\end{aligned}$$

Clause 2 follows from Theorem 3.1. □

From the proof of Theorem 3.4, we see that we only require $\mathsf{E}[Y_{n+1}|V_1, V_2, \mathcal{D}]$ to be independent of $V_2$ for the equivalence of the three and two term expansions. This is the familiar condition first-order ancillarity, Lehmann:1998 Lehmann and Casella (1998), p. 41. For Statement 2, we also require that $\mathrm{Var}[Y_{n+1}|V_1, V_2, \mathcal{D}]$ be independent of $V_2$. We assume this stronger structural condition $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$ because it is easier to verify.

Note that the assumption that $V_1 \perp\!\!\!\perp V_2 \mid \mathcal{D}$ (Assumption 2) and $Y_{n+1} \perp\!\!\!\perp V_2 \mid (V_1, \mathcal{D})$ imply that the condition $(Y_{n+1}, V_1) \perp\!\!\!\perp V_2 \mid \mathcal{D}$ holds for all $V_1$, $V_2$, and $\mathcal{D}$. For many models, such relationships can be easily determined from their description. Theorems 3.3 and 3.4 would rarely apply to a hierarchical Bayes model. However, the hidden Markov Model in Example 2.8 would satisfy all conditions of both theorems.

## 4 General Expansions of the PPV

decomposition

Here, we extend the results of the previous sections to general multi-term expansions of the posterior predictive variance. Without loss of generality, we fix $\mathcal{V} = \{V_1, V_2, \dots, V_K\}$ to be a specific ordering of the entries in $\mathcal{Z}$ and assume that each element in $\mathcal{Z}$ appears in our expansion. That is, each $Z_i$ is *manifest*, no element is *latent*.

A trivial but condensed two-term expansion of the PPV given $\mathcal{V}$ and $\mathcal{D}$ is

$$\begin{aligned}\mathsf{Var}(Y_{n+1}|\mathcal{D}_n)(\mathcal{V}) &= \mathsf{E}_{(V_1,\dots,V_K)}\mathsf{Var}(Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_K)\\ &\quad + \mathsf{Var}_{(V_1,\dots,V_K)}\mathsf{E}(Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_K). \qquad (4.1)\end{aligned}$$

condensed_var

More interesting is the $K+1$-term expansion of posterior predictive variance $\mathsf{Var}[Y_{n+1}|\mathcal{D}]$ w.r.t. $\mathcal{V}$. This is given by:

Conditional_Var_sum

$$\mathsf{Var}(Y_{n+1}|\mathcal{D}_n)(\mathcal{V}) = \mathsf{E}_{(V_1,\dots,V_K)|\mathcal{D}}\mathsf{Var}[Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_K] \qquad (4.2a)$$

eq:g1

$$+ \sum_{k=K}^{2} \mathsf{E}_{(V_1,\dots,V_{k-1})|\mathcal{D}}\mathsf{Var}_{V_k|\mathcal{D},V_1,\dots,V_{k-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_k] \qquad (4.2b)$$

eq:g2

$$+ \mathsf{Var}_{V_1|\mathcal{D}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1]. \qquad (4.2c)$$

eq:g3

The terms in (4.2b) are added in decreasing order of $k$ for notational convenience; this become clear in the sequel.

The inner expectation in the $k$-th in summands in (4.2b) and (4.2c) can be iterated as a sequence of conditional expectations:

$$\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_k] = \mathsf{E}_{V_{k+1}|\mathcal{D},V_1,\dots,V_k}\cdots\cdots\mathsf{E}_{V_K|\mathcal{D},V_1,\dots,V_{K-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_K],$$

and

$$\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1] = \mathsf{E}_{V_2|\mathcal{D},V_1}\cdots\cdots E_{V_K|\mathcal{D},V_1,\dots,V_{K-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1, \dots, V_K],$$

respectively. That is, even though all $V_k$'s appear in the PPV, only $V_1$, $V_2$, ..., $V_k$ appear in the $k$-th term of the sum in (4.2b); the rest of elements are latent.

The expansion in (4.2) depends on the choice of $\mathcal{V}$, i.e., the specific permutation of the elements of $\mathcal{Z}$. The results of Section 3 extend to this general expansion and the uncertainties in the PPV can be quantified.

UQstructural

## 4.1 UQ under Structural Conditions

Suppose that the set of variables $\mathcal{V}$ can be split into $\mathcal{V}_1 = \{V_1, V_2, \ldots, V_m\}$ and $\mathcal{V}_2 = \{V_{m+1}, V_{m+2}, \ldots, V_K\}$. We extend Assumption 1 to the general $K$ and $m$ setting.

asmp:condindext

**Assumption 3.** *$Y_{n+1}$ and $\mathcal{D}$ are conditionally independent of $\mathcal{V}_2$ given $\mathcal{V}_1$, i.e.,*

$$(Y_{n+1}, \mathcal{D}) \perp\!\!\!\perp \mathcal{V}_2 \mid \mathcal{V}_1 \tag{4.3}$$

condIndmath

Under this assumption, Theorem 3.1 admits a straightforward extension to the reduction of a $(K+1)$-term expansion to an $(M+1)$-term expansion given $\mathcal{V}_1$ and $\mathcal{D}$.

thm:gStCond

**Theorem 4.1.** *Under Assumption 3 we have*

$$\mathsf{Var}[Y_{n+1}|\mathcal{D}_n](\mathcal{V}) = \mathsf{Var}[Y_{n+1}|\mathcal{D}_n](\mathcal{V}_1).$$

*Proof.* Since the conditional independence implies $Y_{n+1} \perp\!\!\!\perp \mathcal{V}_2 \mid (\mathcal{D}, \mathcal{V}_1)$, we have

$$\mathsf{E}\left[Y_{n+1}|\mathcal{D},\mathcal{V}\right] = \mathsf{E}\left[Y_{n+1}|\mathcal{D},\mathcal{V}_1\right] \quad \text{and} \quad \mathsf{Var}\left[Y_{n+1}|\mathcal{D},\mathcal{V}\right] = \mathsf{Var}\left[Y_{n+1}|\mathcal{D},\mathcal{V}_1\right].$$

Hence, in (4.2a),

$$\mathsf{E}_{\mathcal{V}|\mathcal{D}}\mathsf{Var}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}] = \mathsf{E}_{\mathcal{V}|\mathcal{D}}\mathsf{Var}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1] = \mathsf{E}_{\mathcal{V}_1|\mathcal{D}}\mathsf{Var}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1].$$

Similarly, it follows that:

$$\begin{aligned}
&\mathsf{E}[Y_{n+1}|\mathcal{D}_n, V_1] = \mathsf{E}_{V_2|\mathcal{D},V_1}\cdots\cdots\mathsf{E}_{V_K|\mathcal{D},V_1,\ldots,V_{K-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}]\\
=&\mathsf{E}_{V_2|\mathcal{D},V_1}\cdots\mathsf{E}_{V_M|\mathcal{D},V_1,\ldots,V_{M-1}}\mathsf{E}_{V_{M+1}|\mathcal{D},\mathcal{V}_1}\cdots\mathsf{E}_{V_K|\mathcal{D},\mathcal{V}_1,V_{m+1},\cdots,V_{K-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1]\\
=&\mathsf{E}_{V_2|\mathcal{D},V_1}\cdots\mathsf{E}_{V_M|\mathcal{D},V_1,\ldots,V_{M-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1].
\end{aligned}$$

The last equality holds because $\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1]$ is free of elements in $\mathcal{V}_2$.

Using the above argument we also see that $K > m$ implies the summands in (4.2b) translate to:

$$\begin{aligned}
&\mathsf{E}_{(V_1,\ldots,V_{k-1})|\mathcal{D}}\mathsf{Var}_{V_k|\mathcal{D},V_1,\ldots,V_{k-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,V_1,\ldots,V_k]\\
=&\mathsf{E}_{(V_1,\ldots,V_{k-1})|\mathcal{D}}\mathsf{Var}_{V_k|\mathcal{D},\mathcal{V}_1,V_{m+1}\ldots,V_{k-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1] = 0,
\end{aligned}$$

since $\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1]$ is free of $V_k$.

Now, collecting all the terms, we get:

$$\begin{aligned}
&\mathsf{Var}(Y_{n+1}|\mathcal{D}_n)(\mathcal{V})\\
=&\mathsf{E}_{\mathcal{V}_1|\mathcal{D}}\mathsf{Var}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1] + \sum_{k=M}^{2} E_{(V_1,\ldots,V_{k-1})|\mathcal{D}}\mathsf{Var}_{V_k|\mathcal{D},V_1,\ldots,V_{k-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1]\\
&+\mathsf{Var}_{V_1|\mathcal{D}}\mathsf{E}_{V_2|\mathcal{D},V_1}\cdots\mathsf{E}_{V_M|\mathcal{D},V_1,\ldots,V_{M-1}}\mathsf{E}[Y_{n+1}|\mathcal{D}_n,\mathcal{V}_1] = \mathsf{Var}[Y_{n+1}|\mathcal{D}_n](\mathcal{V}_1).
\end{aligned}$$

□

## 4.2 UQ under Posterior Independence

We now quantify the uncertainty across different permutations of the elements in $\mathcal{Z}$. Let $\pi$ be a permutation of $\{1,\ldots,K\}$ and $\mathcal{V}^\pi$ denote the corresponding permutation of the elements in $\mathcal{V}$. Our goal is to give conditions under which certain terms in the expansion of $Var[Y_{n+1}|\mathcal{D}](\mathcal{V})$ being zero, imply that certain terms in the expansion of $Var[Y_{n+1}|\mathcal{D}](\mathcal{V}^\pi)$ are also zero.

Without loss of generality, fix $\mathcal{V}$ to be a permutation of the elements in $\mathcal{Z}$. Let $I = \{1,2,\ldots,K\}$. We write the posterior predictive variance in (4.2) as:

$$Var[Y_{n+1}|\mathcal{D}](\mathcal{V}) = T_0^I + T_K^I + \cdots + T_1^I, \tag{4.4}$$ eq:Ts

where

$$\begin{aligned}
T_0^I &= E_{\mathcal{V}|\mathcal{D}} Var\left[Y_{n+1}|\mathcal{D},\mathcal{V}\right],\\
T_k^I &= E_{(V_1,\ldots,V_{k-1})|\mathcal{D}} Var_{V_k|\mathcal{D},V_1,\ldots,V_{k-1}} E[Y_{n+1}|\mathcal{D}_n, V_1,\ldots,V_k], \quad \text{for } k = 2,\ 3,\ \ldots,\ K,\\
T_1^I &= Var_{V_1|\mathcal{D}} E[Y_{n+1}|\mathcal{D}_n, V_1].
\end{aligned}$$

Note that the term $T_0^I$ involves the conditional variance of $Y_{n+1}$. More importantly, for $k \geq 1$, the term $T_k^I$ involves the conditional variance of $V_k$.

Now let $\mathcal{V}^\pi = \{V_1^\pi, V_2^\pi, \ldots, V_K^\pi\}$ be the permuted elements of $\mathcal{Z}$, where $V_j^\pi = V_i$ iff $\pi(i) = j$. Further, in the same way as (4.4) write:

$$Var[Y_{n+1}|\mathcal{D}](\mathcal{V}^\pi) = T_0^\pi + T_K^\pi + \cdots + T_1^\pi, \tag{4.5}$$ eq:Tps

where, as before, the term $T_k^\pi$ involves computing the variance w.r.t. $V_k^\pi$, for all $k \geq 1$.

To generalize Theorem 3.3 to $(K+1)$-term expansions we start with the following. Let $\neg i = \{K, K-1, \ldots, i+1, i-1, \ldots, 1\}$.

condind **Lemma 1.** *Suppose $V_i \perp\!\!\!\perp V_{\neg i} \mid \mathcal{D}$. Then, $\forall u = j+1, \ldots K$*

$$V_j \perp\!\!\!\perp V_u \mid (\mathcal{D}, V_1, V_2, \ldots, V_{j-1}, V_{j+1}, \ldots, V_{u-1}).$$

*Proof.* Include proof. □

thm:genExp **Theorem 4.2.** *In the $(K+1)$-term expansions of $Var[Y_{n+1}|\mathcal{D}](\mathcal{V})$ and $Var[Y_{n+1}|\mathcal{D}](\mathcal{V}^\pi)$ defined above:*

1. *For all $i = 1,\ldots,K$, if $\pi(i) = i$, and $\pi(\{i-1,\ldots,1\}) = \{i-1,\ldots,1\}$, then $T_i^I = 0 \iff T_i^\pi = 0$.*
2. *Under the conditions of Lemma 1, if $\pi(i) = j$, and $\{j-1,\ldots,1\} \subset \pi(\{i-1,\ldots,1\})$, we have that $T_i^I = 0 \implies T_j^\pi = 0$.*

*Proof.* For the first statement, note that

$$\begin{aligned} T_i^I = 0 &\Longleftrightarrow \mathsf{Var}_{V_i|\mathcal{D},V_1,V_2,\ldots,V_{i-1}} \mathsf{E}[Y_{n+1}|\mathcal{D},V_1,V_2,\ldots,V_i] = 0 \\ &\Longleftrightarrow \mathsf{E}[Y_{n+1}|\mathcal{D},V_1,V_2,\ldots,V_i] \text{ is free of } V_i \\ &\Longleftrightarrow \mathsf{E}[Y_{n+1}|\mathcal{D},V_1^\pi,V_2^\pi,\ldots,V_i^\pi] \text{ is free of } V_i^\pi \\ &\Longleftrightarrow \mathsf{Var}_{V_i^\pi|\mathcal{D},V_1^\pi,V_2^\pi,\ldots,V_{i-1}^\pi} \mathsf{E}[Y_{n+1}|\mathcal{D},V_1^\pi,V_2^\pi,\ldots,V_i^\pi] = 0 \\ &\Longleftrightarrow T_i^\pi = 0. \end{aligned} \tag{4.6}$$ eq:perm

For the second statement, because variance is non-negative

$$\begin{aligned} T_i^I = 0 &\Longleftrightarrow \mathsf{Var}_{V_i|\mathcal{D},V_1,V_2,\ldots,V_{i-1}} \mathsf{E}[Y_{n+1}|\mathcal{D},V_1,V_2,\ldots,V_i] = 0 \\ &\Longleftrightarrow \mathsf{E}[Y_{n+1}|\mathcal{D},V_1,V_2,\ldots,V_i] \text{ is free of } V_i \\ &\Longleftrightarrow \frac{\partial}{\partial V_i} E[Y_{n+1}|\mathcal{D},V_1,V_2,\ldots,V_i] = 0 \text{ for all } \mathcal{D}, \end{aligned} \tag{4.7}$$ eq:prm

with mild abuse of notation.

Now, suppose that the set

$$\pi(\{i-1,\ldots,1\}) = \{j', j'-1,\ldots,j+1,j-1,\ldots,1\},$$

From the discussion above:

$$\begin{aligned} &E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_j^\pi] \\ &= E_{V_{j'}^\pi|\mathcal{D},V_1^\pi,\ldots,V_{j'-1}^\pi} E_{V_{j'-1}^\pi|\mathcal{D},V_1^\pi,\ldots,V_{j'-2}^\pi} \cdots E_{V_{j+1}^\pi|\mathcal{D},V_1^\pi,\ldots,V_j^\pi} E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_{j-1}^\pi,V_j^\pi,V_{j+1}^\pi,\ldots,V_{j'}^\pi] \end{aligned} \tag{4.8}$$ eq:perm2

Now, by Lemma 1, each expectation other than $E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_{j-1}^\pi,V_j^\pi,V_{j+1}^\pi,\ldots,V_{j'}^\pi]$ in (4.8) is free of $V_j^\pi$. Now, the dominated derivative theorem gives

$$\begin{aligned} \frac{\partial}{\partial V_j^\pi} E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_j^\pi] &= E_{V_{j'}^\pi|\mathcal{D},V_1^\pi,\ldots,V_{j'-1}^\pi} E_{V_{j'-1}^\pi|\mathcal{D},V_1^\pi,\ldots,V_{j'-2}^\pi} \cdots \\ &\cdots E_{V_{j+1}^\pi|\mathcal{D},V_1^\pi,\ldots,V_j^\pi} \frac{\partial}{\partial V_j^\pi} E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_{j-1}^\pi,V_j^\pi,V_{j+1}^\pi,\ldots,V_{j'}^\pi] \end{aligned}$$

Now, (4.7) implies that:

$$\begin{aligned} &\frac{\partial}{\partial V_j^\pi} E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_{j-1}^\pi,V_j^\pi,V_{j+1}^\pi,\ldots,V_{j'}^\pi] \\ &= \frac{\partial}{\partial V_i} E[Y_{n+1}|\mathcal{D},V_1,V_2,\ldots,V_i] = 0 \text{ for all } \mathcal{D},\ V_1,\ V_2,\ \ldots,\ V_i. \end{aligned}$$

That is:

$$\frac{\partial}{\partial V_j^\pi} E[Y_{n+1}|\mathcal{D},V_1^\pi,\ldots,V_j^\pi] = 0$$

$$\iff E[Y_{n+1}|\mathcal{D}, V_1^\pi, \ldots, V_j^\pi] \text{ is free of } V_j^\pi$$
$$\iff T_j^\pi = Var_{V_j^\pi|\mathcal{D}, V_1^\pi, \ldots, V_{j-1}^\pi} E[Y_{n+1}|\mathcal{D}, V_1^\pi, \ldots, V_j^\pi] = 0.$$

□

**Example 4.1.** *Let $\mathcal{V} = \{V_1, V_2, V_3\}$, ie. $I = \{1,2,3\}$. Suppose the 4-term expansion of PPV is given by:*

$$Var\left[Y_{n+1}|\mathcal{D}, V_1, V_2, V_3\right] = T_0^{\{1,2,3\}} + T_3^{\{1,2,3\}} + T_2^{\{1,2,3\}} + T_1^{\{1,2,3\}}.$$

*Then the following implications are evident from Theorem 4.1.*

1. $T_3^{\{1,2,3\}} = 0$ *iff* $T_3^{\{2,1,3\}} = 0$. *Furthermore, if* $V_3 \perp\!\!\!\perp (V_1, V_2) \mid \mathcal{D}$, *then*

$$T_3^{\{1,2,3\}} = 0 \implies T_2^{\{1,3,2\}} = 0 \implies T_1^{\{3,1,2\}} = 0, \quad \textit{and}$$
$$T_3^{\{1,2,3\}} = 0 \implies T_2^{\{2,3,1\}} = 0 \implies T_1^{\{3,2,1\}} = 0.$$

2. $T_3^{\{1,3,2\}} = 0$ *iff* $T_3^{\{3,1,2\}} = 0$. *Furthermore, if* $V_2 \perp\!\!\!\perp (V_1, V_3) \mid \mathcal{D}$, *then*

$$T_3^{\{1,3,2\}\}} = 0 \implies T_3^{\{1,2,3\}} = 0 \implies T_1^{\{2,1,3\}} = 0, \quad \textit{and}$$
$$T_3^{\{1,3,2\}} = 0 \implies T_2^{\{3,2,1\}} = 0 \implies T_1^{\{2,3,1\}} = 0.$$

3. $T_3^{\{2,3,1\}} = 0$ *iff* $T_3^{\{3,2,1\}} = 0$. *Furthermore, if* $V_1 \perp\!\!\!\perp (V_2, V_3) \mid \mathcal{D}$, *then*

$$T_3^{\{2,3,1\}} = 0 \implies T_3^{\{2,1,3\}} = 0 \implies T_1^{\{1,2,3\}} = 0, \quad \textit{and}$$
$$T_3^{\{2,3,1\}} = 0 \implies T_2^{\{3,1,2\}} = 0 \implies T_1^{\{1,3,2\}} = 0.$$

*For details, see the* **ref to appendix**.

*Put in definition of* $\pi$

## 4.3 More General Expansions of the Posterior Predictive Variance

Let $\mathcal{V_M} = \{V_1, V_2, \ldots, V_M\} \subseteq \mathcal{Z}$ is an ordered set of variables in the model. We now assume that $\mathcal{V_M}$ is manifest, rest of $\mathcal{V} \setminus \mathcal{V_M}$ are latent, and consider an expansion of posterior predictive variance in terms of the elements of $\mathcal{V_M}$.

For all $u \le M \le K$ a $(u+1)$-term expansion of $Var\left[Y_{n+1}|\mathcal{D}\right](\mathcal{V_M})$ can be achieved by first splitting the set $\mathcal{V_M}$ into $u$ mutually exclusive and collectively exhaustive subsets, and then applying the LTV $u$-times to the subsets.

The expressions of the expansions follow directly from (4.2) above, with a change that each conditioning set can now have more than one element. Clearly, for any given $u > 1$, multiple expansions are possible, depending on the permutation of the subsets. Theorems 4.1 and 4.2 will also hold mutatis mutandis.

It is clear that, given a set $\mathcal{Z}$, a large number of expansions of the posterior predictive variance is possible. We now count the number of such expansions under the most general structure.

Cscopecard1 **Proposition 4.1.** *Fix $K$, $M$, $u$ as defined above, and suppose $u \le M \le K$. Then the number of possible $(u{+}1)$-term posterior predictive variance expansions of $Var\left[Y_{n+1}|\mathcal{D}\right](\mathcal{V}_{\mathcal{M}})$ is given by:*

$$u!\binom{K}{M}S(M,u). \tag{4.9}$$

Stirling

*Consequently, for fixed $K$, the total number of posterior predictive variance expansions is given by*

$$\sum_{u=1}^{K}\sum_{M=u}^{K} u!\binom{K}{M}S(M,u). \tag{4.10}$$

totalStirling

*Here, $S(M,u)$ is the Stirling number of the second kind with fixed $u \le M$.*

*Proof.* By definition, $S(M,u)$ is the number of ways to form non-void, disjoint, and exhaustive collections of $u$ subsets from $M$ distinct objects.

We start by observing that for a $(u+1)$-term expansion of the PPV, we must have $M$ disjoint non-void 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 factor of $u!$. Since we can do this for any choice of $M$ manifest variables out of $K$ variables, we get (4.9). Summing over all the possible values of $M$and $u$ gives (4.10). ☐

**Example 4.2.** *For $K = 2$, from (4.10) there are five possibilities. These possibilities can be listed as follows. For $K = 2$, $M$ could be either 1, or 2. For $M = 1$ and $u = 1$, there are two possibilities: either $V_1$ alone or $V_2$ alone is manifest, that is, $V_2$ 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 $V_1$.*

## 5 Term-wise Relative Size of Uncertainty

Draper

In this section we discuss the relative size of the terms in the posterior predictive variance expansions discussed above. **Some discussion on the size of the terms, vis-a-vis** Draper:1995 **(Draper, 1995)**.

**place the two-term expansion for normal mean. The first term is always(?) the leading term. Then we move to the Bayesian Two way Anova, where the first term may not always be the leading term. Then we move to numerical examples.**

Asymptotically, it is easy to see that under standard regularity conditions (independent data with well-behaved conditional densities and $V_k$'s having well-behaved distributions) the leading 'E Var' term in two term expansions is $\mathcal{O}_p(1)$, often simply a constant given by the integral of a variance function. Moreover, the second term is typically going to be $\mathcal{O}_p(1/n)$ for reasons of posterior normality. For three term expansions – which actually covers all cases simply by concatenating individual $V_k$'s on each side of the variance operation – we find that, under regularity conditions, the leading term is $\mathcal{O}_p(1)$ and the last term is $\mathcal{O}_p(1/n)$. However, the middle term is either $\mathcal{O}_p(1)$ if the inner variance does not depend on the data, i.e., is only a function of $V_1$, or $\mathcal{O}_p(1/n)$ if posterior normality holds for $V_2$ for each $V_1 = v_1$.

In this section we present three examples of the proposed expansion of the PPV. The first is a two-way random coefficient model. The second and third are re-analyses of two examples in Draper:1995 Draper (1995).

## 5.1 Two-Way Random Coefficient Model

2wayANOVAeg

check 2 way ANOVA

Consider the Baysian model defined by

$$Y_{ij} = \tau_i + \beta_j + \epsilon_{ij}, \tag{5.1}$$

where $i = 1, \ldots, T$, $j = 1, \ldots, B$ and we set

$$\begin{aligned} \tau_i &\sim N((\tau_0, \sigma_\tau^2) \\ \beta_j &\sim N(\beta_0, \sigma_\beta^2) \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \end{aligned} \tag{5.2}$$

with

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

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

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

the notation $U \perp\!\!\!\perp V$ means that $U$ is marginally independent of $V$.

We assume we have data in $\mathbf{y} = (y_{11}, \ldots, y_{TB})$, where $n = T \times B$ and we want to predict a future $Y_{ij}$ for some prespecified $i$ and $j$ with $1 \leq i \leq T$ and $1 \leq j \leq B$.

Because of the model (5.1), the future $Y_{ij}$ will depend on the entire matrix $\mathbf{y}$. Thus, the predictor automatically borrows information from "similar" observations. So, PI's will provide a quantification of the uncertainty for the $ij$-th cell averaging over the parameters in the model. Models such as (5.1) arise naturally in small-area estimation, poverty mapping, missing plot analysis in agronomy, and Bayesian imputation for missing values in contingency tables amongst other settings.

The three-term expansion of the PPV, with $V_1 = \beta$ and $V_2 = \tau$ is

$$\begin{aligned} \text{Var}(Y_{ij}|\mathbf{y}) = \; & E_{\tau|\mathbf{y}} E_{\beta|\mathbf{y},\tau} \text{Var}(Y_{ij}|\mathbf{y}, \beta, \tau) \\ & + E_{\tau|\mathbf{y}} \text{Var}_{\beta|\mathbf{y},\tau} E(Y_{ij}|\mathbf{y}, \beta, \tau) \\ & + \text{Var}_{\tau|\mathbf{y}} E_{\beta|\mathbf{y},\tau} E(Y_{ij}|\mathbf{y}, \beta, \tau). \end{aligned} \tag{5.3}$$

From the detailed derivations in Appendix B we get:

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

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

The key issue is the relative sizes of these three terms. Term 1 is easy to visualize because it's a constant. Term 2 is next easiest because it does not depend on $B$ or $\sigma_\tau^2$. It is seen that Term 2 is smaller than both $\sigma_\epsilon^2$ and $\sigma_\beta^2$.

Indeed, for fixed $\sigma_\epsilon^2$ and $\sigma_\beta^2$, Term 2 is asymptotically $\mathcal{O}(1/T)$ and in the limit of $\sigma_\beta^2 \to \infty$, Term 2 converges to $\sigma_\epsilon^2/T > 0$ – which is the variance of $\bar{y}_{\cdot j}$.

The behaviour of Term 3 is more complicated. It depends on $B$ through $a$ and $T$ through $b$. By comparing the numerator and the denominator, we see that, asymptotically for large $T$ holding the other constants fixed, Term 3 is $\mathcal{O}(1)$. In the limit as $T \to \infty$, Term 3 convegres to $a^{-1} = (\sigma_\epsilon^2\sigma_\tau^2)/(B\sigma_\tau^2 + \sigma_\epsilon^2)$. Thus, for large values of $T$, Term 3 is independent of $\sigma_\beta^2$ and smaller than both $\sigma_\epsilon^2$ and $\sigma_\tau^2$. Analogous reasoning applied to Term 3 shows that, when all other parameters are fixed, it is $\mathcal{O}(1/B)$ as $B$ increases.

It is now clear that unless both $B$ and $T$ diverge to infinity, Terms 2 and 3 won't simultaneously go to zero. That is, if only one of $B$ or $T$ goes to infinity, at least one of Term 2 or 3 may be a significant fraction of the PPV.

When $T \to \infty$, Term 2 $\to 0$, and the limiting PPV is

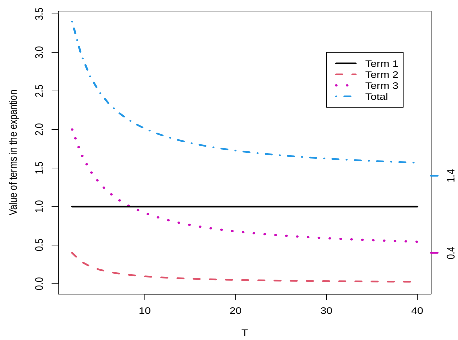
$$\sigma_\epsilon^2 + \frac{1}{a} = \sigma_\epsilon^2 \frac{(B+1)\sigma_\tau^2 + \sigma_\epsilon^2}{B\sigma_\tau^2 + \sigma_\epsilon^2}.$$

That is, under this asymptotic regime, relative size of Term 1 to the total is $\{B\sigma_\tau^2 + \sigma_\epsilon^2\}/\{(B+1)\sigma_\tau^2 + \sigma_\epsilon^2\}$, and that of Term 3 is $\sigma_\tau^2/\{(B+1)\sigma_\tau^2 + \sigma_\epsilon^2\}$.

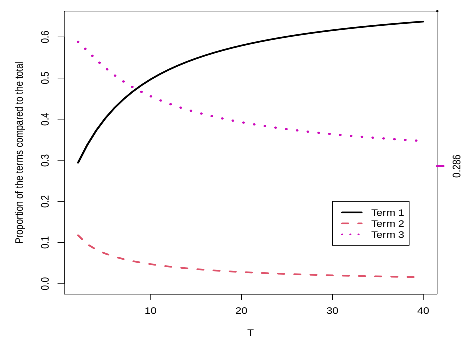
A similar phenomenon can be observed when $B \to \infty$. In this case Term 3 $\to 0$, and the limiting PPV is

$$\sigma_\epsilon^2 \frac{(T+1)\sigma_\beta^2 + \sigma_\epsilon^2}{T\sigma_\beta^2 + \sigma_\epsilon^2},$$

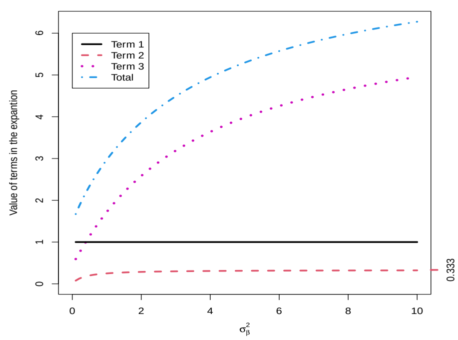
which implies that, the relative size of term 1 to the total is $\{T\sigma_\beta^2 + \sigma_\epsilon^2\}/\{(T+1)\sigma_\beta^2 + \sigma_\epsilon^2\}$ and that of Term 2 to the limiting PPV is $\sigma_\beta^2/\{(T+1)\sigma_\beta^2 + \sigma_\epsilon^2\}$, i.e., neither Term 1 nor Term 2 can be neglected in general.



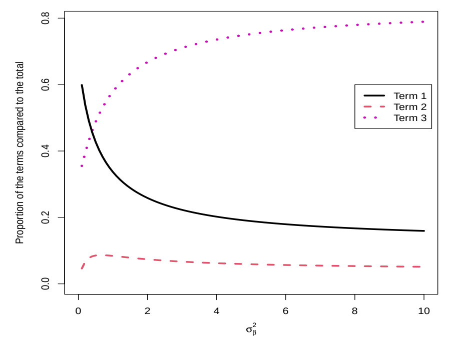

(a) Value of the terms. `fig:termT`




(b) Proportion of the terms to the total. `fig:propT`




(c) Value of the terms. `fig:termB`




(d) Proportion of the terms to the total. `fig:propB`

Figure 3: The value and the proportion of the three terms in 5.3. Figures 3(a) and 3(b) are for $\sigma_\tau^2 = \sigma_\beta^2 = 2$, $\sigma_\epsilon^2 = 1$, $B = 2$, and varying $T$. For Figures 3(c) and 3(d), we have set $B = 2$, $T = 3$, $\sigma_\tau^2 = 5$, $\sigma_\epsilon^2 = 1$, and let $\sigma_\beta^2$ vary.

`2wayANOVAgraphs`

An illustration of the behaviour of the three terms is in Figure 3. In Panel 3(a) we keep $\sigma_\tau^2 = \sigma_\beta^2 = 2$, $\sigma_\epsilon^2 = 1$, $B = 2$, and vary $T$. For values of $T$ larger than 8, Term 1 dominates the other two terms. Term 2 will reduce to zero as $T \to \infty$. Term 3 reduces to $a^{-1} = 0.4$. The PPV itself converges to 1.4. In Panel 3(b), it can be seen that, around $T = 20$, Term 2 can be omitted at a threshold of about 5%. In Panel 3(c), ($B = 2$, $T = 3$, $\sigma_\tau^2 = 5$, $\sigma_\epsilon^2 = 1$) Term 3 dominates, and Term 2 converges to 0.33, as $\sigma_\beta^2 \to \infty$. In Panel 3(d) as $\sigma_\beta^2 \to \infty$, all terms have limits strictly in $(0, 1)$.

In the above we have assumed we have only one replication of $\mathbf{Y} = (Y_{11}, \ldots, Y_{T,B}) = \mathbf{y}$ to predict the next outcomes. More generally, our reasoning can be extended to $n$ replications. The corresponding expressions for the terms in (5.3) will be more complex.

challenger

## 5.2 Application to Challenger disaster data

We apply the PPV decomposition to the Challenger disaster data, where the probability of O-ring failure on a space shuttle is predicted for a given temperature and pressure. This example was considered by Draper:1995 Draper (1995), who pointed out the need for uncertainty of the 'structural' choices. Below, we illustrate that the relative size of the terms in the three-term PPV decomposition indeed depends on the choice and the effectiveness of the parameters.

For the Challenger Space Shuttle disaster, it is widely believed that making the decision to launch the space shuttle at an ambient temperature and pressure at which various components had not been tested, ended up being catastrophic – and could have been avoided had a proper uncertainty analysis 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.

We can apply our techniques to two examples given in Draper:1995 Draper (1995) 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.

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.

Tab_challenger2

| $\mathcal{V}^{(2)}$ | $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ | $m_6$ | $m_7$ | $m_8$ | $m_9$ | $m_{10}$ | $m_{11}$ | $m_{12}$ | $m_{13}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $V_1$ | 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 expansion for within-structure and between-structure variances as

$$Var(p_{t=31}|\mathcal{D}_{23}) = Var_{within} + Var_{between} = 0.0338 + 0.0135 = 0.0473. \tag{5.4}$$

That is, Draper used a two term expansion based on Clause ii) of Prop. **??**. Draper's conclusion was that $.0135/.0473 \approx 28.5\%$ so the uncertainty represented by the second term in (5.4) 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 $V_2$. Details of our computational method are in Appendix **??**.

Considering all sources of uncertainty yields the PPV expansion

$$\begin{aligned} Var(p_{t=31}|\mathcal{D}_{23}) &= E_{V_1}E_{V_2}Var(p_{t=31}|\mathcal{D}_{23},V_1,V_2) + E_{V_1}Var_{V_2}E(p_{t=31}|\mathcal{D}_{23},V_1,V_2) \\ &+ Var_{V_1}E(p_{t=31}|\mathcal{D}_{23},V_1) \\ &= 0.0043 + 0.00087 + 0.00531 = 0.01048. \end{aligned} \tag{5.5}$$

var_challenger_3term

$$\begin{aligned} Var(p_{t=31,s=14.83}|\mathcal{D}_{23}) &= E_{V_1}E_{V_2}Var(p_{t=31}|\mathcal{D}_{23},V_1,V_2) + E_{V_1}Var_{V_2}E(p_{t=31}|\mathcal{D}_{23},V_1,V_2) \\ &+ Var_{V_1}E(p_{t=31}|\mathcal{D}_{23},V_1) \\ &= 0.0043 + 0.0058 + 0.00038 = 0.01048. \end{aligned} \tag{5.6}$$

var_challenger_3term

This is almost three times the variance as obtained by Draper. Thus, 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 see the last term is the smallest. Thus, we conclude that the third term can be taken as zero. Hence, we retain only the terms representing the between-models within-link functions variance and the between-predictions within-models and links variance. A frequentist testing approach confirms this; see Dustin:Ghosh:Clarke:2025 Dustin et al. (2025). So, we would be led to consider a new hierarchical model that did not include $V_1$ and therefore had a two term expansion using only $V_2$ giving a new value of $\text{Var}(Y_{n+1}|\mathcal{D}_n)$. In effect, we would compare this expansion with the first two terms on the right in (5.6) to see which expression for the PPV is more convincing.

So, if we drop $V_1$, and consider the two term expansion of $\text{Var}(Y_{n+1}|\mathcal{D}_n)$ we get

$$\begin{aligned} Var(p_{t=31}|\mathcal{D}_{23}) \quad &= E_{V_2}Var(p_{t=31}|\mathcal{D}_{23},V_2) \\ &+Var_{V_2}E(p_{t=31}|\mathcal{D}_{23},V_2) \\ &= 0.183 + 0.002 = 0.185. \end{aligned} \tag{5.7}$$

Again we see that the second number is quite small compared to the first. So even though we cannot drop $V_2$, we can drop the second term. Note that .185 in (5.7) is only slightly less than .188 in (5.6); this may be because all terms are necessarily non-negative so using a expansion with more terms can appear to increase the PPV purely due to random error. Overall, in this case we are left with Term 1 as the only important term for a PI.

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 expansion like that in **Oil Prices**:

$$Var(Y_{n+1}|\mathcal{D}_n) \quad = \quad E_{V_1}E_{V_2}Var(Y_{n+1}|\mathcal{D}_n,V_1,V_2) + E_{V_1}Var_{V_2}E(Y_{n+1}|\mathcal{D}_n,V_1,V_2)$$

$$+Var_{V_1}E(Y_{n+1}|\mathcal{D}_n, V_1) \tag{5.8}$$

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.

discuss

# 6 Discussion

The main contribution of this paper is to provide an expansion of the posterior predictive variance (PPV) for hierarchical models (HM's) and models that more generally may satisfy conditional independence assumptions that we have called structural. An immediate benefit from this is that we have a conservation law over expansions for the PPV. This is important for two reasons. First, the PPV 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 model 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 expansions start with a fixed hierarchical model and hence a fixed PPV that can be expressed in multiple expansions depending on the how use of the law of total variation is iterated. The various expansions depend on the ordering of the conditioning variables from the levels of the model. We focus on what we call the $C$-scope of a model – the collection of expansions 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. 4.1 we give an explicit expression for the cardinality of the $C$-scope. In Secs. 3 and 4 we give an extensive discussion of when a term being zero can be used to imply terms in other expansions must be zero.

The main methodological implication of our work is that we can more readily use HM's where we might have used Bayesian nonparametrics. Indeed, we can represent any feature of a statistical model with a prior. These features may or may not have any physical analog: in Subsec. 5.2 we use variable selection as a level in an HM and this is part of physical modeling. We also take the link function in a GLM as a feature and this is not in general seen as an aspect of modeling. Elsewhere, see Dustin et al. (2025), we used selection of a shrinkage method as a feature of modeling and this does not really have a physical analog; indeed it corresponds to selecting a prior.

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 HM so that its PPV is not too small relative to the data so that using multiple expansions to prune out levels will be effective. We also want the PPV to be not too large or the pruning will be obvious without analysis. We want the elements of the model list to be close enough to each other that they are plausible but far enough apart from each other that they are distinguishable with the dat we have. Indeed, we may want to construct a HM 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, see Theorem 3.2. Even though we can construct examples of arbitrarily 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 HM should be retained.

One drawback of our procedure is that it we do not have a fully formal way to assess the relative contributions of terms in the expansion. 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, although frequentist tests are possible, see Dustin et al. (2025). 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 with two final observations. First, the interplay between variance and conditioning is not well understood because variance is non-linear. Our methods provide a context where this interplay can be explored. Moreover, as we can see, each level of modeling can be treated as a level in an uncertainty quantification. So, given an ordering on the components in a HM, our analysis here is a step towards assigning an uncertainty to each component of the model.

Second, the treatment we have given for variance can, in principle, be extended to other nonlinear operations, though it looks hard. For instance, Brillinger (1969) gives a way to calculate cumulants of a distribution that can be a posterior quantity. He gives a formula similar to 4.2 and gives examples using this result for sums of variables and mixture distributions. 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 prediction intervals.

calcs3termnormal

## Appendix A: Calculations for the Three Term Normal

Our task is to derive an expression for $\mathsf{Var}(Y_{n+1}|y^n)$ directly. Recall (2.8) and that we have two parameters $\mu$ and $\lambda$ as well as three hyperparameters $\kappa_0$, $\alpha_0$, and $\beta_0$. For simplicity, write $\gamma = \lambda^2$. The conditional density of $y^n$ given $\mu$ and $\lambda^2$ 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

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

*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 $\mu$ means (A.4) becomes

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

Note that the 'extra' $\sqrt{\gamma}$ in (A.1) is absorbed into 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 $\mathsf{Var}(Y_{n+1}|y^n)$ 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 $\mu$) 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 $\mu$ and the exponent in the second factor is

$$y_{n+1}^2 + \kappa_n\mu_n^2 - \frac{(y_{n+1}+\kappa_n\mu_n)^2}{1+\kappa_n}$$

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

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

To complete the derivation of the PPV, write

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

since $\mu$ does not depend on $\gamma$.

calcs2wayANOVA

## Appendix B: Two-Way Bayesian ANOVA

Here we give the details for working out the three term PPV expansion for a two-way random effects ANOVA from Subsec. 5.1.

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

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

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

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

*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, $\tau$ is the vector of $\tau_i$'s, and the subscript + indicates a sum over the appropriate index.

To see this, set up a completing the square in $\beta_j$. That is, write

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\sum_i & \left[\tau_i^2\left(\left(\frac{B}{\sigma_\epsilon^2}+\frac{1}{\sigma_\tau^2}\right)-\frac{\frac{B}{\sigma_\epsilon^2}}{\frac{T}{\sigma_\epsilon^2}+\frac{1}{\sigma_\beta^2}}\right)-2\tau_i\left(\frac{y_{i+}}{\sigma_\epsilon^2}+\frac{\tau_0}{\sigma_\tau^2}\right)\left(\frac{\frac{y_{++}}{\sigma_\beta^2}+\frac{B\beta_0}{\sigma_\beta^2}}{\sigma_\epsilon^2}\right)\right. \\
&- \left.\frac{2B}{\sigma_\epsilon^4\left(\frac{T}{\sigma_\epsilon^2}+\frac{1}{\sigma_\beta^2}\right)}\tau_i\sum_{k\neq i}\tau_k\right] \\
&= \sum_i\left(\tau_i^2(a+Bb)-2\tau_i v_i+2Bb\tau_i\sum_{k\neq i}\tau_k\right) \\
&= \tau^T\begin{pmatrix} a+Bb & Bb & \dots & \dots & Bb \\ Bb & a+Bb & \dots & \dots & Bb \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Bb & Bb & \dots & a+Bb & Bb \\ Bb & Bb & \dots & Bb & a+Bb \end{pmatrix}\tau-2\tau^T V \\
&\equiv \tau^T A+2\tau^T \\
&= \tau^T A\tau-2\tau^T\mu_\tau \qquad \text{(B.8)}
\end{aligned}$$

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

$$A = aI_T + Bb\mathbf{1}\mathbf{1}^T. \qquad \text{(B.9)}$$

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

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

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

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

for some $n\times n$ matrix $\Sigma$. Writing $a_{ij}$ for the elements of $A$ and $\sigma_{ij}^{(-1)}$ for the elements in $\Sigma$ 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

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

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 $\mathrm{Var}(Y_{ij;n+1}|\mathbf{y})$. By two uses of the LTV we have

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

The first term is

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

The second term is

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

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

The third term in (B.14) is

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

$$+\frac{2}{\sigma_\epsilon^2}\left(\frac{T+1}{\sigma_\epsilon^2}+\frac{1}{\sigma_\beta^2}\right)(T-1)\frac{-Bb}{a(a+BbT)}\Bigg). \tag{B.16}$$

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

**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 expansions proposed here.

## References

Bachoc:etal:2019 Bachoc, F., Leeb, H., and Pötscher, B. (2019). "Valid confidence intervals for post-model selection predictions." *Ann. Statist.*, 47: 14751504. 1

Brillinger:1969 Brillinger, D. (1969). "The calculation of cumulants by conditioning." *Ann. Inst. Math. Stat.*, 21: 215–218. 31

Cover:Thomas:2006 Cover, T. and Thomas, J. (2006). *Elements of Information Theory, 2nd Ed.*. New York: John Wiley and Sons. 14

Draper:1995 Draper, D. (1995). "Assessment and Propagation of Model Uncertainty." *J. R. S. S. B*, 57(1): 45–97. 23, 24, 27

stin:Ghosh:Clarke:2025 Dustin, D., Ghosh, S., and Clarke, B. (2025). "Testing for the Important Components of Posterior Predictive Variance." *Stat. Anal. and Data Mining*, 18. 9, 29, 30, 31

Hoeting:etal:1999 Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). "Bayesian Model Averaging: A Tutorial." *Statist. Sci.*, 14(4): 382–417. 9

Lehmann:1998 Lehmann, E. and Casella, G. (1998). *Theory of Point estimation 2nd Ed.*. New York: Springer. 18

Liang:etal:2025 Liang, R., Zhu, W., and Barber, R. F. (2025). "Conformal prediction after data-dependent model selection." *arXiv:2408.07066*. 1

Qian:etal:2025b Qian, C., Liang, F., and Adams, J. (2025a). "Extending Temperature Scaling with Homogenizing Maps." *J. Machine Learning Res.*, 26: 1–46. 3

Qian:etal:2025a Qian, C., Ries, D., Liang, F., and Adams, J. (2025b). "Regression Recalibration by Learning PIT Map Values." *Technometrics*, 67: 481–492. 3

Shen:etal:2004 Shen, X., Huang, H., and Ye, J. (2004). "Inference after model selection." *J. Amer. Stat. Assoc.*, 99: 751–762. 1

Tian:2020 Tian, X. (2020). "Prediction error after model search." *Ann. Statist.*, 48: 763–784. 1