Electronic Journal of Statistics
Vol. 12 (2018) 4640–4668
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1511

Exact and efficient inference for partial Bayes problems

Yixuan Qiu, Lingsong Zhang, and Chuanhai Liu

Department of Statistics, Purdue University
West Lafayette, Indiana 47907, USA
e-mail: yixuanq@purdue.edu*;* lingsong@purdue.edu*;* chuanhai@purdue.edu

Abstract: Bayesian methods are useful for statistical inference. However, real-world problems can be challenging using Bayesian methods when the data analyst has only limited prior knowledge. In this paper we consider a class of problems, called *partial Bayes* problems, in which the prior information is only partially available. Taking the recently proposed inferential model approach, we develop a general inference framework for partial Bayes problems, and derive both exact and efficient solutions. In addition to the theoretical investigation, numerical results and real applications are used to demonstrate the superior performance of the proposed method.

Keywords and phrases: Confidence distribution, empirical Bayes, exact inference, inferential model, partial prior.

Received March 2018.

1. Introduction

In many real-world statistical problems, the information that is available to the data analysts can be organized in a hierarchical structure. That is, there exists some past experience about the parameter(s) of interest, and data relevant to the parameter(s) are also collected. For this type of problems, the standard approach to statistical inference is the Bayesian framework. However, in many applications, the data analysts have only limited prior knowledge. For instance, past experience may indicate that the prior belongs to a known distribution family, but the actual parameters of the prior are unclear. This type of problems have brought many challenges to statisticians; see for example Lambert and Duncan (1986); Meaux, Seaman Jr and Young (2002); Moreno, Bertolino and Racugno (2003). To systematically study such problems that involve partial prior information, in this article we refer to them as *partial Bayes* problems, with the precise definition given in Section 4.1. As a preview of the full characterization, in partial Bayes problems we assume that a genuine prior distribution exists for the model parameters, but due to the limitation of knowledge, some components of the full prior are missing. The target is to use the available data to make valid inference about the parameter of interest without assuming additional subjective hyper-priors for the desired but missing information.

Partial Bayes problems have drawn a lot of attention in statistics literature. One common problem of this type is the case where there exists an unknown prior distribution, either parametric or non-parametric, in a Bayesian hierarchical model. A very popular approach to this type of models is known as the empirical Bayes, which has been first proposed by Robbins (1956) for handling the case with non-parametric prior distributions, and later by Efron and Morris (1971, 1972a,b, 1973, 1975) for parametric prior distributions. Another kind of partial Bayes problems was studied by Xie et al. (2013), in which only the marginal distributions of a parameter vector are known, but the joint prior distribution is missing. For clarity, we refer to this type as the marginal prior problem. In Xie et al. (2013), the solution to the marginal prior problem is based on the confidence distribution approach (Xie, Singh and Strawderman, 2011), which provides a unified framework for meta-analysis.

The empirical Bayes and confidence distribution approaches both have successful real-world applications. However, one fundamental problem in scientific research, the exact inference about the parameter of interest, remains to be an open question for partial Bayes problems. As pointed out by many authors (Morris, 1983; Laird and Louis, 1987; Carlin and Gelfand, 1990), empirical Bayes in general underestimates the associated uncertainty of the interval estimators, so these authors have proposed various methods to correct the bias of the coverage rate. However, even if they have shown better performance, the target coverage rates are still approximately achieved for such methods. The same issue happens in the confidence distribution framework. Confidence distribution provides a novel way to combining different inference results, but these individual inferences may or may not be exact. All of these indicate that the exact inference for partial Bayes problems is highly non-trivial.

Recently, the inferential model framework (Martin and Liu, 2013, 2015a,b,c) has been proposed as a new approach to statistical inference, which not only provides Bayesian-like probabilistic measures of uncertainty about the parameter, but also has an automatic long-run frequency calibration property. In this paper, we use this framework to derive interval estimators for the parameter of interest in partial Bayes problems, and demonstrate their important statistical properties including the exactness and efficiency. When compared with other approaches, we refer to the proposed estimators as partial Bayes solutions for brevity.

The remaining part of this article is organized as follows. In Section 2 we study a hierarchical normal-means model as a motivating example of partial Bayes problems. In Section 3 we provide a brief review of the inferential model framework as the theoretical foundation of our analysis. Section 4 is the main part of this article, where we introduce a general framework for studying partial Bayes problems, and deliver our major theoretical results. We revisit some popular partial Bayes models in Section 5, and conduct simulation studies to numerically compare the proposed solutions with other methods. In Section 6 we consider an application to a basketball game dataset, and finally in Section 7 we conclude with a few remarks. Proofs of theoretical results are given in the appendix.

2. A motivating example

Consider the well-known normal hierarchical model for the observed data $X = (X_1, \ldots, X_n)'$. The model introduces n unobservable means $\mu_1, \ldots, \mu_n$, one for each observation, and assumes that conditional on μ_i's, X_i's are mutually independent with $X_i|\{\mu_1, \ldots, \mu_n\} \sim \mathsf{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$, where the common variance σ^2 is known. In addition, all the μ_i's are i.i.d. with $\mu_i \sim \mathsf{N}(\mu, \tau^2)$ for $i = 1, \ldots, n$, where the variance τ^2 is known but the mean μ is an unknown hyper-parameter.

The problem of interest here is to make marginal inference about the individual means μ_i, and for simplicity we focus on μ_1, as the approach is the same for all individual μ_i's. The aim of the inference is to construct a sample-based interval estimator for μ_1, denoted by $C_\alpha(X)$, which satisfies $P_{\mu_1,X}(C_\alpha(X) \ni \mu_1) \geq 1 - \alpha$ for all μ, a condition given by Morris (1983). The probability $P_{\mu_1,X}$ indicates that the coverage rate is computed over the joint distribution of (X, μ_1). We emphasize that when such a condition holds strictly, $C_\alpha(X)$ is said to be a *valid* or *exact* interval estimator for μ_1 with $100(1-\alpha)\%$ confidence level. This concept of exactness is made more clear in Section 4.1.

The standard empirical Bayes approach to this problem can be found in Efron (2010). It computes the maximum likelihood estimator (MLE) of μ, $\hat{\mu} = \overline{X}$, from the observed data. Plugging $\hat{\mu}$ back into the prior in place of μ, empirical Bayes proceeds with the standard Bayesian procedure to provide an approximate posterior distribution of μ_1, $\mu_1|X \mathrel{\dot\sim} \mathsf{N}\left((1-\omega)X_1 + \omega\overline{X}, (1-\omega)\sigma^2\right)$, where $\omega = \sigma^2/(\tau^2+\sigma^2)$, and the notation "$\dot\sim$" indicates that the distribution is approximate. Accordingly, the $100(1-\alpha)\%$ empirical Bayes interval estimator for μ_1 is obtained as

$$(1-\omega)X_1 + \omega\overline{X} \pm z_{\alpha/2}\sigma\sqrt{1-\omega},$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The partial Bayes solution, derived in Section 5.1.1, has a slightly different formula:

$$(1-\omega)X_1 + \omega\overline{X} \pm z_{\alpha/2}\sigma\sqrt{1-\omega(n-1)/n}. \tag{1}$$

Compared with empirical Bayes, the proposed interval has the same center but is slightly wider for small n. For a numerical illustration, we fix $\alpha = 0.05$, and take $\mu = 0$, $\sigma^2 = \tau^2 = 1$. Figure 1 shows the theoretical coverage rates of both the empirical Bayes solution and the partial Bayes solution as a function of n. It can be seen that the coverage probability of the empirical Bayes interval is less than the nominal value $1-\alpha$, and is close to the target only when n is sufficiently large. On the contrary, the partial Bayes solution correctly matches the nominal coverage rate for all n.

3. A brief review of inferential models

Since our inference for partial Bayes problems is based on the recently developed inferential models, in this section we provide a brief introduction to this new

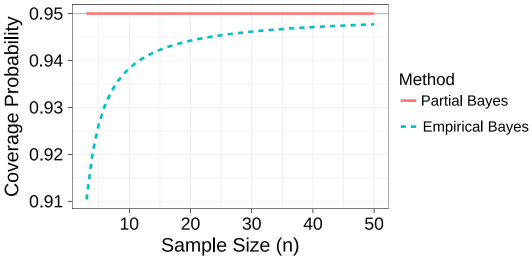

FIG 1. *The coverage probabilities of empirical Bayes (blue dashed curve) and partial Bayes (red solid line) as a function of n. The line for partial Bayes is exactly positioned at the 0.95 level, indicating that it achieves the nominal coverage rate exactly for all n.*

framework, with more details given in Martin and Liu (2013). The inferential model is a new framework designed for exact and efficient statistical inference. The exactness of inference, formally termed as *validity* in the inferential model framework, guarantees that the uncertainty of the generated result is appropriately quantified. On the premise of validity, the framework also provides a number of techniques to efficiently combine information in the data, some of which are mentioned at the end of this section.

Formally, inferential models draw statistical conclusions on an assertion A, a subset of the parameter space Θ, about the parameter of interest θ. For example, the subset $A = \{0\}$ stands for the assertion $\theta = 0$, and $A = (1, +\infty)$ corresponds to $\theta > 1$. In the inferential model framework, two quantities are used to represent the knowledge about A contained in the data: the *belief function*, which describes how much evidence in the data supports the claim that "A is true", and the *plausibility function*, which quantifies how much evidence does not support the claim that "A is false".

Like Fisher's fiducial inference, inferential models make use of auxiliary or unobserved random variables to represent the sampling model. In order to have meaningful probabilistic inferential results, unlike Fisher's fiducial inference, inferential models predict unobserved realizations of the auxiliary variables using random sets, and propagate such uncertainty to the space of θ. Technically, inferential models are formulated as a three-step procedure to produce the inferential results:

Association step This step specifies an association function $X = a(\theta, U)$ to connect the parameter $\theta \in \Theta$, the observed data $X \in \mathbb{X}$, and the unobserved auxiliary random variable $U \in \mathbb{U}$ with U following a known distribution P_U. This relationship implies that the randomness in the data is represented by an auxiliary variable U.

Prediction step Let u^* be the true but unobserved value of U that "generates" the data. This step constructs a valid predictive random set, $\mathcal{S}$, to predict u^*. $\mathcal{S}$ is valid if the quantity $Q_{\mathcal{S}}(u^*) = P_{\mathcal{S}}(\mathcal{S} \ni u^*)$, interpreted as the probability that $\mathcal{S}$ successfully covers u^*, satisfies the condition $P_U(Q_{\mathcal{S}}(U) \geq 1-\alpha) \geq 1-\alpha$, where $U \sim \mathsf{P}_U$.

Combination step This step transforms the uncertainty from the $\mathbb{U}$ space to the Θ space by defining $\Theta_x(\mathcal{S}) = \bigcup_{u\in\mathcal{S}} \Theta_x(u) = \bigcup_{u\in\mathcal{S}}\{\theta : x = a(u,\theta)\}$, a mapping from U back to θ after incorporating the uncertainty represented by $\mathcal{S}$. Then for an assertion A, its belief function is defined as $\mathsf{bel}_x(A) = P\{\Theta_x(\mathcal{S}) \subseteq A | \Theta_x(\mathcal{S}) \neq \varnothing\}$, and similarly, its plausibility function is defined as $\mathsf{pl}_x(A) = 1 - \mathsf{bel}_x(A^c)$.

The plausibility function is very useful to derive frequentist-like confidence regions for the parameter of interest (Martin, 2015, 2017). If we let A be a singleton assertion $A = \{\theta\}$ and denote $\mathsf{pl}_x(\theta) \equiv \mathsf{pl}_x(\{\theta\})$, then a $100(1-\alpha)\%$ frequentist-like confidence region, which is termed as *plausibility region* in inferential models (or *plausibility interval* as a special case), is given by $\mathsf{PR}_x(\alpha) = \{\theta : \mathsf{pl}_x(\theta) > \alpha\}$. It is worth mentioning that the inferential model theory guarantees $\mathsf{PR}_x(\alpha)$ to possess at least $100(1-\alpha)\%$ long-run coverage probability as long as the underlying $\mathcal{S}$ is valid.

Besides validity, the inferential model framework also has a number of extensions for efficient inference. When the model has multiple parameters but only some of them are of interest, the marginal inferential models (Martin and Liu, 2015b) appropriately integrate out the nuisance parameters. For models where the dimension of auxiliary variables is higher than that of the parameters, the conditional inferential models (Martin and Liu, 2015a) are used to combine information in the data such that efficient inference can be achieved. We elaborate the inference procedure for partial Bayes problems in Section 4.2, where both techniques mentioned above are used extensively in our development of exact and efficient inference.

4. Inference for partial Bayes problems

4.1. Model specification

Our attempt here is to provide a simple model framework that is general enough to describe a broad range of partial Bayes problems introduced in Section 1.

Let X be the observed data, whose distribution f relies on an unknown parameter vector θ. The information on θ that comes from the collected data is expressed by the conditional distribution of X given the parameter: $X|\theta \sim f(x|\theta)$. In many cases, we have prior knowledge about θ that can be characterized as a prior distribution $\pi_0(\theta)$. When $\pi_0(\theta)$ is fully specified, standard Bayesian method can be used to derive the posterior distribution of θ. In other cases, there is only partial prior information available. Formally, assume that the parameter θ can be partitioned into two blocks, $\theta = (\tilde{\theta}, \theta^*)$, so that the desirable fully-specified prior of θ can be accordingly decomposed as $\pi_0(\theta) = \pi(\tilde{\theta})\pi^*(\theta^*|\tilde{\theta}) = \pi(\tilde{\theta}|\theta^*)\pi^*(\theta^*)$, where $\pi(\tilde{\theta})$ and $\pi^*(\theta^*)$ are the marginal distributions of $\tilde{\theta}$ and

θ^*, respectively, and $\pi^*(\theta^*|\tilde{\theta})$ and $\pi(\tilde{\theta}|\theta^*)$ are the associated conditional density functions.

We define a partial Bayes problem to be such a Bayesian model in which either $\pi^*(\theta^*|\tilde{\theta})$ or $\pi^*(\theta^*)$ is missing. In other words, the prior information is partial in the sense that only a marginal distribution $\tilde{\theta} \sim \pi(\tilde{\theta})$ or a conditional distribution $\tilde{\theta}|\theta^* \sim \pi(\tilde{\theta}|\theta^*)$ for $\tilde{\theta}$ is available. In general, inference is made on $\tilde{\theta}$ or a component of $\tilde{\theta}$, *i.e.*, $\tilde{\theta}$ can be further partitioned into $\tilde{\theta} = (\eta, \xi)$, with η denoting the parameter of interest and ξ denoting the additional nuisance parameters. In this article we focus on the case that η is a scalar, which is of interest for many practical problems. For better presentation, we summarize these concepts and the proposed model structure in the following table:

Sampling model	$X|\theta \sim f(x|\theta)$
Parameter partition	$\theta = (\tilde{\theta}, \theta^*)$, $\tilde{\theta} = (\eta, \xi)$
Partial prior	$\tilde{\theta} \sim \pi(\tilde{\theta})$ or $\tilde{\theta}|\theta^* \sim \pi(\tilde{\theta}|\theta^*)$
Missing information	$\pi^*(\theta^*|\tilde{\theta})$ or $\pi^*(\theta^*)$
Parameter of interest	η

Despite its simplicity, the above model includes the well-known hierarchical models as an important class of practically useful models. Moreover, the formulation goes beyond the hierarchical models, and also includes the marginal prior problem. As described in Section 1, our target of inference is to construct a sample-based interval $C(X)$ that satisfies some validity conditions. Specifically, the following two types of validity properties are considered:

Definition 1. $C(X)$ is said to be an *unconditionally valid* interval estimator for η with $100(1-\alpha)\%$ confidence level, if $P_{X,\theta}(C(X) \ni \eta) \geq 1-\alpha$ for all $\pi^*(\theta^*)$, where the probability is computed over the joint distribution of (X, θ).

Definition 2. $C(X)$ is said to be a *conditionally valid* interval estimator for η given $H(X)$ with $100(1-\alpha)\%$ confidence level, if $P_{X,\theta|H(X)}(C(X) \ni \eta|H(X) = h) \geq 1-\alpha$ for all $\pi^*(\theta^*)$ and h, where $H(X)$ is a statistic of the data, and the probability is computed over the joint distribution of (X, θ) given $H(X) = h$.

Definition 1 is a rephrasing of the validity condition in Morris (1983), and Definition 2 comes from Carlin and Gelfand (1990). It should be noted that the second condition is stronger than the first, since it can be reduced to Definition 1 by averaging over $H(X)$. In this article, we aim to produce the second type of interval estimators, but the first validity property is studied when different interval estimators for η are compared with each other.

4.2. Procedure of inference

4.2.1. The association step

Constructing the data and prior associations Given the data sampling model $X|\theta \sim f(x|\theta)$, the first association equation can be constructed as $X =$

$a_1(\theta, W_1)$, where $a_1(\cdot)$ is the "data association" function, and W_1 is an unobservable auxiliary variable that has a known distribution. According to the definition of partial prior, θ can be partitioned into $\theta = (\tilde{\theta}, \theta^*)$, and either $\pi(\tilde{\theta})$ or $\pi(\tilde{\theta}|\theta^*)$ is available to the data analyst. In both cases, a unified "prior association" function $a_2(\cdot)$ can be formed as $\tilde{\theta} = a_2(\theta^*, W_2)$, where W_2 is another auxiliary variable independent of W_1. Note that in the first case, $\tilde{\theta}$ has a known marginal distribution $\pi(\tilde{\theta})$, so the prior association can be simplified as $a_2(\theta^*, W_2) \equiv W_2$.

Next, by substituting the prior association into the data association, we get $X = a_1((a_2(\theta^*, W_2), \theta^*), W_1)$. To avoid the over-complicated notations, we simply write this relation as $X = a(\theta^*, W)$, where $W = (W_1, W_2)$. As described in Section 4.1, we are only interested in an element of the $\tilde{\theta}$ vector, so we divide the system of equations $\tilde{\theta} = a_2(\theta^*, W_2)$ into $\eta = a_\eta(\theta^*, V_\eta)$ and $\xi = a_\xi(\theta^*, V_\xi)$, where $a_\eta(\cdot)$ and $a_\xi(\cdot)$ are the decomposed associations, and random vectors V_η and V_ξ are functions of W_2. As a consequence, the partial Bayes model can be summarized by the following system of three equations:

$$X = a(\theta^*, W), \ \eta = a_\eta(\theta^*, V_\eta), \ \text{and } \xi = a_\xi(\theta^*, V_\xi). \tag{2}$$

Note that ξ can be regarded as a nuisance parameter, and (2) is "regular" in the sense of Definition 3 of Martin and Liu (2015b). Then according to the theory of marginal inferential models (Theorems 2 and 3 of Martin and Liu, 2015b), the third equation in (2) can be ignored without loss of efficiency.

Decomposing the data association Since the sample X usually contains multiple observations, the dimension of W can often be very high. In order to reduce the number of auxiliary variables, assume that the relation $X = a(\theta^*, W)$ admits a decomposition

$$T(X) = a_T(\theta^*, \tau(W)), \ \text{and } H(X) = \rho(W) \tag{3}$$

for one-to-one mappings $x \mapsto (T(x), H(x))$ and $w \mapsto (\tau(w), \rho(w))$. Martin and Liu (2015a) shows that this decomposition broadly exists for a large number of models, and the most common way to constructing such a decomposition is to take $T(X)$ as a sufficient statistic for θ^* and $H(X)$ an ancillary statistic. However, it is worth mentioning that the decomposition (3) is more general than the sufficiency reduction; see Section 5.1 of Martin and Liu (2015a) for an example in which a non-trivial sufficient statistic does not exist but (3) is still available. Lastly, in case that such a decomposition is not available, we simply take $H(X) = 1$ and $\rho(W) = 1$.

The equation (3) implies that when the collected data have a realization x, the auxiliary variable $W_H \coloneqq \rho(W)$ is fully observed with the value $h \coloneqq H(x)$. Therefore, by conditioning on $W_H = h$, we are able to predict the remaining auxiliary variables more efficiently via the following two conditional associations:

$$T(X) = a_T(\theta^*, W_T), \quad W_T \coloneqq \tau(W) \sim \mathsf{P}_{W_T|h}, \tag{4}$$

$$\eta = a_\eta(\theta^*, V_\eta), \quad V_\eta \sim \mathsf{P}_{V_\eta|h}, \tag{5}$$

where the notation $Z \sim \mathsf{P}_{Z|h}$ means that the random variable Z has a distribution $\mathsf{P}_{Z|h}$ given $W_H = h$. In the rest of Section 4.2, when we discuss the distribution of a random variable that depends on W_T or V_η, the condition $W_H = h$ is implicitly added.

Obtaining the final association Finally, to make inference about η, the unknown quantity θ^* needs to be marginalized out of the equations. We seek a real-valued continuous function $b(\cdot, \cdot)$ such that when its first argument is fixed to some value t, the mapping $\eta \mapsto b(t, \eta)$ is one-to-one. At the current stage we simply take b as an arbitrary function, and we defer the discussion of its choice in Section 4.4. As a result, associations (4) and (5) are equivalent to

$$T(X) = a_T(\theta^*, W_T), \tag{6}$$

$$b(T(X), \eta) = W_b(\theta^*), \quad W_b(\theta^*) \coloneqq b(a_T(\theta^*, W_T), a_\eta(\theta^*, V_\eta)). \tag{7}$$

Equations (6) and (7) indicate that the only connection between η and θ^* is through $W_b(\theta^*)$, a random variable whose c.d.f. $F_{W_b(\theta^*)|h}$ is indexed by the nuisance parameter θ^*. Therefore, if the function b is chosen such that $F_{W_b(\theta^*)|h}$ does not depend on, or only weakly depends on θ^*, then the connection between η and θ^* is effectively broken. By the theory of marginal inferential models, the association (6) can be ignored, and thus equation (7) completes the association step.

4.2.2. The prediction step

The aim of the this step is to introduce a valid predictive random set $\mathcal{S}_h$ conditional on $W_H = h$ that can predict $W_b(\theta^*)$ with high probability. The following two situations are considered.

The first situation is that $W_b(\theta^*)$ is in fact free of θ^*. This can be easily achieved if θ^* has the same dimension as η, and if the mapping $\eta = a_\eta(\theta^*, V_\eta)$ can be inverted as $\theta^* = a_{\theta^*}(\eta, V_\eta)$. To verify this, plug $\theta^* = a_{\theta^*}(\eta, V_\eta)$ into (4), and we obtain $T(X) = a_T(a_{\theta^*}(\eta, V_\eta), W_T)$, which reduces to a single-parameter inferential model that has a well-defined solution.

The second situation is more general and thus more challenging, in which case $F_{W_b(\theta^*)|h}$ relies on the unknown parameter θ^*. Typically this occurs when the dimension of θ^* is higher than that of η. To deal with this issue, we generalize the Definition 5 of Martin and Liu (2015b) to define the concept of stochastic bounds for tails.

Definition 3. Let Z and Z^* be two random variables with c.d.f. F_Z and F_{Z^*} respectively, and denote by $\mathrm{med}(Z)$ the median of Z. Z is said to be stochastically bounded by Z^* in tails if $F_Z(z) \le F_{Z^*}(z)$ for $z < \mathrm{med}(Z)$, and $F_Z(z) \ge F_{Z^*}(z)$ for $z > \mathrm{med}(Z)$.

The difference between this definition and the one in the literature is that here the medians of Z and Z^* are not required to be zero.

Assume that we have found a random variable W_b^* such that given $W_H = h$, $W_b(\theta^*)$ is stochastically bounded by W_b^* in tails for any θ^*. Note that the first situation discussed earlier can be viewed as a special case, since any random variable is stochastically bounded by itself in tails. To shorten the argument, we only consider this more general case for later discussion. There are various ways to construct such a random variable W_b^*, see the examples in Martin and Liu (2015b). Here we provide a simple approach, by defining the c.d.f. to be

$$F_{W_b^*|h}(z) = \begin{cases} \sup_{\theta^*} F_{W_b(\theta^*)|h}(z), & z < m_h \\ \frac{1}{2}, & z = m_h \\ \inf_{\theta^*} F_{W_b(\theta^*)|h}(z), & z > m_h \end{cases}, \quad m_h = F^{-1}_{W_b(\theta^*)|h}\left(\tfrac{1}{2}\right),$$

provided that the resulting function is a c.d.f..

Given $F_{W_b^*|h}$, a standard conditional predictive random set $\mathcal{S}_h$ can be chosen for the prediction of $W_b(\theta^*)$. For the purpose of constructing two-sided interval estimators, we first define the generalized c.d.f. of a random variable Z as $F_Z^{-1}(u) = \inf\{x : F_Z(x) \geq u\}$, and then construct $\mathcal{S}_h$ as follows:

$$\mathcal{S}_h = \left\{ F^{-1}_{W_b^*|h}(u') : |u' - 0.5| < |U_{\mathcal{S}} - 0.5|, u' \in (0,1) \right\}, \ U_{\mathcal{S}} \sim \mathsf{Unif}(0,1). \tag{8}$$

This completes the prediction step, and other choices of the predictive random set for different purposes are discussed in Martin and Liu (2013).

4.2.3. The combination step

In what follows, to avoid notational confusions we use η to represent the parameter of interest as a random variable, and denote by $\tilde{\eta}$ the possible values of η. In the final combination step, denote by $\Theta_{T(x)}(w)$ the set of $\tilde{\eta}$ values that satisfy the association equation (7) with $T(X) = T(x)$ and $W_b(\theta^*) = w$, *i.e.*, $\Theta_{T(x)}(w) = \{\tilde{\eta} : b(T(x), \tilde{\eta}) = w\}$, and define $\Theta_{T(x)}(\mathcal{S}_h) = \bigcup_{s \in \mathcal{S}_h} \Theta_{T(x)}(s)$. Then the conditional plausibility function for η is obtained as

$$\mathsf{cpl}_{T(x)|h}(\tilde{\eta}) = 1 - P_{\mathcal{S}_h}\left(\Theta_{T(x)}(\mathcal{S}_h) \subseteq (-\infty, \tilde{\eta}) \cup (\tilde{\eta}, +\infty)\right) = P_{\mathcal{S}_h}\left(\Theta_{T(x)}(\mathcal{S}_h) \ni \tilde{\eta}\right), \tag{9}$$

which completes the combination step.

4.3. Interval estimator and validity of inference

In Section 4.2.3 a conditional plausibility function for the η parameter has been derived under the inferential model framework, and in this section it is used to construct the proposed interval estimator. Similar to the construction of plausibility region introduced in Section 3, we define the following set-valued function of x:

$$C_\alpha(x) = \{\tilde{\eta} : \mathsf{cpl}_{T(x)|h}(\tilde{\eta}) \geq \alpha\}. \tag{10}$$

From (9) it can be seen that $\mathsf{cpl}_{T(x)|h}(\tilde{\eta})$ depends on the data on two aspects: the random set $\mathcal{S}_h$ depends on $h = H(x)$, and the association function $\Theta_{T(x)}(w)$

depends on $T(x)$. As a result, we define our partial Bayes interval estimator for η to be $C_\alpha(X)$, obtained by plugging the random sample X into $C_\alpha(x)$.

In the typical case that η is a fixed value, the inferential model theory guarantees that $C_\alpha(X)$ is a valid $100(1-\alpha)\%$ frequentist confidence interval for η. However in our case, the joint distribution of the parameter and data is considered, as in Definitions 1 and 2. Therefore, the validity of $C_\alpha(X)$ does not automatically follow from the inferential model theory, and hence needs to be studied separately. The result is summarized as Theorem 1.

Theorem 1. *With $H(X)$ defined in (3), $C_\alpha(X)$ is a conditionally valid interval estimator for η given $H(X)$ with $100(1-\alpha)\%$ confidence level.*

Recall that if the decomposition (3) is unavailable, we will take $H(X) = 1$ and $\rho(W) = 1$. In such cases, Theorem 1 reduces to the unconditional result corresponding to Definition 1.

4.4. Bayesian-matching and efficiency

Under the guarantee that the proposed interval estimator $C_\alpha(X)$ defined in (10) satisfies the validity condition, we further study another important property, the efficiency of the estimator. Note that Section 4.1 introduces two types of partial prior: either $\pi^*(\theta^*)$ or $\pi^*(\theta^*|\tilde{\theta})$ is missing. In this section we study the former case for simplicity of argument, and the latter one can be dealt with analogously. The efficiency of $C_\alpha(X)$ is demonstrated based on the following two facts that will be verified later.

First, if the "oracle" information $\pi^*(\theta^*)$ is in fact known, then a full prior distribution for θ is available, and the optimal inference for η is via its posterior distribution given the data. For the proposed method, we can show that there exists a predictive random set, denoted by $\mathcal{S}_{h,t}$, which results in an "oracle" interval estimator $C_\alpha^o(X)$ that exactly matches the Bayesian credible interval.

Second, in the actual partial Bayes setting that $\pi^*(\theta^*)$ is unknown, $C_\alpha(X)$ is shown to be a good approximation to $C_\alpha^o(X)$ under mind conditions. This phenomenon indicates that the missing information $\pi^*(\theta^*)$ only results in a minor efficiency loss relative to the "oracle" solution, *i.e.*, the Bayesian credible interval with a full prior.

To show the first point, let $\theta^* = U,\ U \sim \pi^*(\theta^*)$ be the association equation for the marginal distribution of θ^*. Combining it with (6) and (7), we obtain the following three associations:

$$\theta^* = U,\ T(X) = Z_T,\ \text{and}\ b(T(X), \eta) = W_b, \tag{11}$$

where $Z_T = a_T(U, W_T)$ and $W_b = b(Z_T, a_\eta(U, V_\eta))$. Again, the second equation implies that given the data x, Z_T is fully observed with value $t := T(x)$, so the auxiliary variable W_b can be predicted using its conditional distribution given $W_H = h$ and $Z_T = t$, which we denote by $F_{W_b|h,t}$. Similar to the prediction step in Section 4.2.2, we construct a predictive random set $\mathcal{S}_{h,t}$ for W_b by replacing

$F^{-1}_{W_b^*|h}$ with $F^{-1}_{W_b|h,t}$ in formula (8), and proceed with the same combination step to obtain $\mathsf{cpl}_{T(x)|h,t}(\tilde{\eta}) = P_{\mathcal{S}_{h,t}}\left(\Theta_{T(x)}(\mathcal{S}_{h,t}) \ni \tilde{\eta}\right)$.

As a result, the interval estimator for η is obtained as $C^o_\alpha(X)$, where $C^o_\alpha(x) = \{\tilde{\eta} : \mathsf{cpl}_{T(x)|h,t}(\tilde{\eta}) \geq \alpha\}$. This $C^o_\alpha(x)$ and the $C_\alpha(x)$ in (10) are defined by the conditional plausibility functions $\mathsf{cpl}_{T(x)|h,t}$ and $\mathsf{cpl}_{T(x)|h}$, respectively, which only differ in the distributions assigned to the predictive random sets. The following theorem shows that with this slight change, $C^o_\alpha(X)$ matches the Bayesian posterior credible interval.

Theorem 2. *Assuming that $\pi^*(\theta^*)$ is known and η has a continuous distribution function $F_{\eta|x}$ given $X = x$, then $C^o_\alpha(X)$ is optimal in the sense that it matches the Bayesian posterior credible interval, i.e.,*

$$C^o_\alpha(x) = \left(F^{-1}_{\eta|x}(\alpha/2), F^{-1}_{\eta|x}(1-\alpha/2)\right).$$

Theorem 2 implies that given the "oracle" information $\pi^*(\theta^*)$, there exists an "oracle" predictive random set $\mathcal{S}_{h,t}$ for the auxiliary variable W_b with which the inference result attains the optimality. This fact suggests that when $\pi^*(\theta^*)$ is missing, as long as there exists a predictive random set close to $\mathcal{S}_{h,t}$, the resulting interval estimator would be as efficient as the "oracle", at least approximately.

Recall that the predictive random set $\mathcal{S}_{h,t}$ is induced by the distribution $F_{W_b|h,t}$, and when $\pi^*(\theta^*)$ is missing, only $F_{W_b(\theta^*)|h}$ is available. Therefore, the next question we want to answer is under which conditions $F_{W_b(\theta^*)|h}$ is close to $F_{W_b|h,t}$. Note that these two distributions are conditional on the same event $H(X) = h$, and to avoid technical complications we consider the unconditional case $H(X) \equiv 1$, so that $F_{W_b(\theta^*)|h}$ and $F_{W_b|h,t}$ are reduced to $F_{W_b(\theta^*)}$ and $F_{W_b|t}$, respectively, where the former is the c.d.f. of $W_b(\theta^*)$ defined in (7), and the latter stands for the distribution of W_b defined in (11) given $Z_T = t$.

In most real applications, the association relation for $T(X)$ changes with the data size n. To emphasize the dependence on n, in what follows we write $W_{b_n}(\theta^*)$, Z_{T_n}, and W_{b_n} in place of $W_b(\theta^*)$, Z_T, and W_b, respectively. The following definition from Sweeting (1989) and Xiong and Li (2008), which generalizes the usual concept of weak convergence, is needed to study the large sample property of a conditional distribution.

Definition 4. Given two sequences of random variables X_n and Y_n, the conditional distribution function of X_n given Y_n, a random c.d.f. denoted by $F_{X_n|Y_n}$, is said to converge weakly to a non-random c.d.f. F_Z in probability, denoted by $X_n|Y_n \overset{d.P}{\to} Z$, if for every continuous point z of F_Z, $F_{X_n|Y_n}(z) \overset{P}{\to} F_Z(z)$, where $Z \sim F_Z$.

Then we have the following result:

Theorem 3. *Let g_n, h_n, and p_n denote the densities of W_{b_n}, Z_{T_n}, and (W_{b_n}, Z_{T_n}), respectively. Also define $l_n(w,z) = p_n(w,z)/[g_n(w)h_n(z)]$. If (a) for fixed u, $a_T(u, W_{T_n}) \overset{P}{\to} u$, (b) $b(u, a_\eta(u,v)) = v$, and (c) $l_n \to 1$ pointwisely, then $W_{b_n}|Z_{T_n} \overset{d.P}{\to} V_\eta$ and $W_{b_n}(\theta^*) \overset{d}{\to} V_\eta$, where θ^* in $W_{b_n}(\theta^*)$ is seen as a fixed value.*

Remark 1. Conditions *(a)* and *(b)* are intentionally expressed in a simple form. In fact they can be replaced by $a_T(u, W_{T_n}) \overset{P}{\to} f_1(u)$ and $b(f_1(u), a_\eta(u,v)) = f_2(v)$ where f_1 and f_2 are one-to-one functions, and the limiting distribution is changed to $f_2(V_\eta)$ accordingly.

Remark 2. The three conditions are easy to check. Condition *(a)* states that $T(X)$ should be a consistent estimator for θ^* if θ^* is seen as fixed. Condition *(b)* guides the choice of the b function; see below for more discussions. For condition *(c)*, it is shown in the proof that $(W_{b_n}, Z_{T_n}) \overset{d}{\to} (V_\eta, U)$, and a sufficient condition for *(c)* is that the density of (W_{b_n}, Z_{T_n}) also converges to that of (V_η, U), which is satisfied by most parametric models.

To summarize, Theorem 3 indicates that $W_{b_n}(\theta^*)$ and $W_{b_n}|Z_{T_n}$ converge to the same limiting distribution, in which sense the random sets $\mathcal{S}_h$ and $\mathcal{S}_{h,t}$ have approximately identical distributions when n is sufficiently large. As a result, the proposed interval estimator $C_\alpha(X)$ can be seen as an approximation to the "oracle" solution $C_\alpha^o(X)$. Combining Theorem 1 and Theorem 3, it can be concluded that the proposed interval estimator possesses the favorable properties of both validity and efficiency.

Based on the results above, we are now ready to provide some guidelines on the choice of the $b(\cdot,\cdot)$ function introduced in (7). One important fact is that the validity of $C_\alpha(X)$ is unaffacted by $b(\cdot,\cdot)$, as is indicated by Theorem 1; therefore, $b(\cdot,\cdot)$ only influences the efficiency of the estimator. Intuitively, $b(\cdot,\cdot)$ plays the role of combining the information in the data equation (4) and prior equation (5), and meanwhile cancelling the effect of the nuisance parameter θ^*, *i.e.*, $\partial b/\partial\theta^* \approx 0$. Below we introduce two useful techniques that are shown to perform well for many practical models.

First, if the equation $\eta = a_\eta(\theta^*, V_\eta)$ is invertible as $\theta^* = a_{\theta^*}(\eta, V_\eta)$, then plugging it into (4) yields $T(X) = a_T(a_{\theta^*}(\eta, V_\eta), W_T)$, which is an association equation merely involving the paramter η. In this case, a desirable b satisfies $\partial b/\partial\eta \approx 0$. While for many problems this requirement does not hold globally, it can be localized such that $b(T(x), \hat\eta) = 0$ and $\partial b/\partial\eta|_{\eta=\hat\eta} = 0$ hold around a reasonable estimator $\hat\eta$ for η. Inspired by Martin (2015), if $\hat\eta$ is the MLE for η based on the conditional distribution of X_i given η, and $T(X)$ is a sufficient statistic, then we can take $b(T(x),\eta) = \ell(\hat\eta; T(x)) - \ell(\eta; T(x))$, where $\ell(\cdot; T(x))$ is the log-likelihood function. An example of this technique is given in Section 5.2 for a Poisson hierarchical model, and a variant of this method is used in the binomial model in Section 5.3.

Second, in more general settings, $b(\cdot,\cdot)$ is typically chosen by heuristics, and the condition *(b)* of Theorem 3 suggests the following construction. If η has a continuous distribution, then in general we can solve $V_\eta = a_V(\theta^*, \eta)$ from the association $\eta = a_\eta(\theta^*, V_\eta)$. Suppose that from (4), a sensible estimator for θ^*, $\hat\theta^* = t(T(x))$, can be obtained for some function $t(\cdot)$. Define $b(T(x),\eta) = a_V(t(T(x)),\eta)$, and we can show that such a $b(\cdot,\cdot)$ function is only weakly dependent on θ^* since $b(T(X),\eta) = a_V(\hat\theta^*,\eta) \approx a_V(\theta^*,\eta) = V_\eta$. This idea is applied to the normal hierarchical model in Section 5.1.2 where both the mean and variance parameters in the normal prior are missing.

5. Popular models viewed as partial Bayes problems

5.1. The normal hierarchical model

The normal hierarchical model introduced in Section 2 is extremely popular in the empirical Bayes literature, partly due to its simplicity and flexibility; existing inference methods include the naive empirical Bayes (Efron and Morris, 1975; Casella, 1985), the full Bayes method with flat prior (Lindley and Smith, 1972; Deely and Lindley, 1981), the approach used by Morris (1983) and Efron (2010), the bootstrap method (Laird and Louis, 1987), the conditional bias correction method (Carlin and Gelfand, 1990), etc. We set $\sigma^2 = 1$ without loss of generality, since X_i's can always be scaled by a constant to achieve an arbitrary variance. We will consider both the cases where τ^2 is known and unknown, and our parameter of interest is μ_1. To summarize, we write

Sampling model	$X\,\|\,(\tilde{\theta}, \theta^*) \sim \prod_i \mathsf{N}(\mu_i, \sigma^2)$
Partial prior	$\tilde{\theta} = (\mu_1, \mu_2, \ldots, \mu_n)$, $\tilde{\theta}\|\theta^* \sim \prod_i \mathsf{N}(\mu, \tau^2)$
Missing information	$\pi^*(\theta^*)$, $\theta^* = \begin{cases} \mu, & \text{if } \tau \text{ is known} \\ (\mu, \tau^2), & \text{if } \tau \text{ is unknown} \end{cases}$
Parameter of interest	$\eta = \mu_1$

As a first step, this model can be expressed by the following association equations: $\mu_i = \mu + \tau\varepsilon_i$ and $X_i = \mu_i + e_i$ for $i = 1, \ldots, n$, where $\varepsilon_i \overset{iid}{\sim} \mathsf{N}(0, 1)$, $e_i \overset{iid}{\sim} \mathsf{N}(0, 1)$, and e_i and ε_i are independent. An equivalent expression for these associations is $\mu_i = \mu + \tau\varepsilon_i, X_i = \mu + \tau\varepsilon_i + e_i$, in which the data are directly linked to the unknown μ. Since the focus is on μ_1, equations related to $\mu_2, \ldots, \mu_n$ can be ignored.

5.1.1. The case with a known τ^2

This case corresponds to the motivating example presented in Section 2, and we are going to derive formula (1) with $\sigma^2 = 1$. Since τ is known, let $W_i = \tau\varepsilon_i + e_i, i = 1, 2, \ldots, n$, and then the system of associations $X_i = \mu + \tau\varepsilon_i + e_i$ can be rewritten as $\overline{X} = \mu + \overline{W}$ and $X_i - X_1 = W_i - W_1$ for $i = 2, \ldots, n$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\overline{W} = \frac{1}{n}\sum_{i=1}^n W_i$. Therefore, by denoting $T(X) = \overline{X}$ and $H(X) = X_{(-1)} - X_1\mathbf{1}_{n-1}$, where $X_{(-1)} = (X_2, \ldots, X_n)'$ and $\mathbf{1}_{n-1}$ is a vector of all ones, the decomposition in equation (3) is achieved. The associated auxiliary variable for $H(X)$ is $W_H = W_{(-1)} - W_1\mathbf{1}_{n-1}$, where $W_{(-1)} = (W_2, \ldots, W_n)'$.

Next, we keep the following two associations $\overline{X} = \mu + \overline{W}$ and $\mu_1 = \mu + \tau\varepsilon_1$, where $\overline{W} \sim \mathsf{P}_{\overline{W}|h}$ and $\varepsilon_1 \sim \mathsf{P}_{\varepsilon_1|h}$ conditional on $W_H = h \equiv H(x)$. The last step is to take $b(\overline{X}, \mu_1) = \overline{X} - \mu_1$, and the final association equation is $b(\overline{X}, \mu_1) =$

$W_b \coloneqq \overline{W} - \tau\varepsilon_1$. It can be verified that the conditional distribution of W_b given $W_H = h$ is

$$W_b|\{W_H = h\} \sim \mathsf{N}\left(\frac{\tau^2}{1+\tau^2}(\bar{x} - x_1), \frac{n\tau^2+1}{n(\tau^2+1)}\right), \tag{12}$$

and the predictive random set (8) can be constructed accordingly. As a result, the conditional plausibility function for μ_1 is obtained as

$$\mathsf{cpl}_{T(x)|h}(\mu_1) = 2\Phi\left(-\left|\frac{\tau^2}{\tau^2+1}x_1 + \frac{1}{\tau^2+1}\bar{x} - \mu_1\right| \Bigg/ \sqrt{\frac{n\tau^2+1}{n(\tau^2+1)}}\right), \tag{13}$$

where Φ is the standard normal c.d.f., and hence the interval estimator for μ_1 is

$$C_\alpha(X) = \left(\frac{\tau^2}{\tau^2+1}X_1 + \frac{1}{\tau^2+1}\overline{X}\right) \pm z_{\alpha/2}\sqrt{\frac{n\tau^2+1}{n(1+\tau^2)}}. \tag{14}$$

5.1.2. *The case with an unknown τ^2*

Similar to the previous case, the starting point is to decompose the data associations into $T(X)$ and $H(X)$, which can be done in two stages as described below. In the first stage, we keep the association for X_1 and decompose $X_{(-1)}$ instead. Consider the ancillary statistics $H_i(X) = (X_i - \overline{X}_{(-1)})/S_{(-1)}$ for $i = 2, \dots, n$, where $\overline{X}_{(-1)}$ and $S^2_{(-1)}$ are the sample mean and sample variance of $X_{(-1)}$. It is clear that $X_{(-1)}$ has a one-to-one mapping to $(\overline{X}_{(-1)}, S^2_{(-1)}, H_2(X), \dots, H_{n-1}(X))$. Since marginally $X_i \overset{iid}{\sim} \mathsf{N}(\mu, \tau^2+1)$, it is well known that $(\overline{X}_{(-1)}, S^2_{(-1)})$ is a complete sufficient statistic for (μ, τ), and thus is independent of $H_i(X)$ according to Basu's theorem. Therefore, conditioning on $H_i(X)$ does not change the distribution of $(\overline{X}_{(-1)}, S^2_{(-1)})$, and we obtain the following four associations: *(a)* $\mu_1 = \mu + \tau\varepsilon_1$, *(b)* $X_1 = \mu + \tau\varepsilon_1 + e_1$, *(c)* $\overline{X}_{(-1)} = \mu + \tilde{\tau}Z$, and *(d)* $S^2_{(-1)} = (\tau^2+1)M^2_{n-2}$, where $\tilde{\tau} = \sqrt{(\tau^2+1)/(n-1)}$, $Z \sim \mathsf{N}(0,1)$, $M^2_{n-2} \sim \chi^2_{n-2}/(n-2)$, and the auxiliary variables ε_1, e_1, Z, and M^2_{n-2} are mutually independent. Equations *(c)* and *(d)* are derived from the well-known facts that $\overline{X}_{(-1)} \sim \mathsf{N}(\mu, \tilde{\tau}^2)$ and $(n-2)S^2_{(-1)}/(\tau^2+1) \sim \chi^2_{n-2}$.

Then in the second stage, we condition on the following equation, as the auxiliary variable W_H is known to follow a student t-distribution with $n-2$ degrees of freedom:

$$H(X) \coloneqq \sqrt{\frac{n-1}{n}} \cdot \frac{X_1 - \overline{X}_{(-1)}}{S_{(-1)}} = W_H \coloneqq \frac{\tau\varepsilon_1 + e_1 - \tilde{\tau}Z}{\sqrt{n}\tilde{\tau}M_{n-2}} \sim t_{n-2}. \tag{15}$$

As a result, we keep the associations $\mu_1 = \mu + \tau\varepsilon_1$, $\overline{X}_{(-1)} = \mu + \tilde{\tau}Z$, and $S^2_{(-1)} = (\tau^2+1)M^2_{n-2}$, with $\varepsilon_1 \sim \mathsf{P}_{\varepsilon_1|h}, Z \sim \mathsf{P}_{Z|h}$, and $M^2_{n-2} \sim \mathsf{P}_{M^2_{n-2}|h}$

conditional on $W_H = h \equiv H(x)$. Obviously in this case $T(X) = (\overline{X}_{(-1)}, S^2_{(-1)})$, which combined with $H(X)$ completes the decomposition.

Next, by observing that $\overline{X}_{(-1)} - \mu_1$ is free of μ, we can take $b(T(X), \mu_1)$ to be a function of $\overline{X}_{(-1)} - \mu_1$ and $S^2_{(-1)}$, so that the corresponding auxiliary variable $W_b(\tau)$ is indexed by only one unknown parameter τ. Specifically, let

$$\begin{aligned} \tilde{\mu} &= \sqrt{\frac{n-1}{n}} h \left(\frac{n-2}{h^2+n-2} S^{-1}_{(-1)} - S_{(-1)} \right), \\ \tilde{\sigma}^2 &= \max\left\{ n^{-\gamma}, 1 - \frac{(n-1)(n-2)(n-3-h^2)}{n(n-2+h^2)^2} S^{-2}_{(-1)} \right\}, \quad \gamma \in (0, \tfrac{1}{2}), \end{aligned} \tag{16}$$

and then define $b(T(X), \mu_1) = (\overline{X}_{(-1)} - \mu_1 - \tilde{\mu})/\tilde{\sigma}$, where $\tilde{\mu}$ and $\tilde{\sigma}$ are chosen such that $\mathbb{E}(W_b(\tau)|W_H = h) = 0$ and that $W_b(\tau)|\{W_H = h\} \xrightarrow{d} \mathsf{N}(0,1)$. These two conditions ensure that $W_b(\tau)$ will be gradually free of τ when n is large. Next, let $F_{W_b(\tau)|h}$ be the c.d.f. of $W_b(\tau)$ given $W_H = h$, and we can show that

$$\begin{aligned} &F_{W_b(\tau)|h}(s) \\ &\quad = \int_0^{+\infty} \Phi\left(\frac{s\sqrt{\max\{n^{-\gamma}, 1 - c_1\omega/x\}} - c_2\sqrt{\omega}\,(\sqrt{x} - c_3/\sqrt{x})}{\sqrt{1-\omega(n-1)/n}} \right) g(x)\mathrm{d}x, \end{aligned} \tag{17}$$

where $\omega = (1+\tau^2)^{-1}$, $c_1 = (n-2)(n-3-h^2)/\{n(h^2+n-2)\}$, $c_2 = (n-1)h/\sqrt{n(h^2+n-2)}$, $c_3 = (n-2)/(n-1)$, and g is the p.d.f. of $\chi^2_{n-1}/(n-1)$.

Finally, let $\underline{F}(s) = \inf_{\omega \in (0,1)} F_{W_b(\tau)|h}(s)$ and $\overline{F}(s) = \sup_{\omega \in (0,1)} F_{W_b(\tau)|h}(s)$, both computable using numerical methods, and we can show that

$$\mathsf{cpl}_{T(x)|h}(\mu_1) = \min\left\{ 1, 2\left[1 - \underline{F}\left(\frac{x_1 - \mu_1 - \tilde{\mu}}{\tilde{\sigma}}\right)\right], 2\overline{F}\left(\frac{x_1 - \mu_1 - \tilde{\mu}}{\tilde{\sigma}}\right) \right\},$$

and that

$$C_\alpha(X) = \left(\overline{X}_{(-1)} - \tilde{\mu} - \underline{F}^{-1}(1-\alpha/2)\tilde{\sigma},\ \overline{X}_{(-1)} - \tilde{\mu} - \overline{F}^{-1}(\alpha/2)\tilde{\sigma} \right).$$

To numerically compare the partial Bayes solution with other existing methods mentioned in the beginning of this section, we conduct a simulation study in which both hyper-parameters μ and τ^2 are assumed to be unknown, with the same setting in Laird and Louis (1987): the true μ is fixed to 0, and two values of τ, 0.5 and 1, are considered. For the partial Bayes solution, the γ constant in (16) is fixed to be $\frac{1}{3}$. The nominal coverage rate is set to 95%, and data are simulated 10,000 times in order to calculate the empirical coverage percentage and the mean interval width for all the methods compared. The results are summarized in Figure 2.

It is obvious in Figure 2 that among all the methods compared, only the partial Bayes solution achieves the nominal coverage rate for all sample sizes. In terms of interval width, the partial Bayes solution has wider interval estimates than other methods, due to the guarantee of coverage rate; however, as the sample size increases, the gaps between different methods become smaller and smaller, indicating that all methods are efficient asymptotically.

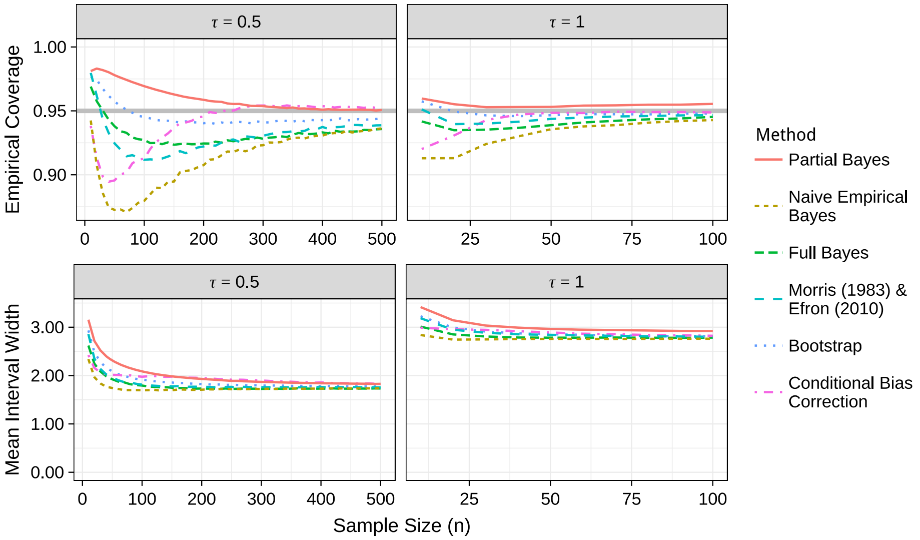

FIG 2. *The empirical coverage percentage (the top two panels) and mean interval width (the bottom two panels) for μ_1 with an increasing sample size n and two parameter settings, among 10,000 simulation runs. For all the methods compared, only the partial Bayes solution guarantees the nominal coverage rate for all n.*

5.2. The Poisson hierarchical model

The Poisson hierarchical model is useful for analyzing discrete data such as counts. Assume that given parameters $\lambda_i > 0$, the observed data $X = (X_1, \dots, X_n)'$ satisfy $X_i | \lambda_i \sim \mathsf{Pois}(\lambda_i t_i), i = 1, \dots, n$, where $t_i > 0$ are known constants. In real-world problems, λ_i can be interpreted, for example, as the rate of events in unit time, and t_i is the length of the time window. It is also assumed that λ_i's follow a common prior, $\lambda_i \overset{iid}{\sim} \gamma\mathsf{Gamma}(s)$, where s is a known shape parameter and γ is an unknown scale parameter. In this setting the parameter of interest is λ_1. This model can also be expressed using the formulation in Section 4.1:

Sampling model	$X \mid (\tilde{\theta}, \theta^*) \sim \prod_i \mathsf{Pois}(\lambda_i t_i)$
Partial prior	$\tilde{\theta} = (\lambda_1, \lambda_2, \dots, \lambda_n)$, $\tilde{\theta} \mid \theta^* \sim \prod_i \gamma\mathsf{Gamma}(s)$
Missing information	$\pi^*(\theta^*)$, $\theta^* = \gamma$
Parameter of interest	$\eta = \lambda_1$

For this Poisson hierarchical model, the data associations and prior associations are given by $X_i = F^{-1}_{\lambda_i t_i}(U_i)$ and $\lambda_i = \gamma V_i$, respectively, with $i = 1, \dots, n$. F^{-1}_{λ} is the generalized inverse c.d.f. of the Poisson distribution with mean λ, $U = (U_1, \dots, U_n)' \overset{iid}{\sim} \mathsf{Unif}(0,1), V = (V_1, \dots, V_n)' \overset{iid}{\sim} \mathsf{Gamma}(s)$, and U and V are independent. After plugging prior associations into data associations and ignoring irrelevant parameters, the following association equations are kept without loss of information:

$$\lambda_1 = \gamma V_1, \text{ and } X_i = F^{-1}_{\gamma V_i t_i}(U_i), \ i = 1, \ldots, n. \tag{18}$$

There is a fundamental difference between this Poisson model and the normal model studied earlier. Due to the discreteness of X_i and the heterogeneity of the t_i values, it is improbable to find a non-trivial function $H(x)$ such that the distribution of $H(X)$ is free of γ. This is an example that the decomposition (3) is not available, and hence we trivially take $H(X) = 1$ and $T(X) = X$.

Next, the b function is chosen using the technique introduced in Section 4.4. Let $\ell(\lambda_1; x)$ denote the log-likelihood function conditional on λ_1, and $\hat{\lambda}_1 = \hat{\lambda}_1(x)$ the MLE for λ_1. Define $b(x, \lambda_1) = \ell(\hat{\lambda}_1; x) - \ell(\lambda_1; x)$, and the final association becomes $b(X, \lambda_1) = W_b(\lambda_1)$, where $W_b(\lambda_1)$ is a random variable by replacing $x = (x_1, \ldots, x_n)'$ with

$$X = \left(F^{-1}_{\lambda_1 t_1}(U_1), F^{-1}_{\lambda_1 t_2 V_2/V_1}(U_2), \ldots, F^{-1}_{\lambda_1 t_n V_n/V_1}(U_n)\right)'$$

in the expression for $b(x, \lambda_1)$.

Let G_{λ_1} be the c.d.f. of $W_b(\lambda_1)$ conditional on λ_1, and then the plausibility function for λ_1 is $\mathsf{pl}_x(\lambda_1) = 1 - G_{\lambda_1}(b(x, \lambda_1))$. Finally, the interval estimator for λ_1 is obtained by inverting the plausibility function, *i.e.,* $C_\alpha(x) = \{\tilde{\lambda} : \mathsf{pl}_x(\tilde{\lambda}) \geq \alpha\}$. The computational details are given in Appendix 7.

We conduct a simulation study with 10,000 data repetitions and 95% nominal coverage rate. All the $t_i's$ are set to 1, and the true value of θ is fixed to 1. Two different values of s, $s = 2, 10$, and a sequence of sample sizes, $n = 10, 15, \ldots, 50$, are considered. There are fewer existing inference methods for the Poisson model than the normal one, and here the partial Bayes solution is compared with the naive empirical Bayes and full Bayes approaches, with the results illustrated in Figure 3.

The pattern of the simulation results is very similar to that of the normal model. As expected, the other two solutions have narrower interval estimates than the partial Bayes solution, but they do not preserve the nominal coverage rate. In contrast, the partial Bayes solution has coverage percentages above 95%, and its interval width is getting close to the other two when sample size increases. This interesting result indicates that even if Theorem 3 no longer applies to this model due to the choice of $T(X) = X$, the interval estimator derived in this section is in fact very efficient.

5.3. The binomial rate-difference model

The last binomial rate-difference model is motivated by a clinical trial study (Xie et al., 2013). It can be described as follows. Assume that two independent binomial samples, X and Y, were collected with $X \sim \mathsf{Bin}(m, p_1)$, and $Y \sim \mathsf{Bin}(n, p_2)$. The available prior information is on the difference of the success rates, $\delta := p_1 - p_2 \sim \pi$, and the task is to make inference about δ. For this model, we have

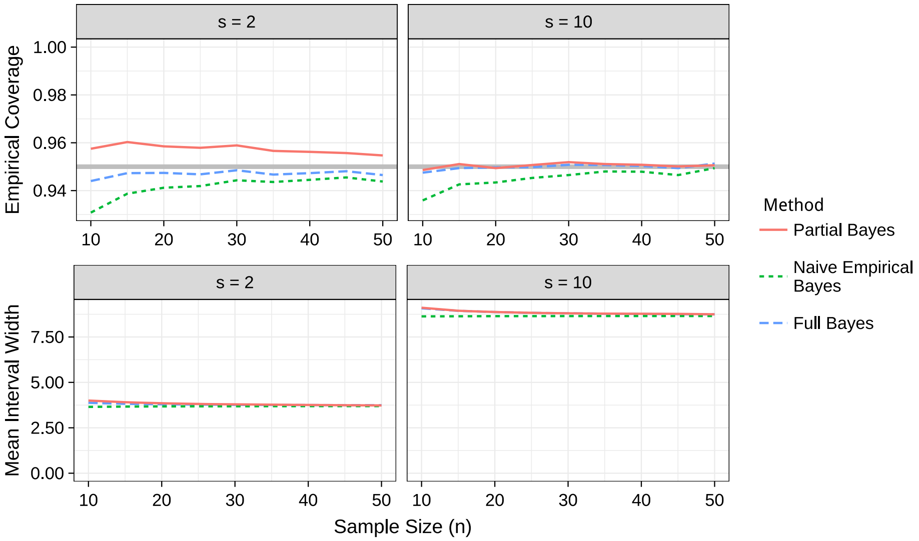

FIG 3. *The empirical coverage percentage (the top two panels) and mean interval width (the bottom two panels) for λ_1 with an increasing sample size n and two parameter settings, among 10,000 simulation runs. Three different solutions are compared, showing that the partial Bayes solution guarantees the nominal coverage rate.*

Sampling model	$X\|(\tilde{\theta}, \theta^*) \sim \mathsf{Bin}(m, p_1)$, $Y\|(\tilde{\theta}, \theta^*) \sim \mathsf{Bin}(n, p_2)$
Partial prior	$\tilde{\theta} = \delta := p_1 - p_2$, $\tilde{\theta} \sim \pi$
Missing information	$\pi^*(\theta^*\|\tilde{\theta})$, $\theta^* = p_1 + p_2$
Parameter of interest	$\eta = \delta$

Obviously, the data association equations of this model are $X = F^{-1}_{m,p_1}(U_1)$ and $Y = F^{-1}_{n,p_2}(U_2)$, and the prior association is $\delta = U$, where $F^{-1}_{k,p}$ is the generalized inverse c.d.f. of $\mathsf{Bin}(k,p)$. The auxiliary variables $U_1, U_2 \overset{iid}{\sim} \mathsf{Unif}(0,1)$, $U \sim \pi$, and U_1, U_2, and U are independent. To simplify the notations, p_1 and p_2 are re-parameterized as $\delta = p_1 - p_2$ and $\tau = p_1 + p_2$. Since p_1 and p_2 must lie in $[0,1]$, τ is further written as $\tau = 1 + (1 - |\delta|)\omega$ to guarantee the range, where $\omega \in (-1,1)$ is an unknown quantity. As a result, $p_1 = p_1(\delta,\omega) = \{1+\delta+(1-|\delta|)\omega\}/2$ and $p_2 = p_2(\delta,\omega) = \{1-\delta+(1-|\delta|)\omega\}/2$ are functions of the new parameters δ and ω.

Similar to the association steps of previously studied models, we first plug the prior association into the data association, resulting in

$$X = F^{-1}_{m,p_1(U,\omega)}(U_1),\ Y = F^{-1}_{n,p_2(U,\omega)}(U_2),\ \text{and } \delta = U.$$

Again due to the discreteness of X and Y, it is unlikely to find a function $H(X,Y)$ such that its distribution is free of ω , so the goal is to seek the b function as in the Poisson model. However, this model has a significant difference with the Poisson case: δ has a genuine prior $\delta \sim \pi$. Our proposal here is to make use of the joint log-density function of (X, Y, δ), denoted as $\ell(\delta, \omega; x, y) = \log f(x, y, \delta; \omega)$, to derive the maximum a posteriori estimator $\hat{\delta}$ as an approximation to δ. Let $(\hat{\delta}, \hat{\omega}) = \arg\max_{\delta,\omega} \ell(\delta, \omega; x, y)$ with $\hat{\delta} = \hat{\delta}(x, y)$ and $\hat{\omega} = \hat{\omega}(x, y)$, and then b is obtained as

$$b(x, y, \delta) = \ell(\hat{\delta}(x, y), \hat{\omega}(x, y); x, y) - \ell(\delta, \hat{\omega}_\delta(x, y); x, y), \tag{19}$$

where $\hat{\omega}_\delta(x, y) = \arg\max_\omega \ell(\delta, \omega; x, y)$.

As a consequence, the final association is $b(X, Y, \delta) = W_b(\omega)$, where $W_b(\omega)$ is obtained by replacing (x, y, δ) with $\left(F^{-1}_{m,p_1(U,\omega)}(U_1), F^{-1}_{n,p_2(U,\omega)}(U_2), U\right)$ in (19). Let G_ω denote the c.d.f. of $W_b(\omega)$, and define $\underline{G}(s) = \inf_{\omega \in (-1,1)} G_\omega(s)$, and then the plausibility function for δ is $\mathsf{pl}_{x,y}(\delta) = 1 - \underline{G}(b(x, y, \delta))$, with the interval estimator $C_\alpha(X, Y)$ defined by $C_\alpha(x, y) = \{\tilde{\delta} : \mathsf{pl}_{x,y}(\tilde{\delta}) \geq \alpha\}$. The computational details are given in Appendix 7.

For a simulation study, the prior of $\delta \equiv p_1 - p_2$ is chosen to have the same distribution as $2\beta - 1$ with $\beta \sim \mathsf{Beta}(a, b)$ for some known value of (a, b). This choice of prior guarantees that the support of $\pi(\delta)$ is $[-1, 1]$. For each simulated δ, the value of $\tau \equiv p_1 + p_2$ is created as $\tau = 1 + (1 - |\delta|)\omega$ with $\omega \sim \mathsf{Unif}(-1, 1)$. Then the corresponding true values of p_1 and p_2 used to simulate the data can be determined accordingly. Two settings of prior distribution parameters, $(a, b) = (2, 2)$ and $(2, 5)$, and a sequence of binomial sizes, $m = n = 20, 30, \ldots, 100$, are considered. Since the typical empirical Bayes methods do not apply to this problem, in Figure 4 we give the results of partial Bayes and confidence distribution solutions.

Similar to the empirical Bayes solutions in the previous two models, the confidence distribution approach does not possess the desired coverage, while partial Bayes provides exact inference results. This is because the confidence distribution method for this model relies on large sample theory, and may not work well for small samples. The interval width of the partial Bayes solution is slightly wider than that of the confidence distribution method, but the difference is only tiny; as expected, the width will decrease as sample size increases, which again indicates the efficiency.

6. Application

In this section we apply the partial Bayes model to a dataset of National Basketball Association (NBA) games. In basketball competitions, a three-point shot, if made, rewards the highest score in one single attempt. Therefore, as the game comes to an end, three-point shots are more valuable for a team that has very limited offensive possessions and needs to overcome the deficit in score. When the game is decided by the last possession, a three-point shot is usually beneficial or even necessary for such teams, and the choice of player that will make the attempt is crucial to the outcome of the game.

Typically, the player to be chosen should have the highest success rate of three-point shots, and historical data can be used to evaluate each player's

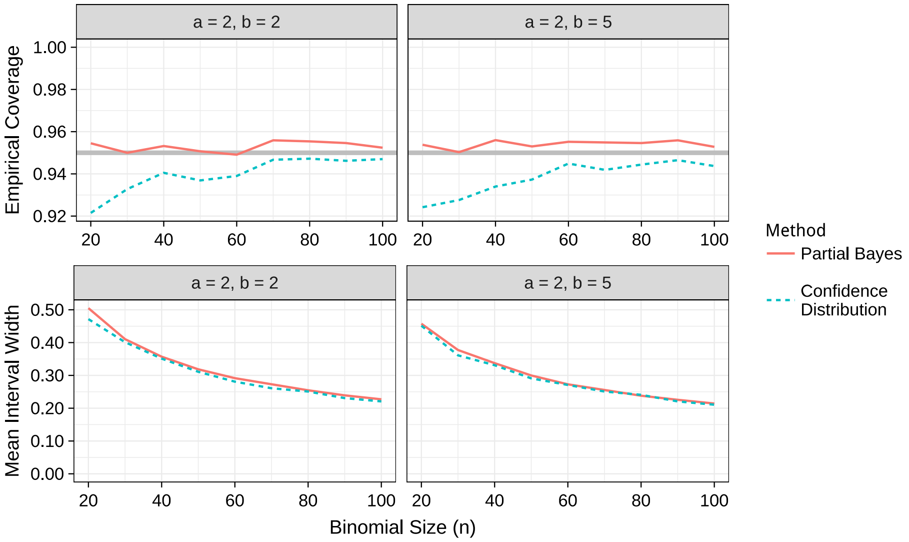

FIG 4. *The empirical coverage percentage (the top two panels) and mean interval width (the bottom two panels) for δ with an increasing binomial size n and two parameter settings, among 10,000 simulation runs. Partial Bayes and confidence distribution solutions are compared, showing that the partial Bayes solution guarantees the nominal coverage for all n.*

performance. If X_i is the number of three-point shots made in n_i attempts by player i, then usually X_i can be modeled by a binomial distribution $\mathsf{Bin}(n_i, p_i)$ or a Poisson distribution $\mathsf{Pois}(n_i p_i)$, where p_i stands for the success rate. In this application we choose the latter one for simplicity. Given this model, a classical point estimator for p_i is $\hat{p}_i = X_i/n_i$, and a $100(1-\alpha)\%$ frequentist confidence interval for p_i is $\left(G_{X_i}\left(\frac{\alpha}{2}\right)/n_i, G_{X_i+1}\left(1-\frac{\alpha}{2}\right)/n_i\right)$, where $G_s(\cdot)$ is the c.d.f. of the $\mathsf{Gamma}(s)$ distribution.

If additional information is available, for example p_i's are assumed to follow a common prior distribution $\pi(p)$, then the efficiency of the inference can be improved by incorporating this prior. This assumption is sensible since the players are in the same team or league, and they are expected to share some common characteristics. By combining the two sources of information — player's own historical statistics, and those of other players in the team or league — a more fair evaluation of players' performance could then be obtained. In what follows, we analyze the three-point shot data obtained from the official NBA website. We first select three players from each team that have the highest three-point goal success rates during the 2015-2016 regular season, and then retrieve the data from each player's last ten games within that season. The number of three-point shots made (X_i) and attempted (n_i) for each player are computed from this dataset.

To take the prior information into account, we first use the empirical Bayes method to analyze this dataset similar to the analysis in Efron and Morris (1975) for baseball games, but with a Poisson model instead of a normal one. The p_i's are assumed to follow a common exponential prior $exp(\theta)$, where $\theta > 0$ stands for the mean. In reality p_i lies in the interval $(0, 1)$, and here the conjugate exponential prior is used mainly for simplicity of computation. It does not harm the analysis in practice since when θ is small the majority of probability mass is on $(0, 1)$, and we manually truncate any interval estimate that is beyond one. Using the marginal distribution of X_i, the MLE of θ is obtained as $\hat{\theta} = 0.410$. As a result, the point estimator for p_i is taken to be the posterior mean $(X_i+1)/(\hat{\theta}^{-1}+n_i)$, and the approximate $100(1-\alpha)\%$ Bayesian credible interval is $\left(G_{X_i+1}\left(\frac{\alpha}{2}\right)/(\hat{\theta}^{-1}+n_i), G_{X_i+1}\left(1-\frac{\alpha}{2}\right)/(\hat{\theta}^{-1}+n_i)\right)$.

Finally, the partial Bayes model in Section 5.2 is used to derive an interval estimator for p_i, and the point estimator is chosen as the value of p_i that maximizes $\mathsf{pl}_x(p_i)$. The comparison of the three methods mentioned above is shown in Figure 5 for five representative players.

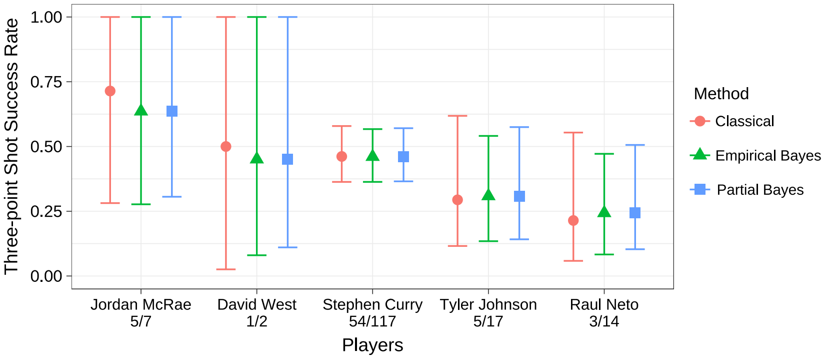

FIG 5. *Comparing three methods for analyzing three-point shot success rates on five representative players among the ninety players studied. Numbers of three-point shots made and attempted are displayed under players' names. The error bars and the dots stand for the 90% interval estimates and the point estimates respectively. The three different shapes of dots represent the three inference methods.*

Among these five players, Jordan McRae and David West are examples of players with high success rates but few number of shot attempts. It is clear that both empirical Bayes and partial Bayes results shrink the classical point estimates towards the grand mean, as an effect of combining individual and league information. To the opposite, for players below the average, such as Tyler Johnson and Raul Neto, their success rates are lifted by a small percentage. Stephen Curry, as a third case, is almost unaffected by the shrinkage. This is because he made a large number of shot attempts, so that his personal performance dominates the overall estimate. It is worth noting that David West has a higher point estimate of success rate than Stephen Curry in the classical method, but their rankings are reversed in empirical Bayes and partial Bayes methods.

The comparison of the three methods also highlights the advantage of the partial Bayes method. It is known that the classical confidence interval is exact, but is wider than that of the other two methods. The empirical Bayes solution is more efficient, but theoretically it is only approximate. The partial Bayes solution, in contrast, combines the advantages of the other two methods, providing both exact and efficient inference results. This example hence suggests that the partial Bayes model framework is useful for real-life data analysis tasks.

7. Conclusion and discussion

This article considers the statistical inference for partial Bayes problems, *i.e.*, Bayesian models without fully-specified prior distributions. We have developed a general model framework for studying such problems, and have provided theoretical justification for both the exactness and the efficiency of the inference results. Compared with other existing methodologies dealing with partial prior information, such as empirical Bayes and confidence distribution, our proposed method has shown superior performance.

Indeed, statisticians and scientists do care about exact inference for such useful models. For example, pioneering work in the empirical Bayes literature, such as Morris (1983); Laird and Louis (1987); Carlin and Gelfand (1990), has revealed the fact that empirical Bayes estimators could underestimate the uncertainty, and these authors all emphasized the importance of providing exact inference for such problems. To some extent our discussion sheds new light on this issue and shows promising results. From this perspective, partial Bayes models are powerful extensions to conventional Bayesian models, as they allow for more flexibility on the prior specifications, and meanwhile avoid sacrificing the exactness of inference. As a result, they can be used to combine different types of information for which other existing methods are difficult.

Of course, “There is no such thing as a free lunch.” The exact and efficient inference for partial Bayes problems is very useful yet challenging. As has been illustrated by the three example models, the construction of the interval estimators sometimes needs to be studied case by case. Also, similar to the hierarchical Bayesian models, partial Bayes solutions usually involve a moderate amount of computations such as sampling and optimization that need to be taken care of. Despite all these obstacles, we believe that the partial Bayes model framework is useful in real data analysis, and we expect that more research along this direction can be fruitful, as far as exact and efficient probabilistic inference concerns.

Appendix

Proof of Theorem 1

Let $Q_h(w) = P_{\mathcal{S}_h}(w \notin \mathcal{S}_h)$, and then for any $(x, w, \tilde{\eta})$ such that $b(T(x), \tilde{\eta}) = w$,

$$\begin{aligned}\mathsf{cpl}_{T(x)|h}(\tilde{\eta}) &= 1 - P_{\mathcal{S}_h}(\Theta_{T(x)}(\mathcal{S}_h) \subseteq (-\infty, \tilde{\eta}) \cup (\tilde{\eta}, +\infty))\\ &= 1 - P_{\mathcal{S}_h}(\tilde{\eta} \notin \Theta_{T(x)}(\mathcal{S}_h))\\ &= 1 - P_{\mathcal{S}_h}(w \notin \mathcal{S}_h) \equiv 1 - Q_h(w). \qquad (20)\end{aligned}$$

Therefore,

$$\mathsf{cpl}_{T(X)|h}(\eta) \geq \alpha \Leftrightarrow Q_h(W_b(\theta^*)) \leq 1-\alpha. \tag{21}$$

First fix θ^*, and let $\mathsf{P}_{T(X),\eta|H(X)=h}$ denote the probability measure of $(T(X),\eta)$ given $H(X)=h$, and then we see that $\mathsf{P}_{T(X),\eta|H(X)=h} \equiv \mathsf{P}_{W_T,V_\eta|h}$. As a result, we apply the probability measure $\mathsf{P}_{W_T,V_\eta|h}$ on both sides of (21), obtaining

$$P_{T(X),\eta|H(X)=h}\left(\mathsf{cpl}_{T(X)|h}(\eta) \geq \alpha\right) = P_{W_b(\theta^*)|h}\left(Q_h(W_b(\theta^*)) \leq 1-\alpha\right).$$

The validity of $\mathcal{S}_h$ implies $P_{W_b(\theta^*)|h}\left(Q_h(W_b(\theta^*)) \geq 1-\alpha\right) \leq \alpha$ for any θ^*. Therefore,

$$P_{X,\eta|H(X)}(C_\alpha(X) \ni \eta|H(X)=h) = P_{T(X),\eta|H(X)=h}\left(\mathsf{cpl}_{T(X)|h}(\eta) \geq \alpha\right) \geq 1-\alpha. \tag{22}$$

Note that (22) is true for any fixed θ^*, so it also holds with $\theta^* \sim \pi^*(\theta^*)$, for any $\pi^*(\theta^*)$.

Proof of Theorem 2

Similar to (20), we have $\mathsf{cpl}_{T(x)|h,t}(\tilde{\eta}) \geq \alpha \Leftrightarrow Q_{h,t}(w) \leq 1-\alpha$, where $(x,w,\tilde{\eta})$ satisfies $b(T(x),\tilde{\eta}) = w$, and $Q_{h,t}(w) = P_{\mathcal{S}_{h,t}}(w \notin \mathcal{S}_{h,t})$. Fixing $t \equiv T(x)$, $\eta \mapsto b(t,\eta)$ is one-to-one by definition, so the mapping must be monotone. Without loss of generality we assume $b(t,\eta)$ is increasing in η, since otherwise we can use $-b$ in place of b.

Let $Z_\eta = a_\eta(U,V_\eta)$, and then it can be shown that

$$\begin{aligned} F_{W_b|h,t}(w) &= P_{W_b|h,t}\left(W_b \leq w | W_H = h, Z_T = t\right) \\ &= P_{Z_T,Z_\eta|h,t}\left(b(Z_T,Z_\eta) \leq b(t,\tilde{\eta}) | W_H = h, Z_T = t\right) \\ &= P_{Z_\eta|h,t}\left(b(t,Z_\eta) \leq b(t,\tilde{\eta}) | W_H = h, Z_T = t\right) \\ &= P_{U,V_\eta|h,t}\left(a_\eta(U,V_\eta) \leq \tilde{\eta} | W_H = h, Z_T = t\right). \end{aligned}$$

By the definition of the decomposition in (3), $W_H = h, a_T(\theta^*,W_T) = t \Leftrightarrow a(\theta^*,W) = x$, and hence $W_H = h, Z_T = t \Leftrightarrow a(U,W) = x$. Also it is clear from the association equations that $(\theta^*,\eta,X) \equiv (U, a_\eta(U,V_\eta), a(U,W))$, so we have $F_{W_b|h,t}(w) = P_{\eta|X=x}(\eta \leq \tilde{\eta}|X=x) = F_{\eta|x}(\tilde{\eta})$.

Finally, let $u = F_{W_b|h,t}(w)$. Since

$$\mathcal{S}_{h,t} = \left\{F^{-1}_{W_{b|h,t}}(u'), u' \in (0,1) : |u'-0.5| < |U_{\mathcal{S}}-0.5|\right\}, \ U_{\mathcal{S}} \sim \mathsf{Unif}(0,1),$$

we have

$$\begin{aligned} Q_{h,t}(w) &= P_{\mathcal{S}_{h,t}}(w \notin \mathcal{S}_{h,t}) = P_{U_{\mathcal{S}}}\left(|u-0.5| \geq |U_{\mathcal{S}}-0.5|\right) = |1-2u| \\ &= |1-2F_{\eta|x}(\tilde{\eta})|, \end{aligned}$$

and hence $\mathsf{cpl}_{T(x)|h,t}(\tilde{\eta}) \geq \alpha \Leftrightarrow Q_{h,t}(w) \leq 1-\alpha \Leftrightarrow \alpha/2 \leq F_{\eta|x}(\tilde{\eta}) \leq 1-\alpha/2$.

Proof of Theorem 3

We first show that $Z_{T_n} \xrightarrow{P} U$ and $W_{b_n} \xrightarrow{P} V_\eta$ under conditions *(a)* and *(b)*. Let P_U be the probability measure of U. Since U and W_{T_n} are independent, we have that for any $\varepsilon > 0$, $P(|Z_{T_n} - U| > \varepsilon) = \int f_n \mathrm{d}\mathsf{P}_U$ where $f_n(u) = P_{W_{T_n}}(|a_T(u, W_{T_n}) - u| > \varepsilon)$. Condition *(a)* indicates that $f_n \to 0$, and then by $|f_n| \le 1$ and the dominated convergence theorem, we have $\int f_n \mathrm{d}\mathsf{P}_U \to 0$, which implies that $Z_{T_n} \xrightarrow{P} U$. Moreover, $Z_{T_n} \xrightarrow{P} U$ implies $(Z_{T_n}, Z_\eta) \xrightarrow{P} (U, Z_\eta)$, where $Z_\eta = a_\eta(U, V_\eta)$. Then by the continuous mapping theorem and condition *(b)* we obtain $W_{b_n} \xrightarrow{P} b(U, a_\eta(U, V_\eta)) = V_\eta$ and $W_{b_n}(\theta^*) \xrightarrow{P} V_\eta$.

Next we prove that $\mathbb{E}(f(W_{b_n})|Z_{T_n}) \xrightarrow{P} \mathbb{E}(f(V_\eta))$ for any bounded continuous function f, where the notation $\mathbb{E}(X|Y)$ stands for the conditional expectation of X given Y. The main tool to prove this result is Theorem 2.1 of Goggin (1994). Let Q_n be a probability measure under which W_{b_n} and Z_{T_n} are independent, *i.e.*, $Q_n((-\infty, w] \times (-\infty, z]) = F_{W_{b_n}}(w) F_{Z_{T_n}}(z)$, where $F_{W_{b_n}}$ and $F_{Z_{T_n}}$ are the corresponding marginal c.d.f.'s. Then for any $\varepsilon > 0$, under the Q_n measure, $P_{Q_n}(|l_n(W_{b_n}, Z_{T_n}) - 1| > \varepsilon) = \int I_{A_n} \mathrm{d}Q_n$, where I_{A_n} is the indicator function of the set $A_n = \{(w, z) : |l_n(w, z) - 1| > \varepsilon\}$. Condition *(c)* implies that $I_{A_n} \to 0$ pointwisely, so by the dominated convergence theorem we have $\int I_{A_n} \mathrm{d}Q_n \to 0$. As a result, under the Q_n measure, $l_n(W_{b_n}, Z_{T_n}) \xrightarrow{P} 1$ and hence $(W_{b_n}, Z_{T_n}, l_n(W_{b_n}, Z_{T_n})) \xrightarrow{d} (V_\eta, U, 1)$. Then Theorem 2.1 of Goggin (1994) claims that $\mathbb{E}(f(W_{b_n})|Z_{T_n}) \xrightarrow{d} \mathbb{E}(f(V_\eta)|U)$ for any bounded continuous function f . Since U and V_η are independent, we have $\mathbb{E}(f(V_\eta)|U) = \mathbb{E}(f(V_\eta))$ and hence $\mathbb{E}(f(W_{b_n})|Z_{T_n}) \xrightarrow{P} \mathbb{E}(f(V_\eta))$.

Finally, Theorem 2.1 of Xiong and Li (2008) shows that $\mathbb{E}(f(W_{b_n})|Z_{T_n}) \xrightarrow{P} \mathbb{E}(f(V_\eta))$ is equivalent to $W_{b_n}|Z_{T_n} \xrightarrow{d.P} V_\eta$, which concludes the proof.

Proof of (12)*,* (13)*, and* (14)

Let $\mathbf{0}_k$ denote the $k \times 1$ zero vector, I_k be the $k \times k$ identity matrix, and J_k be a $k \times k$ matrix with all elements being one. It is easy to show that $(W_b, W_H')' = A(e', \varepsilon')'$, where $A = \begin{pmatrix} \frac{1}{n} & \frac{1}{n}\mathbf{1}_{n-1}' & (\frac{1}{n} - 1)\tau & \frac{\tau}{n}\mathbf{1}_{n-1}' \\ -\mathbf{1}_{n-1} & I_{n-1} & -\tau\mathbf{1}_{n-1} & \tau I_{n-1} \end{pmatrix}$, $e = (e_1, \ldots, e_n)'$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. Since $(e', \varepsilon')' \sim \mathsf{N}(\mathbf{0}_{2n}, I_{2n})$, we have $(W_b, W_H')' \sim \mathsf{N}\left(\mathbf{0}_{2n}, \left(\begin{smallmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{smallmatrix}\right)\right)$, where $\Sigma_{11} = \{1 + (n-1)\tau^2\}/n$, $\Sigma_{12} = \tau^2 \mathbf{1}_{n-1}'$, and $\Sigma_{22} = (\tau^2 + 1)(J_{n-1} + I_{n-1})$.

Simple calculation shows that $\Sigma_{22}^{-1} = (\tau^2 + 1)^{-1}(I_{n-1} - n^{-1}J_{n-1})$, and then according to the property of multivariate normal distribution, we have $W_b | W_H = h \sim \mathsf{N}(\tilde{\mu}, \tilde{\sigma}^2)$, where $\tilde{\mu} = \Sigma_{12}\Sigma_{22}^{-1}h = \tau^2(\tau^2 + 1)^{-1}(\bar{x} - x_1)$, and $\tilde{\sigma}^2 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = n^{-1}(1 + \tau^2)^{-1}(n\tau^2 + 1)$.

Let $F_{W_b|h}$ denote the c.d.f. of $\mathsf{N}(\tilde{\mu}, \tilde{\sigma}^2)$, then

$$\begin{aligned} \mathcal{S}_h &= \left\{ F_{W_b|h}^{-1}(u'), u' \in (0,1) : |u' - 0.5| < |U_{\mathcal{S}} - 0.5| \right\}, \ U_{\mathcal{S}} \sim \mathsf{Unif}(0,1) \\ &= \{ z : |(z - \tilde{\mu})/\tilde{\sigma}| < |Z_{\mathcal{S}}| \}, \ Z_{\mathcal{S}} \sim \mathsf{N}(0,1). \end{aligned}$$

Therefore, define $Q_h(s) = P_{\mathcal{S}_h}(s \notin \mathcal{S}_h)$, and we get $Q_h(s) = 2\Phi(|(s-\tilde{\mu})/\tilde{\sigma}|)-1$. From (20) we have $\mathsf{cpl}_{T(x)|h}(\mu_1) = 1 - Q_h(w)$ where $w = b(T(x), \mu_1) = \bar{x} - \mu_1$. As a result, $\mathsf{cpl}_{T(x)|h}(\mu_1) = 2 - 2\Phi(|(\bar{x}-\mu_1-\tilde{\mu})/\tilde{\sigma}|) = 2\Phi\left(-|(\bar{x}-\mu_1-\tilde{\mu})/\tilde{\sigma}|\right)$, which reduces to (13). The interval estimator then follows directly.

Proof of (16) *and* (17)

Let $U = (\tau\varepsilon_1 + e_1 - \tilde{\tau}Z)/(\sqrt{n}\tilde{\tau})$, and then it is easy to verify that $(U, e_1)' \sim \mathsf{N}(\mathbf{0}, \Sigma)$, where $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ and $\rho = (\sqrt{n}\tilde{\tau})^{-1}$. Since U, e_1, and M_{n-2}^2 are independent, the joint density function of (U, e_1, M_{n-2}^2) can be written as

$$g_0(u,z,x) \propto \exp\left\{-\frac{1}{2}(u,z)\Sigma^{-1}(u,z)'\right\} x^{\frac{n}{2}-2} \exp\left\{-\frac{n-2}{2}x\right\}.$$

Let $W_e = e_1/\sqrt{M_{n-2}^2}$. Note also that $W_H = U/\sqrt{M_{n-2}^2}$, so with the transformation of variables $s = x/\sqrt{z}, h = y/\sqrt{z}, t = z$, the joint density of (W_e, W_H, M_{n-2}^2) is

$$\begin{aligned} g(s,h,t) &\propto \exp\left\{-\frac{1}{2}(s\sqrt{t}, h\sqrt{t})\Sigma^{-1}(s\sqrt{t}, h\sqrt{t})'\right\} t^{\frac{n}{2}-2} \exp\left\{-\frac{n-2}{2}t\right\} \cdot t \\ &= \exp\left\{-\frac{1}{2}t(s,h)\Sigma^{-1}(s,h)'\right\} t^{\frac{n}{2}-1} \exp\left\{-\frac{n-2}{2}t\right\}. \end{aligned}$$

For simplicity of notations let $\Sigma^{-1} = \left(\begin{smallmatrix} A & B \\ B & A \end{smallmatrix}\right)$, where $A = n(\tau^2+1)(n\tau^2+1)^{-1}, B = -(n\tau^2+1)^{-1}\sqrt{n(n-1)(\tau^2+1)}$, and then the joint density of (W_e, M_{n-2}^2) given $W_H = h$ is

$$\begin{aligned} g(s,t|h) &\propto \exp\left\{-\frac{1}{2}t(As^2 + 2Bsh + Ah^2)\right\} t^{\frac{n}{2}-1} \exp\left\{-\frac{n-2}{2}t\right\} \\ &= \exp\left\{-\frac{A}{2}t\left(s + \frac{B}{A}h\right)^2\right\} \cdot t^{\frac{n}{2}-1} \cdot \exp\left\{-\frac{1}{2}(h^2+n-2)t\right\}. \end{aligned} \tag{23}$$

Integrating s out gives $g(t|h) \propto t^{(n-1)/2-1} \exp\left\{-(h^2+n-2)t/2\right\}$, which corresponds to the $2(h^2+n-2)^{-1}\mathsf{Gamma}((n-1)/2)$ distribution. (23) also shows that given $W_H = h$ and $M_{n-2}^2 = t$, the density function of W_e is $g(s|h,t) \propto \exp\left\{-At(s+hB/A)^2/2\right\}$, implying the $\mathsf{N}\left(-hB/A, (At)^{-1}\right)$ distribution.

As a consequence, given $W_H = h$, the random variables M_{n-2}^2 and W_e can be expressed as $M_{n-2}^2 = C\tilde{M}^2$ and $W_e = -hB/A + (AM_{n-2}^2)^{-1/2}\tilde{Z}$, where $C = (n-1)(h^2+n-2)^{-1}$, $\tilde{M}^2 \sim \chi_{n-1}^2/(n-1)$, $\tilde{Z} \sim \mathsf{N}(0,1)$, and $\tilde{M}^2$ and $\tilde{Z}$ are independent. Therefore, $e_1 = W_e\sqrt{M_{n-2}^2} = h\sqrt{\omega(n-1)/n} \cdot \sqrt{C}\tilde{M} + \sqrt{1-\omega(n-1)/n} \cdot \tilde{Z}$.

Now consider the distribution of $\overline{X}_{(-1)} - \mu_1$. It is easy to see that $\overline{X}_{(-1)} - \mu_1 = e_1 - \sqrt{n}\tilde{\tau}W_H M_{n-2}$, so given $W_H = h$,

$$\begin{aligned}\mathbb{E}(\overline{X}_{(-1)} - \mu_1 | W_H = h) &= \mathbb{E}(e_1 | W_H = h) - \sqrt{n}\tilde{\tau} h \mathbb{E}(M_{n-2} | W_H = h) \\ &= h\sqrt{C}\left(\sqrt{\omega(n-1)/n} - 1/\sqrt{\omega(n-1)/n}\right)\mathbb{E}(\tilde{M}).\end{aligned}$$

Also $\mathbb{E}(S_{(-1)} | W_H = h) = \sqrt{C/\omega} \cdot \mathbb{E}(\tilde{M})$, $\mathbb{E}(S_{(-1)}^{-1} | W_H = h) = \sqrt{\omega/C} \cdot \mathbb{E}(\tilde{M}^{-1}) = (n-1)(n-2)^{-1}\sqrt{\omega/C} \cdot \mathbb{E}(\tilde{M})$, so with the $\tilde{\mu}$ given in (16), we can show that $\mathbb{E}(\overline{X}_{(-1)} - \mu_1 - \tilde{\mu} | W_H = h) = 0$. Similarly, it can be calculated that

$$\mathrm{Var}(\overline{X}_{(-1)} - \mu_1 - \tilde{\mu} | W_H = h) = 1 - \frac{(n-1)(n-2)(n-3-h^2)}{n(n-3)(n-2+h^2)}\omega,$$

and an unbiased and consistent estimator for ω is $\hat{\omega} = (n-3)(h^2+n-2)^{-1}S_{(-1)}^{-2}$. Therefore, with the $\tilde{\sigma}$ in (16), $W_b(\tau) | W_H = h \xrightarrow{d} \mathsf{N}(0,1)$ for any $\tau > 0$. The $n^{-\gamma}$ term is used to guarantee that the variance is always positive.

Finally, the auxiliary variable to predict is

$$W_b(\tau) = \frac{c_2\sqrt{\omega}\left(\tilde{M} - c_3\tilde{M}^{-1}\right) + \sqrt{1-\omega(n-1)/n}\tilde{Z}}{\sqrt{\max\left\{n^{-\gamma}, 1 - c_1\omega\tilde{M}^{-2}\right\}}},$$

and (17) follows immediately.

Computation for the Poisson hierarchical model

We first obtain the expression for $\ell(\lambda_1; x)$. Given λ_1, $X_1 \sim \mathsf{Pois}(\lambda_1 t_1)$, $X_i = F^{-1}_{\lambda_1 t_i V_i / V_1}(U_i)$, and X_1 and $X_{(-1)} = (X_2, \ldots, X_n)'$ are independent. Marginally X_i follows a negative binomial distribution $\mathsf{NB}(s,p)$ with probability mass function $p(x) \propto p^s(1-p)^x$, where $p = 1/(1+\gamma)$. Therefore, the joint density of $X_{(-1)}$ and V_1 is

$$\begin{aligned}p(x_2, \ldots, x_n, v_1 | \lambda_1) &= \prod_{i=2}^{n}\left\{\frac{\Gamma(x_i+s)}{\Gamma(s)}p_i^s(1-p_i)^{x_i}\right\} \cdot \frac{1}{\Gamma(s)}v_1^{s-1}e^{-v_1}, \\ p_i &= \frac{1}{1+\lambda_1 t_i / v_1},\end{aligned}$$

and hence the density of $X_{(-1)}$ is

$$p(x_2, \ldots, x_n | \lambda_1) = \int_0^{+\infty}\prod_{i=2}^{n}\left\{\frac{\Gamma(x_i+s)}{\Gamma(s)}p_i^s(1-p_i)^{x_i}\right\} \cdot \frac{1}{\Gamma(s)}v_1^{s-1}e^{-v_1}\mathrm{d}v_1.$$

As a result, $\ell(\lambda_1; x) = x_1\log(\lambda_1) - \lambda_1 t_1 + \log p(x_2, \ldots, x_n | \lambda_1) + C$, where C is some constant unrelated to λ_1, and the MLE for λ_1 can be obtained using standard optimization methods.

To obtain G_{λ_1}, the c.d.f. of $W_b(\lambda_1)$, we first use Monte Carlo method to simulate U and V to get a random sample of $W_b(\lambda_1)$, and then G_{λ_1} is approximated by $\hat{G}_{\lambda_1}$, the empirical c.d.f. of $W_b(\lambda_1)$. Finally, the interval estimator is computed using a grid search on $\mathsf{pl}_x(\lambda_1)$.

Computation for the binomial rate-difference nodel

It is easy to show that $\ell(\delta,\omega;x,y) = x\log p_1 + (m-x)\log(1-p_1) + y\log p_2 + (n-y)\log(1-p_2) + \log\pi(\delta)$, where $p_1 = \{1+\delta+(1-|\delta|)\omega\}/2$ and $p_2 = \{1-\delta+(1-|\delta|)\omega\}/2$.

Keeping δ fixed, $\hat{\omega}_\delta = \arg\max_\omega \ell(\delta,\omega;x,y)$ can be obtained by solving the equation

$$\frac{\partial\ell}{\partial\omega} = \left(\frac{x}{p_1} - \frac{m-x}{1-p_1} + \frac{y}{p_2} - \frac{n-y}{1-p_2}\right)\cdot\frac{1}{2}(1-|\delta|) = 0. \tag{24}$$

Since $p_1 = p_2+\delta$, (24) reduces to a cubic equation $ap_2^3+bp_2^2+cp_2+d=0$, where $a=m+n$, $b=-(x+y)-m(1-\delta)-n(1-2\delta)$, $c=x-m\delta+y(1-2\delta)-n(\delta-\delta^2)$, and $d=y(\delta-\delta^2)$. The solution should be sought within the range $\max(0,-\delta) < p_2 < \min(1,1-\delta)$. As a result, $(\hat{\delta},\hat{\omega}) = \arg\max_{\delta,\omega}\ell(\delta,\omega;x,y)$ is obtained by computing $\hat{\omega}_\delta$ over a grid of δ values.

The remaining part of the computation proceeds similarly to the Poisson model, by simulating (U_1,U_2,U) and computing the distribution of $W_b(\omega)$, and hence the details are omitted.

Acknowledgments

The authors would like to thank the editor, an associate editor, and two reviewers for their valuable and insightful comments, which have greatly helped improve this article. The authors also thank Ryan Martin for early discussions on this topic. Chuanhai Liu's work was partially supported by the National Science Foundation grant DMS-1316922.

References

Carlin, B. P. and Gelfand, A. E. (1990). Approaches for Empirical Bayes Confidence Intervals. *Journal of the American Statistical Association* **85** 105-114. MR1137356

Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *The American Statistician* **39** 83-87. MR0789118

Deely, J. J. and Lindley, D. V. (1981). Bayes empirical bayes. *Journal of the American Statistical Association* **76** 833–841. MR0650894

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press. MR2724758

Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators – Part I: The Bayes Case. *Journal of the American Statistical Association* **66** 807-815. MR0323014

Efron, B. and Morris, C. (1972a). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347. MR0334386

Efron, B. and Morris, C. (1972b). Limiting the Risk of Bayes and Empirical Bayes Estimators – Part II: The Empirical Bayes Case. *Journal of the American Statistical Association* **67** 130-139. MR0323015

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its Competitors – An Empirical Bayes Approach. *Journal of the American Statistical Association* **68** 117-130. MR0388597

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association* **70** 311-319.

Goggin, E. M. (1994). Convergence in Distribution of Conditional Expectations. *Ann. Probab.* **22** 1097–1114. MR1288145

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes Confidence Intervals Based on Bootstrap Samples. *Journal of the American Statistical Association* **82** 739-750. MR0909979

Lambert, D. and Duncan, G. T. (1986). Single-parameter inference based on partial prior information. *Canadian Journal of Statistics* **14** 297–305. MR0876755

Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological)* **34** 1-41. MR0415861

Martin, R. (2015). Plausibility Functions and Exact Frequentist Inference. *Journal of the American Statistical Association* **110** 1552-1561. MR3449054

Martin, R. (2017). A mathematical characterization of confidence as valid belief. *arXiv preprint arXiv:1707.00486*.

Martin, R. and Liu, C. (2013). Inferential Models: A Framework for Prior-Free Posterior Probabilistic Inference. *Journal of the American Statistical Association* **108** 301-313. MR3174621

Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 195–217. MR3299405

Martin, R. and Liu, C. (2015b). Marginal Inferential Models: Prior-Free Probabilistic Inference on Interest Parameters. *Journal of the American Statistical Association* **110** 1621-1631. MR3449059

Martin, R. and Liu, C. (2015c). *Inferential Models: Reasoning with Uncertainty*. Chapman & Hall/CRC. MR3618727

Meaux, L., Seaman Jr, J. and Young, D. (2002). Statistical inference with partial prior information based on a Gauss-type inequality. *Mathematical and computer modelling* **35** 1483–1488. MR1916028

Moreno, E., Bertolino, F. and Racugno, W. (2003). Bayesian inference under partial prior information. *Scandinavian Journal of Statistics* **30** 565–580. MR2002228

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* **78** 47-55. MR0696849

Robbins, H. (1956). An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*

157–163. MR0084919

Sweeting, T. J. (1989). On conditional weak convergence. *Journal of Theoretical Probability* **2** 461–474. MR1011199

Xie, M., Singh, K. and Strawderman, W. E. (2011). Confidence Distributions and a Unifying Framework for Meta-Analysis. *Journal of the American Statistical Association* **106** 320-333. MR2816724

Xie, M., Liu, R. Y., Damaraju, C. V. and Olson, W. H. (2013). Incorporating external information in analyses of clinical trials with binary outcomes. *The Annals of Applied Statistics* **7** 342-368. MR3086422

Xiong, S. and Li, G. (2008). Some results on the convergence of conditional distributions. *Statistics & Probability Letters* **78** 3249 - 3253. MR2479485