*Statistical Science*
2011, Vol. 26, No. 2, 271–287
DOI: 10.1214/10-STS349


# Estimating Random Effects via Adjustment for Density Maximization[1]

**Carl Morris and Ruoxi Tang**

*Abstract.* We develop and evaluate point and interval estimates for the random effects $\theta_i$, having made observations $y_i|\theta_i \overset{ind}{\sim} N[\theta_i, V_i], i = 1, \ldots, k$ that follow a two-level Normal hierarchical model. Fitting this model requires assessing the Level-2 variance $A \equiv \mathrm{Var}(\theta_i)$ to estimate shrinkages $B_i \equiv V_i/(V_i + A)$ toward a (possibly estimated) subspace, with $B_i$ as the target because the conditional means and variances of $\theta_i$ depend linearly on $B_i$, not on $A$. Adjustment for density maximization, ADM, can do the fitting for any smooth prior on $A$. Like the MLE, ADM bases inferences on two derivatives, but ADM can approximate with any Pearson family, with Beta distributions being appropriate because shrinkage factors satisfy $0 \le B_i \le 1$.

Our emphasis is on frequency properties, which leads to adopting a uniform prior on $A \ge 0$, which then puts Stein's harmonic prior (SHP) on the $k$ random effects. It is known for the "equal variances case" $V_1 = \cdots = V_k$ that formal Bayes procedures for this prior produce admissible minimax estimates of the random effects, and that the posterior variances are large enough to provide confidence intervals that meet their nominal coverages. Similar results are seen to hold for our approximating "ADM-SHP" procedure for equal variances and also for the unequal variances situations checked here.

For shrinkage coefficient estimation, the ADM-SHP procedure allows an alternative frequency interpretation. Writing $L(A)$ as the likelihood of $B_i$ with $i$ fixed, ADM-SHP estimates $B_i$ as $\hat{B}_i = V_i/(V_i + \hat{A})$ with $\hat{A} \equiv \mathrm{argmax}(A * L(A))$. This justifies the term "adjustment for likelihood maximization," ALM.

*Key words and phrases:* Shrinkage, ADM, Normal multilevel model, Stein estimation, objective Bayes.

## 1. INTRODUCTION

This concerns approximate frequentist, Bayesian, and objective Bayesian inferences for a widely applied two-level Normal hierarchical model. At Level-1, for $i = 1, \ldots, k$, unbiased estimates $y_i$ are observed with means $\theta_i$ and with known variance $V_i$. In practice the $\{V_i\}$ usually are unequal, perhaps with $V_i = \sigma^2/n_i$ and $\sigma^2$ known or accurately estimated. Thus

$$y_i|\theta_i \overset{ind}{\sim} N[\theta_i, V_i], \quad i = 1, \ldots, k. \tag{1}$$

In practice each Level-1 value $y_i$ here represents a sufficient statistic or a summary unbiased estimate based on the $n_i$ observations taken from the $i$th of the $k$ units (e.g., a hospital, a small area, or a teaching unit).

Level-2 specifies a Normal model for the random effects $\theta_i$, each with its own $r$-dimensional predictor variables $x_i$ so that for $\beta$ and an unknown variance $A \ge 0$,

$$\theta_i|\beta, A \overset{ind}{\sim} N[\mu_i = x_i'\beta, A], \quad i = 1, \ldots, k. \tag{2}$$

The case $r = 0$ corresponds to $\beta$ fully known and then it may be convenient to set $\beta = 0$ and $\mu_i = 0$, WLoG.

*Carl Morris is Professor, Department of Statistics, Harvard University, One Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: morris@stat.harvard.edu). Ruoxi Tang obtained his Ph.D. from Harvard's Statistics Department and is with a New York investment firm, Bloomberg L.P., 731 Lexington Ave., New York, New York 10022, USA.*

[1]Discussed in 10.1214/11-STS349A and 10.1214/11-STS349B; rejoinder at 10.1214/11-STS349REJ.

If $r \geq 1$, $X \equiv (x_1', x_2', \dots, x_k')'$ as a known $k \times r$ matrix, assumed to have full rank $r$.

The marginal distribution of $y = (y_1, \dots, y_k)'$, given $\beta$ and $A$, and the conditional distribution of $\theta_i$ follow from the above, so that

$$y_i | \beta, A \overset{ind}{\sim} N[x_i'\beta, V_i + A], \quad i = 1, \dots, k, \tag{3}$$

$$\theta_i | y_i, \beta, A \overset{ind}{\sim} N[(1 - B_i)y_i + B_i \mu_i, V_i(1 - B_i)], \quad i = 1, \dots, k, \tag{4}$$

where $\mu_i \equiv x_i'\beta$, and $B_i \equiv \frac{V_i}{V_i + A}$ is a "shrinkage factor."

When $r \geq 1$, the vector $\beta$ is assumed throughout to follow Lebesgue's flat prior on $[0, \infty)$, so

$$p(\beta, A)\, d\beta\, dA \propto d\beta \pi(A)\, dA. \tag{5}$$

Using this flat prior density for $\beta$ is equivalent to restricted maximum likelihood (REML). When $\pi(A)$ is proper, the posterior distribution for this prior is proper (it integrates finitely) if $k \geq r$. When $\pi(A)$ is improper, a larger $k$ is needed, with $k \geq r + 3$ sufficing for the main distributions $\pi(A)$ of interest here. When $r = 0$, as assumed initially, or when $r \geq 1$ and with $\beta$ integrated out, we can focus on the main issue of dealing with the (nuisance) variance component $A = \text{Var}(\theta_i)$ and how to make inferences about the shrinkages $B_i$.

Widely used programs like HLM, ML3 and SAS use MLE/REML methods to fit this model, while software for fully Bayesian inferences is available via BUGS and MLwiN (Rasbash et al., 2001). Maximum likelihood and REML obtain an estimate $\hat{A}$ that maximizes the likelihood function of $A$ (or marginal likelihood in the REML case). Asymptotically ($k$ large), maximum likelihood provides optimal estimates of $A$, leading to convergence of estimates via frequentist and Bayesian approaches. However, the standard errors assigned by MLE and REML methods to the random effect estimates and the corresponding interval estimates can lead to confidence intervals with much smaller than their nominal confidences, even asymptotically. This happens with MLE and REML methods not only because $A$ can be underestimated so that shrinkages are overestimated, but also because these procedures do not account for the fact that $A$ has been estimated.

Maximum likelihood and REML estimates of $A$ not infrequently produce $\hat{A} = 0$, in which case shrinkage MLEs are $\hat{B}_i = 1$. Examples occur in every field, as for the 8 schools data (Gelman et al., 2004), and in small area estimation (Bell, 1999). Then, per typical usage, the variance estimates may be taken to be $V_i(1 - \hat{B}_i) = 0$ when $r = 0$, leading to zero-width or overly narrow confidence intervals of $\theta_i$. As will be seen in Section 4, even when $\hat{A} > 0$ and this situation is avoided, overfitting via MLE and REML can be considerable and nominal 95% confidence intervals for $\theta_i$ might have true coverages in the 50–80% range.

The procedures developed here to fit the two-level model above offer computational ease comparable to maximum likelihood and REML methods, being based on differentiating the (adjusted) likelihood function twice. When $k$ is small or moderate, however, the adjustment provides much better standard errors and interval coverages. "Better" coverage is meant in the Level-2 frequentist sense of averaging over the data and the Level-2 model (2), for all fixed $\beta$, $A$, as illustrated in the equal variances case of Figure 6, Section 4.

Central to this development is the ADM procedure, "adjustment for density maximization" (Morris, 1988b), albeit not then with the ADM label. ADM can be used with any Pearson family (Normal, Gamma, Inverted Gamma, Beta, $F$, $t$ or skew-$t$) to approximate another distribution with a one-dimensional density. One merely multiplies the density by an adjustment which is determined by the Pearson family, and then makes the argmax function produce the mean, not the mode, of the Pearson distribution. As seen in (4), posterior means and variances of the random effects are linear functions of the shrinkage factors $B_i$, not of $A$, so it is desirable to estimate the posterior mean of $B_i$, and not the mode of $B_i$ or the mean of $A$. Shrinkage factor distributions are skewed and lie in [0,1], both of which make a Beta distribution approximate better than a Normal. Fitting Beta distributions via ADM is described in Section 2.3.

Estimating shrinkage factors via ADM will be seen to reduce to maximizing the posterior density of $A$ (or the marginalized density, if necessary), after having multiplied this density by $A$. This adjustment has several benefits, which include prevention of estimating $A$ as 0, and overestimating $A$ by just enough to account for the convex dependence of $B_i$ on $A$. ADM methods have been used successfully before to improve inferences of random effects in other multilevel models, as in Christiansen and Morris (1997) for a Poisson multilevel model.

The main procedure here approximates a formal posterior distribution stemming from the flat prior $\pi(A) = 1$ on $A \geq 0$ in (5). This flat prior on $A$, in conjunction with (2), induces Stein's harmonic prior (SHP) (30) on the random effects (Stein, 1981) and a minimax admissible estimator. (Stein's prior on $\boldsymbol{\theta}$ for $k \geq 3$, $d\boldsymbol{\theta}/\|\boldsymbol{\theta}\|^{(k-2)}$, is harmonic except at the origin, so it actually is "superharmonic." The shorter term "harmonic" is used here for simplicity of discourse.) The ADM approximations are seen in Section 3 to approximate closely the exact posterior means and variances of the random effects. Buttressed with the examples of Sections 3 and 4, our assessments show, by frequency standards, so for all fixed hyperparameters $A \ge 0$ and $\beta$, that the ADM-SHP combination outperforms commonly used MLE and REML procedures for estimating the random effects (1)–(5).

The ADM approximations of Section 2.7 apply to any smooth prior density $\pi(A)$, including the scale-invariant prior densities $\pi(A)$ on $A$

$$\pi(A)\,dA \propto A^{c-1}\,dA, \quad c > 0. \tag{6}$$

These receive some specific attention, but our frequency evaluations are limited to the special choice in (6) of $c = 1$ for which $A \sim \text{Unif}(0, \infty)$. Stein's harmonic prior not only produces safe frequency procedures for squared-error point estimation, but the posterior variances of $\theta_i$ are large enough to serve as a basis for confidence intervals centered at the posterior means (Stein, 1981; Morris 1983b, 1988a; Christiansen and Morris, 1997). Hierarchically, the uniform formal prior $\pi(A) = 1$ is suggested by the fact that the renowned James–Stein estimator is the posterior mean, exactly, if this flat prior is extended (inappropriately) to $A \sim \text{Unif}[-V, \infty)$ (Morris, 1977, 1983b).

Section 2 starts with the "equal variances case," Stein's setting (James and Stein, 1961) for which $V_1 = \cdots = V_k (\equiv V)$. Although equal variances are unusual in practice, this situation provides a rich and meaningful structure that has been studied widely because of its relative simplicity for mathematical investigation. Among other advantages, when $r > 0$ and the unknown means $\mu_i$ must be estimated, the equal variances situation allows easy recovery of risks and coverage probabilities merely by translating these quantities from the simpler $(k - r)$-dimensional situation when shrinkages are toward known means $\mu_i = 0$. Also with equal variances, ADM approximations to Bayes rules are easily developed for the range of scale-invariant priors (6), merely by solving a quadratic equation for $A$.

Section 2 continues by extending these ADM rules for the "unequal variance case" (the variances $V_i$ differ, as is common in practice). Section 2.8 introduces a new, more general approximation for the posterior means and variances, which allows any $r \ge 0$ so that shrinkages can be toward an estimated regression. With computational and programming methods similar to those of REML, noticeably more accurate procedures emerge.

Section 3 examines how well ADM methods approximate the exact Bayes rule. These approximations are good for small values of $k$ and they become exact as $k \to \infty$. Even the data analyst who insists on exact computations can find such approximations useful because of increased speed, even if only for doing preliminary analyses.

For the case $c = 1$ when $A$ is flat, Section 4 evaluates the resulting ADM-SHP procedure's performance in repeated sampling for relative mean squared errors and for interval coverages. In the equal variances case, and in the unequal variance examples considered, nominal coverages are achieved or exceeded for any $k \ge r + 3$. MLE and REML procedures cannot do this.

## 2. ADJUSTMENT FOR DENSITY MAXIMIZATION

This section starts by examining the inadequacy of MLE methods as a basis for inferences about shrinkage factors $B_i$ and random effects, and why the ADM approach for shrinkage constants should be better. For most of this section $r = 0$, the dimension of $\beta$, so that $\beta$ and all $\mu_i \equiv E(\theta_i)$ are assumed known. Thus, the only unknown Level-2 (nuisance) parameter is $A$, the between groups variance that governs the shrinkage factors $B_i \equiv \frac{V_i}{V_i + A}$. With $r = 0$, (3) and (4) simplify slightly to

$$\begin{aligned} &y_i \mid A \sim N(\mu_i, V_i + A), \\ &\text{with shrinkage factor } B_i = \frac{V_i}{V_i + A}, \text{ and} \\ &\theta_i \mid y_i, A \sim N\big((1 - B_i) y_i + B_i \mu_i, V_i(1 - B_i)\big). \end{aligned} \tag{7}$$

Let $S_i \equiv (y_i - \mu_i)^2 \sim (V_i + A)\chi_1^2$ independently. $\mathbf{S} \sim (S_1, \dots, S_k)'$ is a (minimal, if all $V_i$ differ) sufficient statistic for $A \ge 0$. Then $\hat{A}_i \equiv S_i - V_i$ for $i = 1, \dots, k$ are independent unbiased estimates of $A$ with $\text{Var}(\hat{A}_i) = 2(V_i + A)^2$. One could average these $\hat{A}_i$, weighted by the reciprocal of these variances to estimate $A$, iteratively until convergence, with a negative estimate of $A$ reset to 0. This produces $\hat{A}_{\text{MLE}}$, the MLE of $A$ (Efron and Morris, 1975).

In the equal variances case, $S_+ \equiv \sum_{i=1}^k S_i$ is complete and sufficient for $A$, $S_+ \sim (V + A)\chi_k^2$. Then $\hat{A}_{\text{unb}} \equiv \frac{1}{k} \sum \hat{A}_i = \frac{S_+}{k} - V$ is unbiased for $A$. Of course, $\hat{A}_{\text{unb}}$ can be negative, and $P(\hat{A}_{\text{unb}} < 0) = P(\chi_k^2 \le kB)$, where the equal shrinkages are $B \equiv \frac{V}{V+A}$. Because $k$ exceeds the median of $\chi_k^2$, $P(\chi_k^2 \le kB) > 1/2$

if $B$ is near 1 so that $A$ is near to zero. This inequality holds for any $k$ if $A \le \frac{2V}{3k}$, in which case $\hat{A}_{\text{unb}} < 0$ and $\hat{A}_{\text{MLE}} = 0$ more often than not. This issue of $\hat{A}_{\text{MLE}}$ being zero or quite small has received theoretical attention at least since Morris (1983b), and has been recognized for some time in practice (Bell, 1999), because its occurrence is not rare. Still, the problem has yet to be sufficiently recognized so as to be avoided in practice, and avoided in widely used software.

When $r = 0$ the likelihood function is proportional to

$$L_0(A) \equiv \left\{\prod_{i=1}^{k}(V_i + A)^{-1/2}\right\} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{k} S_i/(V_i + A)\right\}. \tag{8}$$

This is positive at $A = 0$ and decreasing near 0 if the $S_i$'s are small enough to make the exponential term be nearly constant. Then 0 is a local maximum and if $\hat{A}_{\text{MLE}} = 0$ Fisher's information cannot be used to assess the variance of the MLE. Furthermore, when $\hat{A}_{\text{MLE}} = 0$, the MLE of $\text{Var}(\theta_i | y, A) = V_i(1 - B_i)$ also is zero. An unwary data analyst who uses this for the width of a confidence interval would assert that $\theta_i = \mu_i$ with arbitrarily high confidence.

The left panel of Figure 1 illustrates a case when the logarithm of the posterior density of $A$, equivalently the log-likelihood $\log(L_0(A))$ since $A$ has a flat prior, cannot use Fisher's observed information to estimate the variance of $A$ since $\hat{A}_{\text{MLE}} = 0$, there is no stationary point, and the second derivative is not negative. The situation for these data is much improved by using ADM to arrive at the adjusted log-likelihood in the middle and right panels of Figure 1.

## 2.1 Comparing ADM and MLE Methods

MLE methods, viewed from a Bayesian (posterior probability) perspective, amount to finding the posterior mode of a parameter's distribution and its variance (reciprocal of observed information) when the parameter has a flat prior distribution. Normal distributions are used to approximate the MLE's distribution based on two derivatives of the log-likelihood. That works well when the likelihood is approximately Normal, for example, with large samples, but it works poorly when likelihoods are quite non-Normal, as can happen when estimating shrinkage factors.

Morris (1988b), on approximating posterior distributions, showed how to fit any prespecified Pearson family (Normal, Gamma, F, Beta, t, etc.) to a density (but also a likelihood function) by calculating two derivatives of the "adjusted" (posterior) density function. The adjustment, multiplying by the quadratic or linear function that generates the particular Pearson family, makes the maximizer approximate the mean of the parameter, and not its mode. For a nearly symmetric bell-shaped distribution or likelihood, the Normal is the best Pearson approximation, the adjustment is a constant. Then the mode agrees with the mean and the MLE is the ADM. For skewed likelihoods, the statistician may be able to choose a better approximating Pearson family, for example, the Beta family for shrinkage factors.

The following factors compare the ADM and its fitting process, perhaps starting with a flat prior on $A$, with that of the MLE.

1. Simplicity. An ADM fit is accomplished via a complexity level comparable to the MLE, that is, both require two derivatives.

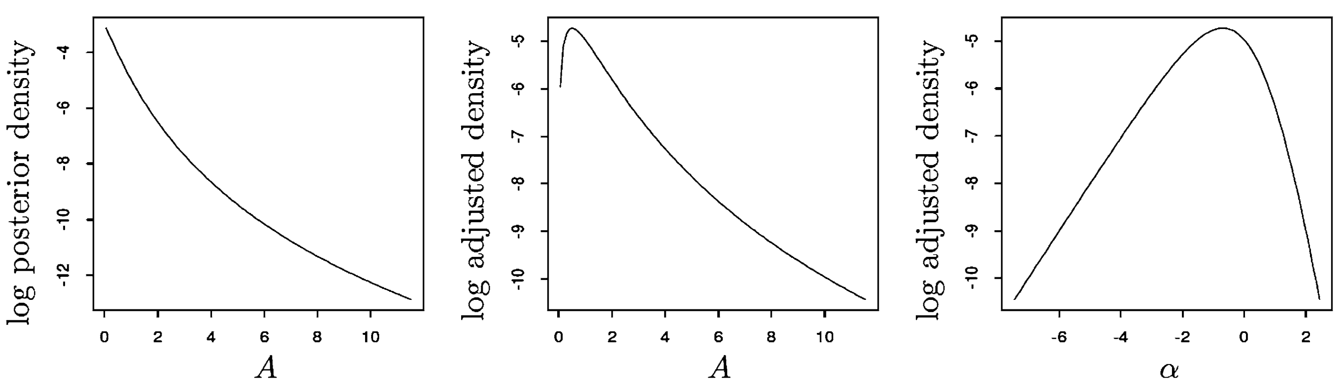


FIG. 1. *An equal variances example* ($V_i = 1$) *with the MLE on the boundary,* $S_+ = 8, k = 10, r = 0$. *The left panel plots the log posterior density for* $A$, *which is the log-likelihood for a flat prior on* $A$. *The middle panel plots the log adjusted density against* $A$, $\log(A * L_0(A))$ *in this case,* $L_0$*-likelihood function* (*see text and Section* 2.4). *The right panel shows the log adjusted density versus* $\alpha \equiv \log A$, *which looks more quadratic*.

2. Normality. If a Normal distribution is chosen for the matching Pearson family, the ADM approach agrees exactly with the MLE, and the variances in both cases are estimated by using Fisher's observed information.
3. Asymptotics. No matter which Pearson distribution is chosen, ADM provides the same asymptotic inferences (for large $k$) as the MLE. This holds because each Pearson family has an asymptotic Normal limit.
4. Linear expectations. While various transformations of a parameter can be considered for the MLE, ADM targets the mean. For example, shrinkage factors $B_i$ enter linearly in (4), so we approximate their means and variances, not $A$ or some other function of $B_i$.
5. Likelihoods? The ADM procedure could be termed ALM (Adjustment for Likelihood Maximization), to parallel with MLE language. ALM and MLE both work best when a version of the parameter is chosen to represent vague prior information, giving a relatively flat prior. We will see that ADM-SHP amounts to maximizing not the likelihood of $A$, as the MLE does, but the likelihood after adjustment via multiplication by $A$. Li and Lahiri (2010) proposed using "adjusted maximum likelihood estimator" that is identical to ADM if $r = 0$. They showed its advantages in small area estimation for estimating shrinkages and for constructing parametric bootstrap prediction intervals.
6. Multivariate ADM? Adjustments for density maximization agree with the MLE for approximations via the Multivariate Normal. The paucity of non-Normal multivariate Pearson families restricts ADM's extensions of the MLE to univariate parameters. However, hybrid extensions are possible, and here we use a multivariate Normal to approximate the $r$-dimensional vector $\beta$ and a Beta distribution for a shrinkage factor.

Given a prior distribution on $A \geq 0$, say $\pi(A)\,dA$ (proper or not), and still with $r = 0$, knowledge of

$$\text{(9)} \qquad \hat{B}_i \equiv E_\pi[B_i|y] \quad \text{and} \quad v_i \equiv \mathrm{Var}_\pi(B_i|y)$$

enables computation of two moments of $\theta_i$, which with $r = 0$ ($\mu_i$ known) are

$$\text{(10)} \qquad E[\theta_i|y] = (1 - \hat{B}_i)y_i + \hat{B}_i\mu_i,$$

$$\text{(11)} \qquad \mathrm{Var}(\theta|y) = V_i(1 - \hat{B}_i) + v_i(y_i - \mu_i)^2.$$

The second variance component in (11) often is not represented in MLE applications, understating variances and encouraging overconfidence.

### 2.2 How Maximum Likelihood Can Distort Shrinkage and Random Effects Inferences

Each of the following issues can cause overassessment of the information in the data. This perfect storm can have serious consequences when $k$ is small or moderate.

1. Nonlinearity. The posterior means and variances of the random effects are linear in $B_i$, not in $A$. $B_i(A) = V_i/(V_i + A)$ is a convex function of $A$, even if $\hat{A}$ were unbiased for $A$, one sees, Jensen's inequality, which states that $B_i(E[A|\mathbf{y}]) > E[B_i|\mathbf{y}]$, indicates that the plug-in shrinkage estimate would be biased too large. This is why the James–Stein estimator that shrinks according to $\hat{B}_{\mathrm{JS}} \equiv \frac{(k-2)V}{S_+}$, uses the $k-2$ in its numerator, and not $k$ (as in the MLE), and leads to smaller mean squared errors than when using MLE shrinkages $\hat{B} = \frac{kV}{S_+}$.
2. Boundary limits. Normal approximations to $B_i$ put positive probability outside the boundaries of the interval [0, 1].
3. Boundary pileup. While $0 \leq B_i \leq 1$ is guaranteed, even in the equal variances case $V/(V + \hat{A}_{\mathrm{unb}}) = \frac{kV}{S_+} > 1$ is possible. The MLE cannot exceed 1, but $\hat{B}_{\mathrm{MLE}} = \min(1, kV/S_+) = 1$ with positive probability. This pileup happens despite there being no prior distribution on $A$, other than $A = 0$ with certainty, that can allow $E[B|\mathbf{y}] = 1$ for any observation $\mathbf{y}$.
4. Skewness. $L_0(A)$ tends to be right-skewed, substantially when the modal value of $A$ is small. Alternatively, choose a fixed $i$ and replace $A$ by $B_i$ in the likelihood by substituting $A = \frac{1-B_i}{B_i}V_i$ in $L_0(A)$. The resulting likelihood function of $B_i$ will be left-skewed. Approximating such a skewed likelihood by a symmetric (Normal) distribution overstates the magnitude of $B_i$. A Beta density better approximates an asymmetric likelihood.
5. Zero variances. The MLE approach assesses $\mathrm{Var}(\theta_i|\mathbf{y}, A)$ as being $V_i(1 - \hat{B}_{i,\mathrm{MLE}})$. When $\hat{A}_{\mathrm{MLE}} = 0$, this approach in effect attributes perfect certainty to $A = 0$ and that $\theta_i = \mu_i$.
6. Variance components. Estimating the variance of $\theta_i$ by plugging into $V_i(1 - B_i)$ overlooks the variance component $v_i = \mathrm{Var}(B_i|y)$ which would account for the uncertainty in $A$ when estimating $B_i$. Ignoring the term $v_i(y_i - \mu_i)^2$ amounts to setting $v_i = 0$.

All six of these biases produces overconfidence. The unknown variance $A$ is underestimated, shrinkage $B_i$ is overestimated, and $\mathrm{Var}(B_i|\mathbf{y})$ is underestimated.

### 2.3 ADM, Adapted to Beta Distributions

The applications here require approximating the means and variances of the shrinkage factors $B_i$, $0 \le B_i \le 1$. Beta distributions are constrained to [0, 1], so are the obvious approximating Pearson distribution. Consider an exact Beta distribution for $B$ with $B \sim \text{Beta}(a_1, a_0)$ and density

$$(12)\quad f(B)\,dB = \frac{\Gamma(a_1)\Gamma(a_0)}{\Gamma(a_1+a_0)} B^{a_1-1}(1-B)^{a_0-1}\,dB.$$

Maximizing over $B$ gives $\hat{B} = \frac{a_1-1}{a_1+a_0-2}$, the mode (if $a_1, a_0 \ge 1$), not the mean. The "adjustment" for the Beta distribution maximizes the product $(B(1-B))f(B)$, giving $\hat{B} = \frac{a_1}{a_1+a_0}$, the mean of the Beta($a_1$, $a_0$) distribution. Maximizing a Beta density after multiplying by $B(1-B)$ produces the mean, not the mode.

Now let

$$(13)\quad \ell(B) = \log\{B(1-B)f(B)\}$$

$$(14)\quad = a_0 \log B + a_1 \log(1-B).$$

This is a concave function, maximized uniquely at a point interior to (0,1). We have $\ell'(B) = \frac{a_1}{B} - \frac{a_0}{1-B} = 0$ at $\hat{B} = \frac{a_1}{a_1+a_0}$. Then

$$(15)\quad -\ell''(B)|_{B=\hat{B}} = \frac{a_1}{\hat{B}^2} + \frac{a_0}{(1-\hat{B})^2} = \frac{a_1+a_0}{\hat{B}(1-\hat{B})}.$$

Thus, given $\hat{B}$ and $-\ell''(\hat{B}) > 0$ allows one to recover $a_1$ and $a_0$ via $a_1 + a_0 = -\ell''(\hat{B}) \cdot \hat{B}(1-\hat{B})$ and $a_1 = \hat{B}(a_1+a_0)$.

If $f(B)$ is a Beta($a_1, a_0$) density, exactly, then

$$E(B) = \hat{B},$$

$$(16)\quad v \equiv \text{Var}(B) = \frac{\hat{B}(1-\hat{B})}{a_1+a_0+1} = \frac{\hat{B}(1-\hat{B})}{1+\hat{B}(1-\hat{B})(-\ell''(\hat{B}))}.$$

If a density $f(B)$ is not exactly Beta but it lies near to a Beta density, the ADM approach proceeds similarly, based on two derivatives of $\log(B(1-B)f(B))$, and approximates $E[B] = \int_0^1 Bf(B)\,dB$ by $\hat{B}$, the maximizer of this adjusted density. The variance Var($B$) is approximated by (16), starting with

$$(17)\quad \ell(B) \equiv \log\{B(1-B)f(B)\}.$$

That is, ADM for a Beta approximation first finds $\hat{B} = \text{argmax}(\ell(B))$. Then it determines $-\ell''(\hat{B})$ and uses that to approximate Var($B$) by $\frac{\hat{B}(1-\hat{B})}{1+\hat{B}(1-\hat{B})(-\ell''(\hat{B}))}$.

This Beta distribution approximation to a density on [0,1] is exact if the original density is a Beta exactly, and it will be a good approximation if the match is close. Its asymptotic accuracy can be evaluated favorably (Morris, 1988b, with discussion).

It is useful when fitting shrinkages $B_i = B_i(A)$ to reexpress the results just outlined in terms of $A$, or equivalently in terms of its logarithm $\alpha = \log(A)$, being sure to include the Jacobian in the posterior density. Instead of using derivatives of $-\ell(B)$, the "invariant information" will be calculated, defined by

$$(18)\quad \text{inv.info} \equiv -\frac{d^2\ell(B)}{d\{\text{logit}(B)\}^2}\bigg|_{B=\hat{B}}.$$

The derivative $d\,\text{logit}(B)/dB = d\log(\frac{B}{1-B})/dB = 1/(B(1-B))$, which gives

$$(19)\quad \frac{d^2\ell(B)}{d\{\text{logit}(B)\}^2} = B^2(1-B)^2\ell''(B) + B(1-B)(1-2B)\ell'(B).$$

As $\ell'(\hat{B}) = 0$, we have $\text{inv.info} = (\hat{B}(1-\hat{B}))^2 \cdot (-\ell''(\hat{B}))$.

Thus, if $f(B)$ is (nearly) a Beta density $B \sim \text{Beta}(a_1, a_0)$, then $E[B] = \frac{a_1}{a_1+a_0} = \hat{B}$ with $\hat{B} = \text{argmax}(\ell(B))$, and the (approximate) variance is

$$(20)\quad \text{Var}(B) = \frac{\hat{B}(1-\hat{B})}{a_1+a_0+1} = \frac{(\hat{B}(1-\hat{B}))^2}{\text{inv.info} + \hat{B}(1-\hat{B})}.$$

Use of this invariant information is especially valuable because of the identity

$$(21)\quad -\frac{d^2\ell(B)}{d\{\text{logit}(B)\}^2} = -\frac{d^2\ell(B(A))}{d\{\log(A)\}^2} = -\frac{d^2\ell(B(A(\alpha)))}{d\alpha^2}.$$

This follows from $d\{\text{logit}(B)\} = d\log(\frac{V}{A}) = -d\alpha$ with $\alpha \equiv \log(A)$. The invariant information is the negative second derivative with respect to $\alpha$ of $\ell_2(\alpha)$, being the log density written as a function of $\alpha$:

$$(22)\quad \text{inv.info} = -\frac{d^2\ell(B)}{d\{\text{logit}(B)\}^2}\bigg|_{B=\hat{B}} = -\frac{d^2\ell(B(A))}{d(\log(A))^2}\bigg|_{A=\hat{A}} = -\frac{d^2\ell_2(\alpha)}{d\alpha^2}\bigg|_{\alpha=\hat{\alpha}}.$$

Thus, inv.info agrees with Fisher's observed information, but only if the parameter is $\alpha \equiv \log(A)$.

### 2.4 ADM for Estimating Shrinkage Constants

Now return to the Normal model with $r = 0$ and likelihood function $L_0(A)$. Suppose $A \geq 0$ has a prior density $\pi(A)$, not necessarily proper, and consider the shrinkage coefficient for component $i$, $1 \leq i \leq k$, $B_i = \frac{V_i}{V_i+A}$. The posterior density for $B_i$, given $\mathbf{y}$, is proportional to $L_0(A)\pi(A)\,dA \equiv f(B_i)\,dB_i$, where $A = V_i(1-B_i)/B_i$ and $dA = -V_i\,dB_i/B_i^2$. Then $f(B_i) \equiv L_0(A)\pi(A)V_i/B_i^2$ is proportional to the density of $B_i$. To apply ADM, define

$$\ell_0(B_i) \equiv \log\bigl(B_i(1-B_i)f(B_i)\bigr) \tag{23}$$

$$= \log(A\pi(A)L_0(A)) \equiv \ell(A). \tag{24}$$

Still thinking of $A$ as a function of $B_i$,

$$\frac{d\ell(A)}{dB_i} = \frac{dA}{dB_i}\frac{d\ell(A)}{dA} = \frac{-V_i}{B_i^2}\ell'(A). \tag{25}$$

The following theorem summarizes what has just been demonstrated about the ADM approximation by a Beta distribution for $B_i = V_i/(V_i + A)$, starting with a posterior density on $A$ that is proportional to $L_0(A)\pi(A)$.

THEOREM 1. *Given a prior density $\pi(A)$ and a likelihood function $L_0(A)$, the ADM procedure for a Beta distribution approximates the first two posterior moments of $B_i$ as*

$$E[B_i|\mathbf{y}] = \hat{B}_i = \frac{V_i}{V_i + \hat{A}}, \tag{26}$$

*where* $\hat{A} = \operatorname{argmax}(\ell(A))$, $\ell(A) \equiv \log(A\pi(A)L_0(A))$, *and*

$$v_i \equiv \operatorname{Var}(B_i|\mathbf{y}) = \frac{(\hat{B}_i(1-\hat{B}_i))^2}{\text{inv.info} + \hat{B}_i(1-\hat{B}_i)}, \tag{27}$$

*with* inv.info $\equiv -\ell''(\hat{A})\hat{A}^2$.

Neither $\hat{A}$ nor the invariant information depends on $i$ or on $V_i$.

### 2.5 Priors for Good Frequency Performance

Admissible rules, which are Bayes and extended Bayes rules (per the "fundamental theorem of decision theory"), can provide good frequency properties if they are based on priors that let the data speak. One way to do that restricts to scale invariant improper priors $\pi(A)\,dA = A^{c-1}\,dA$, $0 < c \leq 1$. As discussed earlier, given $k$, these priors with $c \geq c_k > 0$ ($c_k < 1/2$, but not too small) produce estimators of $\theta_i$ whose posterior means are minimax estimators for squared-error loss in the equal variance setting, so that for all vectors $\boldsymbol{\theta}$ (fixed),

$$E\sum_{i=1}^{k}\{(1-\hat{B}(S_+))y_i - \theta_i\}^2/V < k, \tag{28}$$

$$\hat{B}(S_+) \equiv E\left[\frac{V}{V+A}\Big|S_+\right]. \tag{29}$$

The choice $c = 0$, so $\pi(A)\,dA = dA/A$, puts essentially all mass at $A$ nearly 0, making $\hat{B}(S_+) = 1$ with certainty, no matter what the data say. This choice must be avoided, but sometimes it is not. As $c$ increases, shrinkages $\hat{B}(S_+)$ decrease. For $c = 1$ and for some smaller values, down to $c_k$, minimax and admissible estimators result.

Our preference $A \sim \text{Uniform}(0,\infty)$ is equivalent to Stein's harmonic prior, that is, for $\boldsymbol{\theta} \in \mathbb{R}^k$, $k \geq 3$, the (improper) measure on $\boldsymbol{\theta}$ is seen to be $d\boldsymbol{\theta}/\|\boldsymbol{\theta}\|^{(k-2)}$. This is the density of $\boldsymbol{\theta}$ if, independently for $i = 1,\ldots,k$, $\theta_i|A \sim N(0,A)$ and $A \sim \text{Unif}[0,\infty)$, as seen from

$$\int_0^\infty e^{-1/2\|\boldsymbol{\theta}\|^2/A}\frac{dA}{A^{k/2}} \propto \|\boldsymbol{\theta}\|^{2-k}. \tag{30}$$

This prior with $c = 1$, that is, $A \sim \text{Unif}[0,\infty)$, is strongly suggested in the equal variance case by the fact that the James–Stein shrinkage constant $\hat{B} = \frac{k-2}{S_+}$ is precisely the posterior mean $E[\frac{V}{V+A}|S_+]$ if $A \sim \text{Unif}[-V,\infty)$. Lopping off the impossible part where $A < 0$ leads to $A \sim \text{Unif}[0,\infty)$ (Morris, 1983a). That the James–Stein estimator is asymptotically optimal for large $\|\theta\|$ further suggests its use, that is, choosing $c = 1$. Still in the equal variances case, some values of $c < 1$, for example $c = 1/2$, shrink harder, which lowers the summed mean squared error if $\|\boldsymbol{\theta}\|^2$ is suspected not to be large. Experience with this flat prior on $A$ has borne out its good frequency properties in a variety of situations, also including for unequal variances. Supporting evidence is given in Sections 3 and 4.

### 2.6 Exact Moments for the Uniform Prior in the Equal Variances Case

The exact posterior means and variances of $B = V/(V + A)$ for $c = 1$, $A$ being uniform (Morris, 1983a), are as follows. Denote $m \equiv (k - r - 2)/2$, so $m = (k-2)/2$ when $r = 0$. If $r > 0$, the dimension of $\beta$, then the one can shrink toward the $r$-dimensional fitted subspace determined by $\hat{\beta} \equiv (X'X)^{-1}X'y$. In the $(k-r)$-dimensional space orthogonal to the range of $X$, shrinkage is toward the 0-vector. We therefore can focus on that $k - r$ subspace with $r = 0$ and $k$ replacing $k - r$ (or think of shrinkage as toward a known, fixed vector $\mu$ as here). Now with $y_i \sim N(\mu_i, V + A)$, let $S_+ \equiv \sum_{i=1}^{k}(y_i - \mu_i)^2$, and let $T \equiv S_+/2V$. The James–Stein estimate is $\hat{B}_{\text{JS}} \equiv m/T = (k - r - 2)V/S_+$. Let $M_m(T)$ be the moment generating function of a Beta$(1, m)$ distribution at $T$, a confluent hypergeometric function (Abramowitz and Stegun, 1964),

$$\begin{aligned} M_m(T) &\equiv \int_0^1 \exp[(1-B)T]\, dB^m \\ &= \Gamma(m+1)T^{-m}\exp(T)P(\chi^2_{2m} \le 2T). \end{aligned} \tag{31}$$

Then (Morris, 1983a),

$$\begin{aligned} \hat{B}_{\text{exact}} &\equiv E[B|S] \\ &= \frac{m}{T}\big(1 - 1/M_m(T)\big) \\ &= \frac{(k-r-2)V}{S_+} \\ &\quad \cdot \frac{P(\chi^2_{2m+2} \le S_+/V)}{P(\chi^2_{2m} \le S_+/V)}, \end{aligned} \tag{32}$$

$$\begin{aligned} v_{\text{exact}} &\equiv \text{Var}(B|S) \\ &= \frac{1}{m}\hat{B}^2_{\text{exact}} - (\hat{B}_{\text{JS}} - \hat{B}_{\text{exact}}) \\ &\quad \cdot \left(1 - \frac{m+1}{m}\hat{B}_{\text{exact}}\right). \end{aligned} \tag{33}$$

With $r = 0$, it follows that

$$\begin{aligned} \hat{\theta}_{\text{exact},i} &\equiv E[\theta_i|\mathbf{y}] \\ &= (1 - \hat{B}_{\text{exact}})y_i + \hat{B}_{\text{exact}}\mu_i, \end{aligned} \tag{34}$$

$$\begin{aligned} s^2_{\text{exact},i} &\equiv \text{Var}(\theta_i|\mathbf{y}) \\ &= V(1 - \hat{B}_{\text{exact}}) + v_{\text{exact}}(y_i - \mu_i)^2. \end{aligned} \tag{35}$$

The elegance of these formulas for the equal variances case is striking. Unfortunately, this disappears in the unequal variances case that invariably arises in practice, which motivates the search for relatively simple alternatives to exact calculations.

### 2.7 ADM for Shrinkages, Equal Variances Case

Maximum likelihood estimates have optimal asymptotic properties, but the small and moderate sample sizes ($k$) that arise in hierarchical modeling applications may be too small for the MLE to perform well. The mode of $A$, or more relevantly of $B$, may be quite inadequate approximations to the posterior mean that corresponds to a flat prior that makes the likelihood agree with the posterior density. Figure 1 provides a simple example for equal variances, scaled for a sample size $k = 10$ with shrinkage toward zero ($r = 0$) and a sufficient statistic $S_+ = 8$. $S_+ = 8$ is the mode of a $\chi^2_{10}$ distribution, and also is the largest value of $S$ that makes the James–Stein shrinkage estimate $\hat{B}_{\text{JS}} = 1$. Likelihood graphs like this are not uncommon in practice, even when unequal variances occur. The rightmost panels, which have made an adjustment to the likelihood, make it possible for two derivatives to capture the distribution, whereas there is no hope of this with the unadjusted left panel.

Figure 2 plots estimated shrinkages $\hat{B}$ against $T = S_+/2V$, for values of $k = 4, 10, 20$, each panel showing three different estimation methods: the exact shrinkage estimate for the flat harmonic prior $c = 1$, SHP (solid curve); the ADM approximation to the same prior (dotted); and the MLE $= \min(1, (m + 1)/T) = \min(1, k/S_+)$. The MLE shrinks much more heavily than the other two methods when $T$ (or $S_+$) is

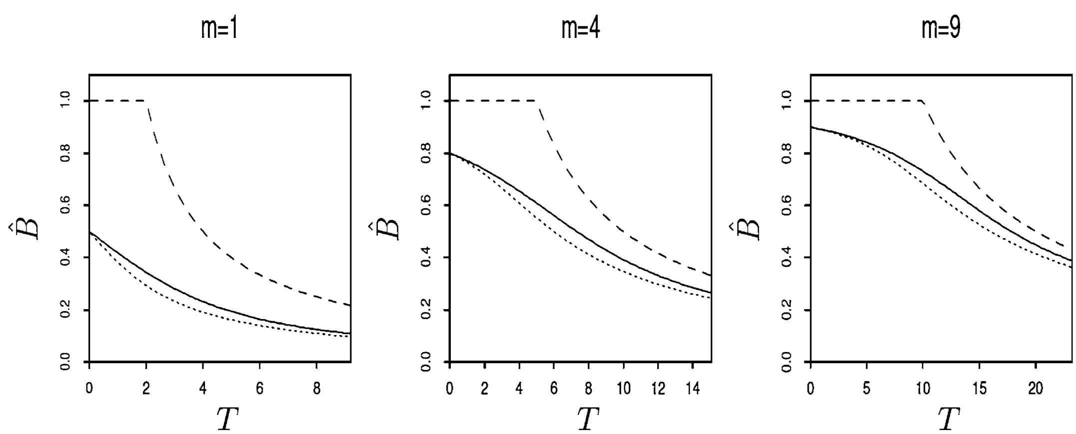


FIG. 2. *Plot of $\hat{B}$ versus $T \equiv S_+/2V$ from three different methods, with $m = 1, 4, 9$ ($k = 4, 10, 20$), respectively. The solid line is from the exact calculation, the dotted line is from ADM, and the dashed line is the MLE.*

small. The ADM shrinkage curves are fairly close to the exactly computed expected shrinkage in each case, but are slightly more conservative.

When $\beta$ is unknown so that $r > 0$, the marginal distribution of $A$ is gotten by integrating $\beta$ out of the joint posterior density of $\beta$ and $A$ (which is done in the next section, and extended to unequal variances). The marginal density is neatly written in this equal variances case in terms of the sum of squared residuals, $S_+ \equiv \sum_i (y_i - \hat{y}_i)^2$ and $\hat{y} \equiv X\hat{\beta}$ as

$$p(A|y) \propto (V + A)^{-(k-r)/2} \cdot \exp\left\{-\frac{S_+}{2(V + A)}\right\}\pi(A). \tag{36}$$

For $\pi(A) \propto A^{c-1}$, the logarithm of the adjusted density (multiplying by $A$) is

$$\ell_2(A|y) \equiv c \log A - (m + 1)\log(V + A) - \frac{TV}{V + A}, \tag{37}$$

$T \equiv S_+/2V$. With no covariates, $r = 0$, this equation continues to hold with $m = (k - 2)/2$.

Now,

$$\begin{aligned}\frac{d\ell_2(\alpha)}{d\alpha} &= A\frac{d\ell_2}{dA} \\ &= -\big((m + 1 - c)A^2 \\ &\quad - (2c + T - m - 1)VA - cV^2\big) \\ &\quad /(V + A)^2.\end{aligned} \tag{38}$$

The numerator of (38) is a convex quadratic function of $A$ (with $m + 1 - c > 0$) which is negative at $A = 0$. It therefore has two real roots, one negative and unacceptable. The positive root is the ADM estimator $\hat{A}$. Then,

$$\begin{aligned}\hat{B} &\equiv \frac{V}{V + \hat{A}} \\ &= \frac{2(m - c + 1)}{T + m + 1 + \sqrt{(T - m - 1)^2 + 4cT}}.\end{aligned} \tag{39}$$

Note that $\hat{B}$ is monotone decreasing in $T$ and that $\hat{B}$ reaches its maximum, $1 - c/(m + 1) < 1$ at $T = 0$. Shrinkage is bounded away from 100% if $c > 0$, for example, if $c = 1$ and $r = 0$ the maximum shrinkage is $(k - 2)/k$. These shrinkages $\hat{B}$ decrease as $c$ increases and as $c \to 0$ in (39), $\hat{B} \to \min((m + 1)/T, 1)$. Of course $c = 0$ is not allowed because then the posterior guarantees 100% shrinkage, no matter what the data say.

Define $\alpha \equiv \log(A)$ and $\hat{\alpha} \equiv \log \hat{A}$. Then for any $c$, the invariant information = inv.info satisfies

$$\begin{aligned}\text{inv.info} &= -\frac{d^2\ell_2}{d\alpha^2}\bigg|_{\alpha=\hat{\alpha}} \\ &= m(1 - \hat{B})^2 + \hat{B}^2 + (1 - c)(1 - 2\hat{B}).\end{aligned} \tag{40}$$

Matching the first and second derivatives of the two densities (i.e., of the adjusted density and of a Beta($a_1$, $a_0$) density) gives

$$a_1 = \frac{\text{inv.info}}{1 - \hat{B}}, \quad a_0 = \frac{\text{inv.info}}{\hat{B}},$$

and this Beta distribution has variance

$$\begin{aligned}v &= \frac{\hat{B}(1 - \hat{B})}{a_0 + a_1 + 1} \\ &= \frac{\hat{B}^2(1 - \hat{B})^2}{m(1 - \hat{B})^2 + (1 - c) + (2c - 1)\hat{B}}.\end{aligned} \tag{41}$$

When $c = 1$, the ADM approximations in this equal variances case to the posterior moments of $B = V/(V + A)$ are

$$\hat{B} = \frac{2(k - r - 2)V}{S_+ + kV + \sqrt{(S_+ - kV)^2 + 8S_+V}}, \tag{42}$$

$$v = \frac{\hat{B}^2(1 - \hat{B})^2}{m(1 - \hat{B})^2 + \hat{B}}. \tag{43}$$

For the SHP case $c = 1$ in Figure 2, $\hat{B}$ is plotted as a function of $T$, showing that the ADM estimate of $B$ shrinks slightly less than the exactly computed $B$, while it matches exactly at $T = 0$, and asymptotes to the exact value for large $T$. The MLE produces much larger shrinkages.

Figure 3, as in Figure 2, also shows graphs for the SHP ($c = 1$) and with curves for $m = 1, 4, 9$ (e.g., if $r = 0$, then for $k = 4, 10, 20$). It reveals that the ADM approximation to $v$ corresponds well with the exact posterior variance of a shrinkage factor, each as a function of its own shrinkage $\hat{B}$. In both cases the shrinkage $\hat{B}$ decreases monotonically as the sufficient statistic $S_+$ rises. Figure 3 shows ADM's excellent ADM approximation of the exact variance, and that it becomes exact as $T$ nears 0 (where maximal shrinkage in both cases is for $\hat{B} = m/(m + 1) = (k - r - 2)/(k - 2)$).

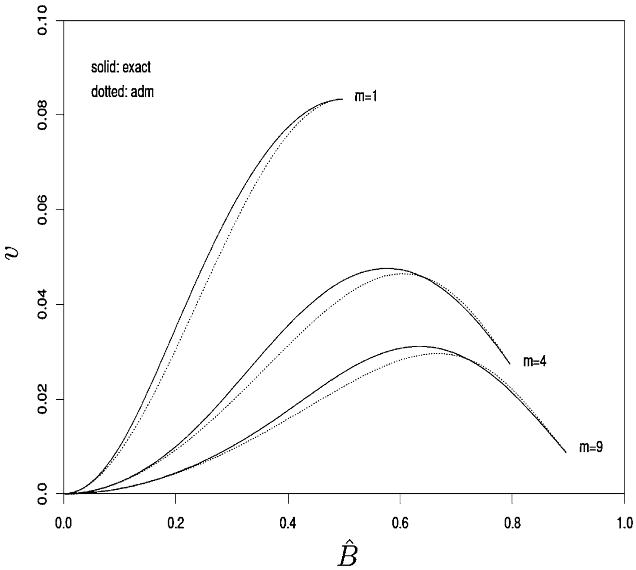


FIG. 3. *Plot of $v$ versus its own $\hat{B}$ from two different methods. The solid line is from the exact method, that is, formulas* (32) *and* (33), *and the dotted line is from the approximate method, formulas* (42) *and* (43).

For any $r \geq 0$ in this equal variance case, the preceding estimates of the shrinkages and of their variances provide the following estimates of the means and the variances of the random effects $\theta_i$ in terms of the ADM approximations to the posterior moments $\hat{B}$ and $\hat{\beta} \equiv (X'X)^{-1}X'y$:

$$\hat{\theta}_i \equiv \hat{E}(\theta_i|y) = (1-\hat{B})y_i + \hat{B}x_i'\hat{\beta}, \tag{44}$$

$$s_i^2 \equiv \widehat{\mathrm{Var}}(\theta_i|y) = V(1-\hat{B}) + V\big(x_i'(X'X)^{-1}x_i\big)\hat{B} + v(y_i - x_i'\hat{\beta})^2. \tag{45}$$

Note that $s_i^2$ depends on $i$ by increasing proportionally to the squared residual, as one would expect because mis-estimation of $B$ hardly matters when $(y_i - x_i'\hat{\beta})^2$ is small. These results are seen most easily by using a least squares regression predictor in the $r$-dimensional range space of $X$, and shrinking to 0 in the $(k-r)$-dimensional orthogonal subspace. The extension to the unequal variance case, which is next, is more complicated.

### 2.8 The Unequal Variances Case With Regression

An ADM approach to fitting our general model starts by integrating out the $\{\theta_i\}$ to get, in matrix notation,

$$y|\beta, A \sim N_k(X\beta, D_{V+A}), \tag{46}$$

where $D_{V+A} \equiv \mathrm{diag}(V_i + A)$ is a $k$-by-$k$ diagonal matrix. With $\beta$ having a flat prior on $R^r$, standard calculations with (46) lead to

$$\hat{\beta}_A \equiv E(\beta|y, A) = (X'D_{V+A}^{-1}X)^{-1}X'D_{V+A}^{-1}y. \tag{47}$$

With $A$ known, $\hat{\beta}_A$ is at once both the posterior mean and the weighted least squares estimate of $\beta$. The full distribution, given $A$, is

$$\beta|A, y \sim N_r(\hat{\beta}_A, (X'D_{V+A}^{-1}X)^{-1}). \tag{48}$$

The objective is to make inferences about the vector $\theta = (\theta_1, \ldots, \theta_k)$ with conditional distribution

$$\theta|\beta, A, y \sim N_k\big((I - B_A)y + B_A X\beta, (I - B_A)V\big). \tag{49}$$

This is (4) in matrix notation, with $I$ the $k$-by-$k$ identity matrix, $V \equiv \mathrm{diag}(V_1, \ldots, V_k)$ and $B_A \equiv \mathrm{diag}(B_i = V_i/(V_i + A))$. Integrating out $\beta$, with help from (47), it follows that

$$\theta|A, y \sim N_k\big((I - B_A)y + B_A X\hat{\beta}_A, (I - B_A)V + V^{1/2}B_A^{1/2}P_A B_A^{1/2}V^{1/2}\big), \tag{50}$$

where in (50) $P_A$ is a $k \times k$ projection matrix of rank $r$,

$$P_A \equiv D_{V+A}^{-1/2}X(X'D_{V+A}^{-1}X)^{-1}X'D_{V+A}^{-1/2}. \tag{51}$$

When $A$ has prior density element $\pi(A)\,dA$, the posterior density of $A$, given $y$, follows:

$$p(A|y) \propto |D_{V+A}|^{-1/2}|X'D_{V+A}^{-1}X|^{-1/2} \cdot \exp\big(-\tfrac{1}{2}(y - X\hat{\beta}_A)' \cdot D_{V+A}^{-1}(y - X\hat{\beta}_A)\big). \tag{52}$$

The logarithm of this adjusted posterior density, with $\alpha = \log(A)$, is

$$l(\alpha) = \log(A\pi(A)) - \frac{1}{2}\sum_1^k \log(V_i + A) - \frac{1}{2}\log|X'D_{V+A}^{-1}X| - \frac{1}{2}(y - X\hat{\beta}_A)'D_{V+A}^{-1}(y - X\hat{\beta}_A). \tag{53}$$

Denote $\hat{\alpha} \equiv \mathrm{argmax}(l(\alpha))$, set $\hat{A} = \exp(\hat{\alpha})$, and define $\mathrm{inv.info} \equiv -l''(\hat{\alpha})$. Then the ADM approximation, with $\hat{B}_i \equiv V_i/(V_i+\hat{A})$, is $B_i \equiv \frac{V_i}{V_i+A} \sim$ Beta with approximate mean $E(B_i) = \hat{B}_i = V_i/(V_i+\hat{A})$ and variance $v_i = \mathrm{Var}(B_i) = \{\hat{B}_i(1-\hat{B}_i)\}^2/\{\text{inv.info}+\hat{B}_i(1-\hat{B}_i)\}$, both moments depending on the prior $\pi(A)$. Maximizing $\ell(\alpha)$ and determining its second derivative at $\hat{\alpha}$, the negative of the invariant information, can be done by numerical methods, by Newton's method (which requires matrix derivatives), or by other means that include an EM technique available in Tang (2002).

Given $\hat{A}$ and the values $\{\hat{B}_i, v_i\}, i = 1, \dots, k$, one could insert $\hat{A}$ into (50) to estimate both posterior moments of the $\theta_i$. However, that underestimates the variance and makes no use of the $\{v_i\}$, so we proceed as follows, leading to a main theorem.

Define $\hat{\beta}$ as $\hat{\beta}_A$ evaluated at $\hat{A}$ and $\hat{y} \equiv X\hat{\beta}$. Then from (50), and approximating $\hat{\beta}_A$ by $\hat{\beta}$,

$$E(\theta_i|A, y) \doteq y_i - B_i(y_i - \hat{y}_i), \tag{54}$$

$$\mathrm{Var}(E(\theta_i|A, y)) \doteq v_i(y_i - \hat{y}_i)^2. \tag{55}$$

To minimize complications in making our final approximations to $E(\theta_i|y)$ and $\mathrm{Var}(\theta_i|y)$, we neglect variations of $\hat{\beta}_A$ in (47) and $P_A$ in (51) as $A$ varies around $\hat{A}$. This is exact in the equal variances case because both $\hat{\beta}_A$ and $P_A$ do not depend on $A$, and it will be nearly true if the $\{V_i\}, i = 1, \dots, k$ differ only slightly. With unequal variances both $\hat{\beta}_A$ and (51) involve weights that depend on $\{V_i + A\}$. If $\hat{A}$ is near $A$, as happens when $k$ is large, then $\frac{V_i+\hat{A}}{V_i+A}$ is near 1. With data, one can evaluate

$$\mathrm{Var}\left\{\left(\frac{V_i+\hat{A}}{V_i+A}\right)\Big|y\right\} = \mathrm{Var}\left(\frac{B_i}{\hat{B}_i}\Big|y\right) = \frac{v_i}{\hat{B}_i^2}. \tag{56}$$

These variances may be acceptably small, and $v_i/\hat{B}_i^2$ diminishes as $1/k$ as $k \longrightarrow \infty$.

THEOREM 2. *Assume the model* (1), (2), *and the prior in* (5). *Write* $\hat{B}_i$ *and* $v_i$ *as the ADM approximations to* $E(B_i|y) = E(\frac{V_i}{V_i+A}|y)$ *and to* $\mathrm{Var}(B_i|y)$. *Assume* $E(\hat{\beta}_A|y) \doteq \hat{\beta} \equiv \hat{\beta}_A$ *and* $E(P_A|y) \doteq P_{\hat{A}}$. *Then for* $i = 1, \dots, k$

$$E(\theta_i|y) \doteq (1-\hat{B}_i)y_i + \hat{B}_i x_i'\hat{\beta} \equiv \hat{\theta}_i, \tag{57}$$

$$\begin{aligned}\mathrm{Var}(\theta_i|y) \doteq{}& \bigl(1-(1-p_{i,i})\hat{B}_i\bigr)V_i\\ &+ v_i(y_i-\hat{y}_i)^2.\end{aligned} \tag{58}$$

*Here* $p_{i,i}$ *is the* $i$*th diagonal term in* $P_{\hat{A}}$.

PROOF. Equation (57) follows from (50), (54) and $E(\hat{\beta}_A|y) = \hat{\beta}$, since

$$E(\theta_i|y) = E\{(1-B_i)y_i + B_i\hat{y}_i\}|y.$$

Now use EVE's law (total variation) to get, from (50) and (55),

$$\mathrm{Var}(\theta_i|y) = E\,\mathrm{Var}(\theta_i|A, y) + \mathrm{Var}(E\theta_i|A, y) \tag{59}$$

$$\begin{aligned}&= E\{(1-B_i)V_i + B_i p_{i,i}V_i\}|y\\ &\quad + v_i(y_i-\hat{y}_i)^2,\end{aligned} \tag{60}$$

which is (58).

In our experience, these regression approximations when $\pi(A) = A^{c-1}$, and $c = 1$ especially, have been quite satisfactory. Tang (2002) provides a basis for making more precise approximations to $E(P_A|y)$ and to $E(\hat{\beta}_A|y)$ based on matrix and determinant derivatives. In the equal variance case, the theorem's two moments are exact provided exact formulas for $E(B_i|y)$ and $\mathrm{Var}(B_i|y)$ are used. However, Normality of $\theta_i|A, y$ does not hold exactly for $\theta_i$ after averaging over $A|y$, although that Normal approximation is commonly made. □

## 3. APPROXIMATION ACCURACY

### 3.1 Approximation Accuracy of Shrinkages and the Random Effects

Figures 2 and 3 show in the equal variance setting that even for small samples like $k = 4, 10, 20$, the ADM approximation of the first two exactly computed posterior moments of $B$ is quite good. Our end goal, however, is verifying this leads to good approximations of the posterior means and variances of each random effect $(\theta_i, i = 1, \dots, k)$.

First, in the equal variance situation with $r = 0$, we compare the weighted average of posterior mean squared error of the $\theta_i$ values via the ADM approximation with this measure with the "exact" posterior mean. Let us measure the difference of their mean squared errors, given the data $y$, by computing

$$E\left\{\sum_{i=1}^k (\hat{\theta}_i - \theta_i)^2|y\right\} \tag{61}$$

for the ADM approximation, with the expectation calculated exactly, when $\pi(A) = 1$. Now

$$\begin{aligned}&E\left\{\sum_{i=1}^k(\hat{\theta}_i-\theta_i)^2|y\right\}\\ &\quad= E\left\{\sum_{i=1}^k(\hat{\theta}_i-\hat{\theta}_{e,i})^2|y\right\} + E\left\{\sum_{i=1}^k(\hat{\theta}_{e,i}-\theta_i)^2|y\right\}\\ &\quad= \sum_{i=1}^k(\hat{\theta}_i-\hat{\theta}_{e,i})^2 + \sum_{i=1}^k s_{e,i}^2,\end{aligned}$$

where the subscript $e$ denotes estimates done exactly (see Section 2.6), with $s^2_{e,i}$ is given in (35). Therefore

$$\frac{\sum_{i=1}^{k}(\hat{\theta}_i - \hat{\theta}_{e,i})^2}{\sum_{i=1}^{k} s^2_{e,i}} \tag{62}$$

measures how well the ADM approximation works for random effects estimates, smaller values indicating better approximations. The highest (worst) ratio is 1.1% which occurs for $k$ near 20, and for 60% shrinkage. Greater accuracy holds for $k < 20$ and for $k > 20$. Thus, in the equal variances setting, the conditional mean squared errors of the ADM approximation and the exact estimator of $\theta$ never differ by more than 1.1%.

Now, still with $\pi(A) = 1$, consider the unequal variance case and ADM's accuracy for approximating the exact Bayes estimator of $\theta$. The following example involves two groups of variances for the $y_i$ values, and estimates the unknown mean vector $\mu_1 = \cdots = \mu_{10}$ in the second level (so $k = 10, r = 1$). Five "small" variances are set at $V_1 = V_2 = \cdots = V_5 = 0.55$, and five "large" ones at $V_6 = \cdots = V_{10} = 5.5$. Their maximum-to-minimum variance ratio is a factor of 10, and their harmonic mean is 1.0 (for convenience only). Shrinkages $B_1 = \cdots = B_5 < B_6 = \cdots = B_{10}$ are toward the nine-dimensional subspace orthogonal to the unit vector. We calculated exact and ADM means and variances of these shrinkages, which depend on the separate values of the two-dimensional statistic $T_1, T_2$ (these two sums of squares are standardized by their respective $2V_i$, each summed over its respective subgroup of size 5, both centered on their common fitted grand mean).

Figure 4 concerns shrinkages for the first five components with small variances, $V_j = 0.55$, and Figure 5 shows shrinkages for the five components with large variances, $V_j = 5.5$. The left panels of each figure show shrinkage factor patterns for three different rules: the MLE (dashed curve), the exactly computed shrinkage using the harmonic prior for which $A$ has a flat density (solid curve), and the ADM approximations to that shrinkage factor (dotted curve). These are graphed as a function of $T_1$ (Figure 4) and $T_2$ (Figure 5) with separate displays, each conditional on one of four different values of the opposite $T_j$.

Both figures show that the MLE has quite large shrinkages, just as for equal variances. The relationship between the ADM approximation and the exactly computed expected shrinkage that the ADM approximates is similar to what was seen in the equal variance case. The right-hand panels of each figure show good agreement between the ADM variance approximation and the exactly computed variances $v_i$ when each is plotted against its own shrinkage. The maximum shrinkages for ADM and the exact rule are limited to values $< 1$, curtailing the horizontal axes for plots of $\mathrm{Var}(B_i)$.

To summarize for the prior $\pi(A) = 1$, the ADM approximations of exact shrinkage factors for posterior means and variances of shrinkage factors are slightly conservative, but generally are in good agreement with the exact values obtained in the equal variance case. Similar results hold for the unequal variance case when variances $V_i$ differ by a factor of 10 and when $r = 1$.

## 4. COVERAGE PROBABILITIES AND RISK FUNCTIONS

Confidence interval coverage rates for $\theta_i$ are evaluated next for the two main procedures of Section 2, both based on assuming $A > 0$ has a flat prior $\pi(A) = 1$ so that the posterior density is the likelihood function. One procedure, labeled "exact" here, evaluates the exactly computed posterior means and variances of $\theta_i$, given $y$, as in (34) and (35) for the equal variances case, and otherwise by numerical integration. It then assigns a Normal distribution with these two moments to determine a posterior interval. The second approach uses Normal distributions in the same way, but centered and scaled via the ADM approximations of these two moments in (44) and (45), or when $r \geq 0$ and with unequal variances, as in (57) and (58). Normal distributions are not exact for $\theta_i$, since the actual distributions are skewed (right-skewed for relatively large $y_i$, and left-skewed for small $y_i$). This matters less in repeated sampling evaluations that randomize over $y$, making skewnesses average to zero for each $i$.

For all $i = 1, \dots, k$, we seek two-tailed frequency coverage probabilities as a function of $A$:

$$\Pr\left[\frac{(\theta_i - \hat{\theta}_i)^2}{s_i^2} \leq (z^*)^2 | A\right], \tag{63}$$

when the nominal coverage is 95%, so $z^* = 1.96$. Each procedure studied uses its own estimate $s_i^2$ of the conditional variance of $\theta_i$. A related measure directly assesses how well each $s_i^2$ envelops the expected squared error, given $A$, with values $\leq 1$ indicating that $s_i$ assigns sufficiently large intervals:

$$E\{(\theta_i - \hat{\theta}_i)^2 / s_i^2 | A\}. \tag{64}$$

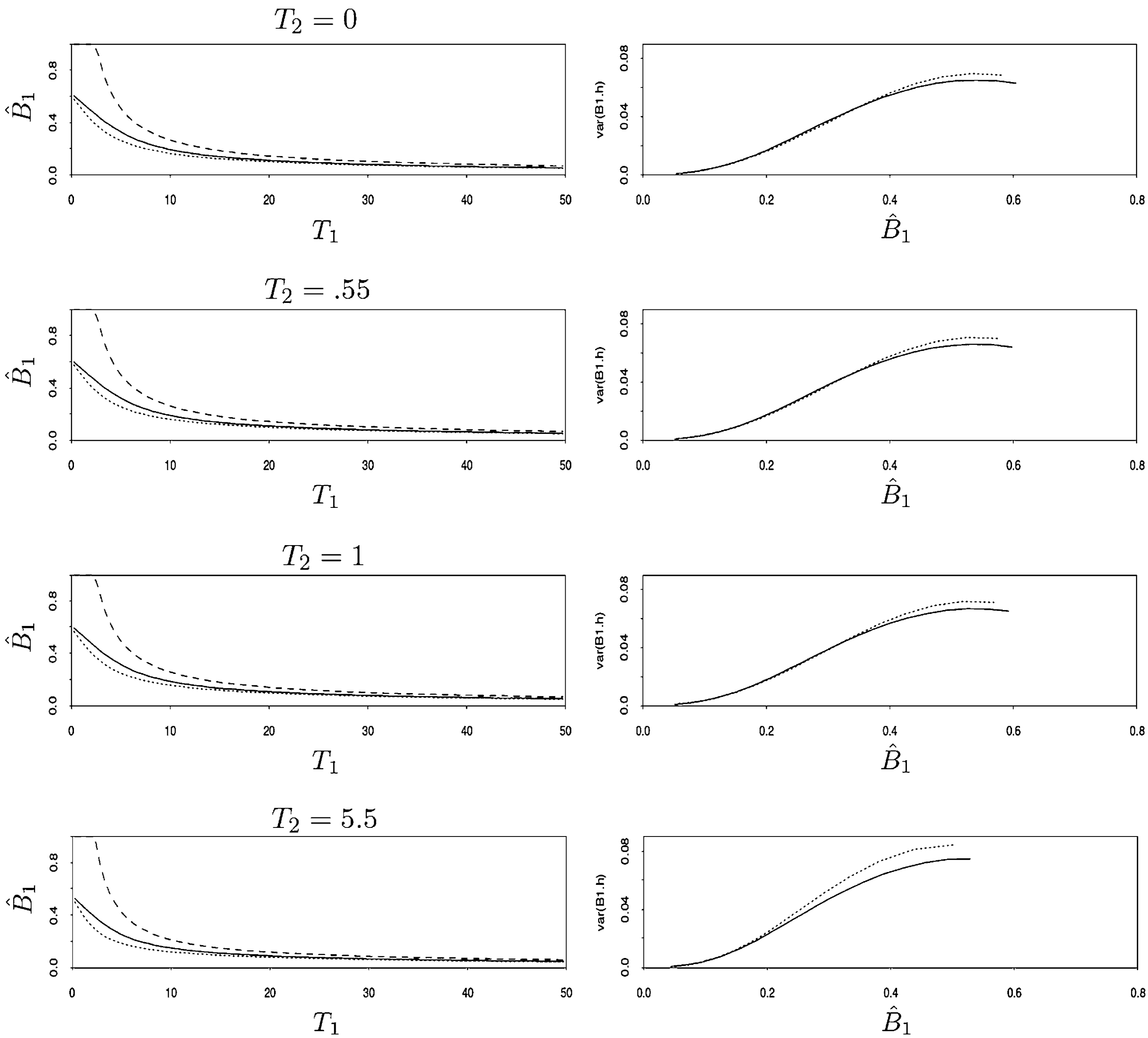


FIG. 4. *Approximation accuracy for two groups of variances, here for the small variance group* ($k = 10$, $r = 1$, $V_1 = \cdots = V_5 = 0.55$, $V_6 = \cdots = V_{10} = 5.5$). *The left-hand side plots* $\hat{B}_1$ *against* $T_1$, *with* $T_2$ *fixed at various values* (*which correspond to* $A = 0, 0.55, 1, 5.5$). *The right-hand side plots* $\mathrm{Var}(B_1|\mathrm{data})$ *against* $\hat{B}_1$. *Solid line is from the exact method, dotted line from ADM approximation, long dashed line is MLE.*

Details of the simulation are in Tang (2002), where Rao–Blackwellization increased the accuracy by evaluating some conditional Normal distributions exactly, given $A$ and $y$. That is, for (64),

$$\Pr\left\{\frac{(\hat{\theta}_i - \theta_i)^2}{s_i^2} \le (z^*)^2 | A\right\}$$
$$= E\left[\Pr\left\{\frac{(\hat{\theta}_i - \theta_i)^2}{s_i^2} \le (z^*)^2 | y, \beta, A\right\}\middle| A\right]$$
$$= E\left[\Phi\left\{\frac{\hat{\theta}_i - (1 - B_i)y_i - B_i x_i'\beta + z^* s_i}{\sqrt{V_i(1 - B_i)}}\right\} - \Phi\left\{\frac{\hat{\theta}_i - (1 - B_i)y_i - B_i x_i'\beta - z^* s_i}{\sqrt{V_i(1 - B_i)}}\right\}\middle| A\right].$$

### 4.1 Equal Variances Example

Figure 6 plots the actual coverage probabilities for the three confidence interval procedures, each against the possible "true" $B$ values, for three equal variance

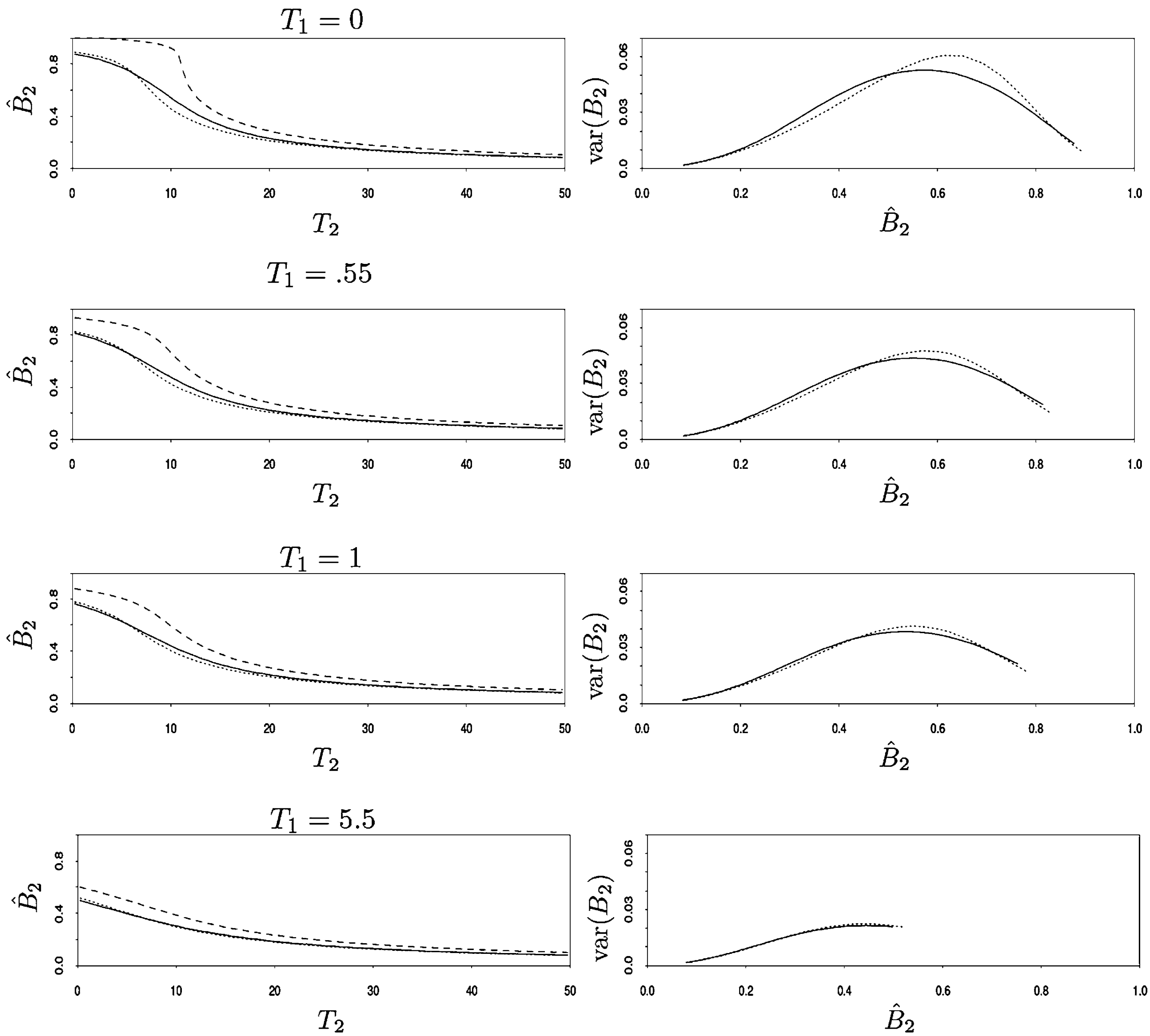


FIG. 5. *Approximation accuracy for two groups of variances, here for the large variance group* ($k = 10$, $r = 1$, $V_1 = \cdots = V_5 = 0.55$, $V_6 = \cdots = V_{10} = 5.5$). *The left-hand side plots* $\hat{B}_2$ *against* $T_2$, *with* $T_1$ *fixed at various values* (*which correspond to* $A = 0, 0.55, 1, 5.5$). *The right-hand side plots* $\mathrm{Var}(B_2|\mathrm{data})$ *against* $T_2$. *Solid line is from the exact method, dotted line from ADM approximation, long dashed line is MLE.*

procedures always with $r = 0$, and for $k = 4$ ($m = 1$), $k = 10$ ($m = 4$) and $k = 20$ ($m = 9$). For each $B = 0.005, 0.015, \ldots, 0.995$, 1000 data sets were generated and the interval procedures for "exact," its ADM approximation, and the MLE were evaluated and averaged to estimate the coverage probabilities. Confidence intervals for the MLE were determined simply by taking each variance to be the MLE $V(1 - B)$. These MLE coverages are plotted with long dashes in Figure 6. When shrinkage $B$ is large, these MLE intervals give poor coverages, ultimately dropping to just under 50%, as shown in Section 2.

The graph of Figure 6 is redone in the first row of Figure 7, but without the MLE. That allows an amplified scale that shows the slight differences in coverage rates between the "exact" rule and its ADM-SHP approximation. The ADM-SHP coverages meet or exceed $\geq 0.95$ for all $A$ (within simulation error). The "exact" procedure's coverages can be slightly nonconservative, but its lowest coverage is at least 0.945

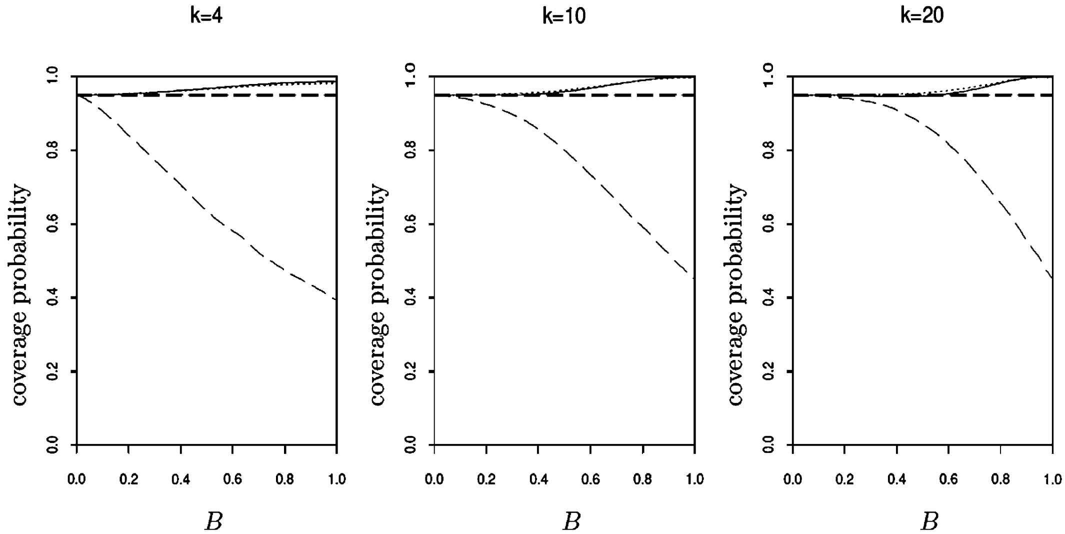


FIG. 6. *Plot of coverage probabilities of* $\theta_i$ (*random effects*) *for equal variances without regression, each against the true shrinkage factor. In the three graphs* ($k = 4, 10, 20$), *both the "exact" Bayes* (*solid curve*) *and its ADM-SHP approximation* (*dotted curve*) *achieve approximately the nominal* 0.95 *coverage rates* (*as indicated by the bold dashed horizontal line*), *or higher. The MLE* (*long dashes*) *can be markedly nonconservative, especially with large true shrinkages B* (*A near* 0). *As A approaches* 0, *MLE coverages fall below* 50%, *however large k might be.*

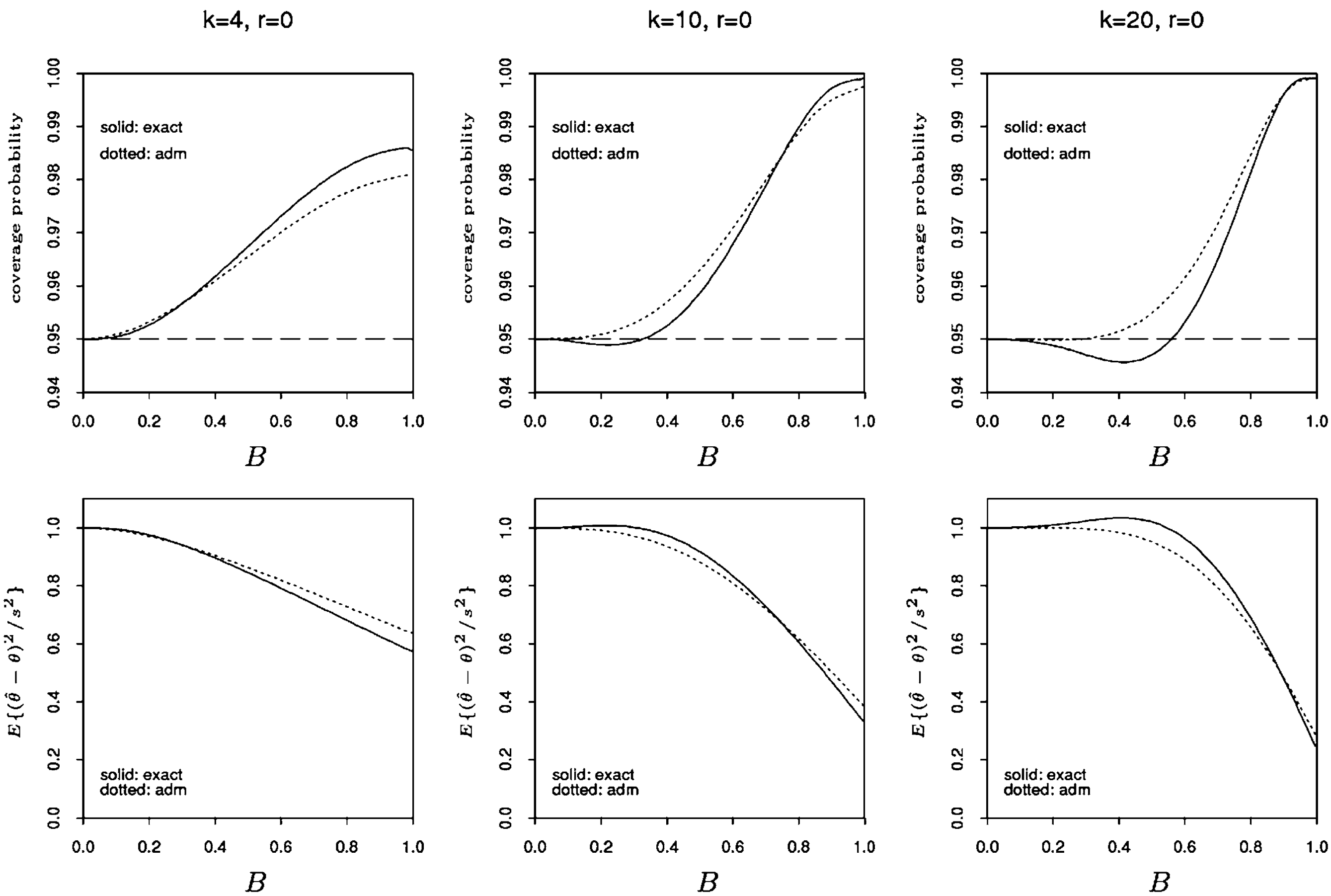


FIG. 7. *Plot of coverage probabilities and standardized risk functions for equal variances without regression. The first row plots coverage probabilities against true values of* $B = V/(V + A)$ *on a larger scale than in Figure* 6, *with ADM-SHP coverages being the dotted curves. The second row plots the expected value of the loss function calibrated by* $s_i^2$ (64).

(when $k = 20$ and $B = 0.4$) for all $k$ shown. The ADM-SHP intervals achieve (or exceed) their nominal 0.95 coverage rates by having slightly wider intervals than "exact," due to ADM's reduced shrinkage estimates and its larger variance estimates $v$, as studied in Section 3. As $B$ increases both methods become quite conservative, with coverages well above 0.95.

The bottom row of Figure 7 plots the function (64) against $B$ to compare the two different methods. Values less than 1.0 indicate that the estimated variances $s_i^2$ average to as much as or more than the average mean square. This further suggests that the interval coverages will (nearly) provide the nominal coverage (95%) for all values of $A > 0$.

### 4.2 An Unequal Variances Example: Two Groups of Variances

We return to the unequal variances example of Section 3 with $k = 10$, $r = 1$, $V_1 = \cdots = V_5 = 0.55$, and $V_6 = \cdots = V_{10} = 5.5$. For this simulation, 100 data sets were generated for each of 50 values $B_0 = 0.01, 0.03, \ldots, 0.99$, where $B_0 \equiv V_0/(V_0 + A)$ and $V_0 = 1$ is the harmonic mean of the $V_i$. Nominal 95% confidence intervals for each $\theta_i$ were evaluated for each data set. The confidence rates and average calibrated losses (64) then were averaged over the simulated values.

Figure 8 plots coverages of the ADM-SHP intervals and calibrated risk functions (64) for $\theta_1$ and for $\theta_{10}$ as $B_0 = 1/(1 + A)$ varies. The upper left panel of Figure 8 plots the coverage probabilities against $B_0$ for the group of five with small variances $V_i = 0.55$, and the

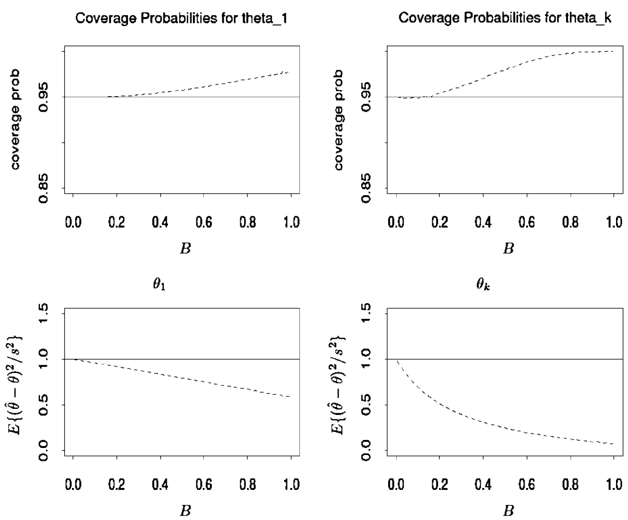


FIG. 8. *Plot of ADM-SHP coverages and expected value of average calibrated losses against $B_0 = 1/(1 + A)$. Here $k = 10, r = 1$, $V_1 = \cdots = V_5 = 0.55$, $V_6 = \cdots = V_{10} = 5.5$.*

upper right for the remaining group of five with large variances $V_i = 5.50$. As $B_0$ increases and $A$ decreases, coverage rates generally increase. Coverages achieve or exceed their nominal 0.95 levels (within simulation error), while for small $A$ and big $B_0$, coverages for the large variance group substantially exceed both their nominal rate and the coverages for the small variance group. The calibrated risks are less than 1.0 in Figure 8 which show that the intervals are wide enough to be conservative, although they may be excessively conservative for the large variance group. One remedy could be using the scale-invariant prior $c = 0.5$, which makes $\sqrt{A}$ flat. Coverages rates for the exact version of SHP were not evaluated for this unequal variance case, and that can be time-consuming for repeated sampling. Simple and fast computing, plus a procedure's transparency, are reasons for finding simple and accurate approximations.

## 5. CONCLUSIONS

Why might a Bayesian or objective Bayesian statistician who has settled on prior distribution $\pi(A)$ on $A$ consider approximating with ADM? There are several reasons, beyond the general observation that any procedure used in an application is an approximation.

1. Speed of convergence is valuable with big data sets, especially if a procedure is to be used repeatedly for model selection and model checking. The approximations here avoid MCMC burn-ins. Speed also makes it feasible to simulate many times, for example, for bootstrapping, or to check a procedure's operating characteristics.
2. Data analysts may need to obtain the same results each time a particular model is re-fit to the same data, which stochastic approximations do not do.
3. MLE methods always will play a central role in statistics. For the model of this paper, ADM maintains the spirit of MLE while making small sample improvements.
4. Using ADM to help fit shrinkage factors extends to multilevel generalized linear models, for example, to fit a Poisson model (Christiansen and Morris, 1997). In such more complicated non-Normal models, MCMC and exact numerical integration may be more difficult or impossible, giving MLE and ADM a greater advantage of ease. Then the frequency properties of ADM can be checked with each data application by simulating or bootstrapping from the fitted multilevel model. However, that will not reveal how well ADM approximates the exact Bayes procedure.

5. Multiplying a likelihood by $A$ before maximizing combines neatly with EM methods as used to find the MLE of $A$ (Dempster, Laird and Rubin, 1977). With ADM, EM would avoid infinite loops that occur when the MLE $\hat{A} = 0$.
6. Data analysts always will need well-checked, prepackaged, documented, widely known and available procedures for fitting models.
7. Statistical software programmers should find it easy to program and adopt the ADM-SHP formulas, for example, the formulas of Section 2.8, in standard software. For example, ADM could be an option in SAS PROC MIXED along with MLE and REML.

Barring prior information that $A$ is likely to be small, the ADM-SHP methods developed here for making inferences, especially interval estimates, about the random effects in a two-level Normal regression model will have better frequency performance over the entire range of $A \geq 0$ than MLE and REML methods. Our derivation has benefited from viewing Stein's harmonic prior SHP on the random effects $\theta_i$ as arising from a uniform mixture over $A$ of the Level-2 Normal distribution (2), that is, according to $\pi(A) = 1$.

With this formal (improper) prior, the posterior density on $A$ agrees with the marginalized likelihood function $L(A)$. That justifies the term "adjustment for likelihood maximization" when "ALM" is restricted to point estimation of a shrinkage factor. The results here go on to use the flat $\pi(A) = 1$ prior and conditional (Bayesian) reasoning as a guide to accounting for variability of the shrinkage factors $B_i$ and ultimately, of the random effects $\theta_i$. ADM approximates the exact Bayes procedures with considerable accuracy, given that it retains the (relative) ease of MLE/REML calculations, that is, by using two derivatives of the adjusted log-likelihood $\log(A\ L(A))$. Of course the adjustment here more generally would adjust by using the multiplier $\pi(A)$ if $\pi(A) \neq 1$. While more testing is needed for unequal variances cases, the confidence intervals for random effects arising from the ADM-SHP combination here thus far have met or exceeded their nominal coverages if $k - r \geq 3$. Still, the search should continue for priors on $A$ that will provide even better frequency interval coverages.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge funding for this project provided in part by NSF Grant DMS-97-05156, and for many helpful suggestions made by the Editors, the Associate Editor and a referee.

## REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions*. *Applied Mathematics Series* **55**. National Bureau of Standards, Washington, DC.

BELL, W. (1999). Accounting for uncertainty about variances in small area estimation. Available at www.census.gov/did/www/saipe/publications/conference.html.

CHRISTIANSEN, C. L. and MORRIS, C. N. (1997). Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.* **92** 618–632. MR1467853

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

EFRON, B. and MORRIS, C. N. (1975). Data analysis using Stein's estimator and it's generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC Press, Boca Raton, FL. MR2027492

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. Probab.* **I** 361–379. Univ. California Press, Berkeley. MR0133191

LI, H. and LAHIRI, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *J. Multivariate Anal.* **101** 882–892. MR2584906

MORRIS, C. N. (1977). Inverval estimation for empirical Bayes generalizations of Stein's estimator. In *Proceedings of the Twenty-Second Conference on the Design of Experiments in Army Research Development and Testing*. ARO Report 77-2.

MORRIS, C. N. (1983a). Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis, and Robustness (Madison, Wis., 1981)*. *Publ. Math. Res. Center Univ. Wisconsin* **48** (G. E. Box, T. Leonard and C.-F. Wu, eds.) 25–50. Academic Press, Orlando, FL. MR0772762

MORRIS, C. N. (1983b). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

MORRIS, C. (1988a). Determining the accuracy of Bayesian empirical Bayes estimates in the familiar exponential families. In *Statistical Decision Theory and Related Topics IV* **1** (*West Lafayette, Ind.*, 1986) (S. Gupta and J. Berger, eds.) 251–263. Springer, New York. MR0927105

MORRIS, C. N. (1988b). Approximating posterior distributions and posterior moments. In *Bayesian Statistics 3* (J.-M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 327–344. Oxford Univ. Press, New York. MR1008054

RASBASH, J., BROWNE, W., GOLDSTEIN, H., YANG, M., PLEWIS, I., HEALY, M., WOODHOUSE, G., DRAPER, D., LANGFORD, I. and LEWIS, T. (2001). *A User's Guide to MLwiN*. *Centre for Multilevel Modelling*. Institute of Education, Univ. London.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

TANG, R. (2002). Fitting and evaluating certain two-level hierarchical models. Ph.D. thesis, Harvard Univ. MR2703395