Statistical Science
2012, Vol. 27, No. 1, 115–134
DOI: 10.1214/11-STS363

Shrinkage Estimation in Multilevel Normal Models

Carl N. Morris and Martin Lysy

Abstract. This review traces the evolution of theory that started when Charles Stein in 1955 [In *Proc. 3rd Berkeley Sympos. Math. Statist. Probab.* **I** (1956) 197–206, Univ. California Press] showed that using each separate sample mean from $k \geq 3$ Normal populations to estimate its own population mean μ_i can be improved upon uniformly for every possible $\mu = (\mu_1, \ldots, \mu_k)'$. The dominating estimators, referred to here as being "Model-I minimax," can be found by shrinking the sample means toward any constant vector. Admissible minimax shrinkage estimators were derived by Stein and others as posterior means based on a random effects model, "Model-II" here, wherein the μ_i values have their own distributions. Section 2 centers on Figure 2, which organizes a wide class of priors on the unknown Level-II hyperparameters that have been proved to yield admissible Model-I minimax shrinkage estimators in the "equal variance case." Putting a flat prior on the Level-II variance is unique in this class for its scale-invariance and for its conjugacy, and it induces Stein's harmonic prior (SHP) on μ_i.

Component estimators with real data, however, often have substantially "unequal variances." While Model-I minimaxity is achievable in such cases, this standard requires estimators to have "reverse shrinkages," as when the large variance component sample means shrink less (not more) than the more accurate ones. Section 3 explains how Model-II provides appropriate shrinkage patterns, and investigates especially estimators determined exactly or approximately from the posterior distributions based on the objective priors that produce Model-I minimaxity in the equal variances case. While correcting the reversed shrinkage defect, Model-II minimaxity can hold for every component. In a real example of hospital profiling data, the SHP prior is shown to provide estimators that are Model-II minimax, and posterior intervals that have adequate Model-II coverage, that is, both conditionally on every possible Level-II hyperparameter and for every individual component $\mu_i, i = 1, \ldots, k$.

Key words and phrases: Hierarchical model, empirical Bayes, unequal variances, Model-II evaluations, Stein's harmonic prior.

1. INTRODUCTION: STEIN AND SHRINKAGE ESTIMATION

Charles Stein [23] stunned the statistical world by showing that estimating k population means $\mu = (\mu_1, \ldots, \mu_k)'$ with their sample means $y = (y_1, \ldots, y_k)'$ is inadmissible. That result assumes $k \geq 3$ independent Normal distributions and a sum of mean squared component errors risk function. With Willard James [14], he provided a specific shrinkage estimator,

Carl N. Morris is Professor of Statistics, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: morris@stat.harvard.edu). Martin Lysy is PhD student, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: lysy@stat.harvard. edu).

the James–Stein estimator, which dominates the sample mean vector very substantially.

This first section introduces the history of the James–Stein minimax estimator and its extensions when "equal variances" prevail, with "Model-I" evaluations that are conditional on μ. However, "Model-I" does not allow certain practical needs to be met, such as valid confidence intervals. Section 2 shows how this has been rectified by enlarging "Model-I" to "Model-II" wherein random effects distributions are assigned in Level-II. The resulting framework enables repeated sampling (frequency based) interval estimates [9, 17] and frees practitioners from determining and specifying valid relative weights for each squared-error component loss, upon which Model-I minimax estimators depend critically. Model-II even supports developing admissible minimax shrinkage estimators via posterior mean calculations by simplifying the specification of prior distributions, proper and otherwise, on the Level-II parameters. The centerpiece of Section 2 is Figure 2, which graphically organizes some priors on the Level-II variance that lead to minimax estimators. Stein's harmonic prior (SHP) in Figure 2 corresponds to an admissible shrinkage estimator that provides acceptable frequency coverage intervals in Model-II evaluations.

Section 3 reviews the unequal variances case that arises regularly in practice, but for which mathematical evaluations are difficult. The previous sections are meant especially to provide the background needed for more research on the operating characteristics in repeated sampling of unequal variances procedures, while Section 3 shows why that is needed. It is shown in Section 3 why, in substantially unequal variances settings, the Model-II random effects framework works well while the Model-I perspective provides inappropriate ("reversed") shrinkage patterns. The SHP prior in that setting leads to estimators and to formal posterior intervals for μ_i, $i = 1, \ldots, k$, that appear to provide approximate (or conservative) frequency confidence intervals with respect to Model-II evaluation standards that for each individual μ_i approximates or exceeds its nominal 95% coverage, no matter what the true Level-II variance.

The data in Table 1 provide an equal variance example based on a 1992 medical profiling evaluation of $k = 10$ New York hospitals. We are to consider these as Normally-distributed indices of successful outcome rates for patients at these 10 hospitals following coronary artery bypass graft (CABG) surgeries. The indices are centered so that the New York statewide average outcome over all hospitals lies near 0. Larger estimates y_i indicate hospitals that performed better for these surgeries. For example, Hospital 10 was more than 2 standard deviations above the statewide mean. All 10 sample means have nearly the same variances, which we have scaled so the common variance is about $V = 1.00$. The variances V_i must be the same in order to meet the equal variance assumption upon which the James–Stein estimator is based. This "equal variance" case enables various mathematical calculations that are difficult, if not impossible, for the widely encountered "unequal variances" situation.

TABLE 1
Hospital profiling data and James–Stein shrinkage estimates for $k = 10$ NY hospitals

i	y_i	sd_i	V_i	$\hat{B}_{\mathrm{JS}}$	$\hat{\mu}_{\mathrm{JS},i}$
1	−2.15	1.0	1.0	0.688	−0.67
2	−0.34	1.0	1.0	0.688	−0.11
3	−0.08	1.0	1.0	0.688	−0.02
4	0.01	1.0	1.0	0.688	0.00
5	0.08	1.0	1.0	0.688	0.02
6	0.57	1.0	1.0	0.688	0.18
7	0.61	1.0	1.0	0.688	0.19
8	0.86	1.0	1.0	0.688	0.27
9	1.11	1.0	1.0	0.688	0.35
10	2.05	1.0	1.0	0.688	0.64

The vector of sample means y has total mean squared error (risk) as an estimator of μ given by

$$E\sum_{i=1}^{k}(y_i - \mu_i)^2 = \sum_{i=1}^{k} V_i = kV.$$

This unbiased estimator is minimax, since its constant risk is the limit of the risks of a sequence of proper Bayes' rules (see, e.g., Theorem 18 of Chapter 5 in [3]).

In the simplest situation, the James–Stein estimator "shrinks" y_i toward an arbitrarily preassigned constant μ_0. It is appropriate to set $\mu_0 = 0$ in this case because we have recentered the CABG indices to have NY statewide mean equal to 0. Then with $\mu_0 = 0$, the sum of squared residuals for these data,

$$S = \sum_{i=1}^{10} y_i^2 / V = 11.62,$$

would have a $\chi^2_{(10)}$ distribution if the hypothesis that all values of $\mu_i \equiv \mu_0 = 0$ were true, thereby failing to reject the null at even the 30% level. However, most members of the medical community would not believe that all hospitals are equally effective, and many in the

statistical community would be reluctant to think that the first and last hospitals in the list, whose quality estimates differ by more than 2 standard deviations from 0, should be declared to have the same underlying quality as all the others.

On the other hand, S isn't far from its expectation $k = 10$ if all the μ_i are 0, and some extreme rates would occur, at least in part, because of randomness. Thus, regression-toward-the-mean (RTTM), that is, shrinkage toward μ_0, would be expected if more data were to appear for these hospitals.

RTTM is anticipated if one believes that there is some similarity among the hospitals, and that sampling variation is part of the reason for the extreme hospitals. That is, the hospital with the highest quality index with $y_{10} = 2.05$ probably has a true mean μ_{10} smaller than 2.05 because

$$E\Big[\max_{1\le i\le k} y_i|\mu\Big] > \max_{1\le i\le k} E[y_i|\mu_i] = \max_{1\le i\le k} \mu_i \ge \mu_{10}$$

(by Jensen's inequality and convexity of the maximum function). So we expect in this case that the observed maximum $y_{10} = 2.05$ exceeds μ_{10}, and a shrunken estimator is in order. The two-level Model-II, soon to be described, anticipates and models RTTM, leading to shrinkage estimation.

Following earlier notation set in a series of papers by Efron and Morris, for example, [9], about Stein's estimator and its generalizations, we denote shrinkage factors by the letter B (often with subscripts). The James–Stein shrinkage coefficient for this setting is calculated as

$$\hat{B}_{\mathrm{JS}} = (k-2)/S,$$

which for these data is $\hat{B}_{\mathrm{JS}} = 8/11.62 = 0.688$. This estimator then shrinks the usual unbiased estimates y_i toward $\mu_0 = 0$ according to

$$\hat{\mu}_{\mathrm{JS},i} = (1-\hat{B}_{\mathrm{JS}})y_i + \hat{B}_{\mathrm{JS}}\mu_0 = (1-\hat{B}_{\mathrm{JS}})y_i.$$

Based on this shrinkage estimate, future observations are being predicted to regress about 68.8% of the way toward 0. Column 5 of Table 1 lists the shrunken values

$$(1-0.688)\times y_i + 0.688\times 0 = 0.312\times y_i$$

for each hospital, the James–Stein estimate of the mean. For example, the estimate of Hospital 10's quality index is reduced from 2.05 standard deviations above the New York mean to 0.64 standard deviations. The RTTM effect is strong for these 10 hospitals, which are estimated to be more similar than different, with only 31.2% of the weight allocated to each hospital's own estimate. Figure 1 illustrates the shrinkage pattern.

The parameter μ_i can be thought of as the quality index that would result for hospital i if that hospital theoretically could have performed a huge number of CABG surgeries in 1992. Whether the JS estimator of quality is a better estimator of μ than y for these data cannot be guaranteed because the true values of μ aren't known. However, one can calculate an unbiased estimator of the expected risk (i.e., for sum of squared errors) of the JS estimator [14]. This unbiased estimator of the risk is

$$\hat{R} = V\big(k-(k-2)\hat{B}_{\mathrm{JS}}\big),$$

a function of y only through S. That y is inadmissible and that the JS estimate is "minimax" (risk never exceeds kV) follows because $\hat{B}_{\mathrm{JS}} > 0$ for all data sets. This proves minimaxity, that the risk of $\hat{\mu}_{\mathrm{JS}} = (\hat{\mu}_{\mathrm{JS},1},\ldots,\hat{\mu}_{\mathrm{JS},k})'$ as a function of μ is

$$E[\hat{R}|\mu] = V\big(k-(k-2)E\hat{B}_{\mathrm{JS}}\big) < kV.$$

For these data, $\hat{R} = 1.00\times(10-8\times 0.688) = 4.496$. This is a large reduction in mean squared error, less than half of $kV = 10$, the risk of the separate unshrunken estimates y_i. In fact, the smallest possible

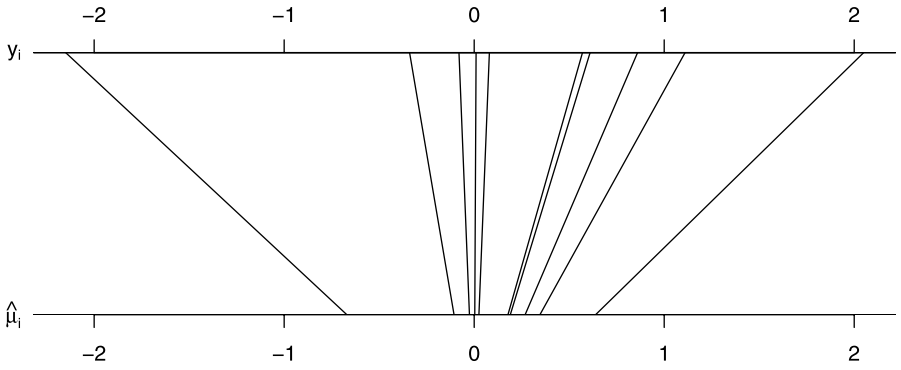

FIG. 1. *Unbiased (top) versus James–Stein (bottom) estimates for* 10 *NY hospitals.*

value of the risk for the JS estimator is $2V$, when $\mu = 0$, for any value of $k \geq 3$, thus offering very substantial possible improvements on y.

The JS estimator can be extended in the equal variance setting to cover more general situations. For example, as Stein and others showed, one can shrink the y_i toward the grand mean of the data, $\overline{y} = \sum y_i/k$. With these hospital data this would shrink toward $\overline{y} = 0.272$ (which differs by less than one standard error of the overall average for the 10 hospitals from the assumed mean 0). More generally, if along with each y_i one collects a vector of $r > 0$ covariate vectors x_i (possibly including the intercept), each y_i can be shrunk toward its regression prediction $x_i'b$, where

$$b = (X'X)^{-1}X'y$$

and X is the $k \times r$ covariate matrix with columns x_i, $i = 1, \ldots, k$. Doing this forfeits r degrees of freedom, so that

$$\hat{B}_{\mathrm{JS}} = (k - r - 2)/S,$$

with S now replaced by

$$S = \sum_{i=1}^{k} (y_i - x_i'b)^2/V.$$

The James–Stein estimates of the μ_i then become

$$\begin{aligned}\hat{\mu}_{\mathrm{JS},i} &= (1 - \hat{B}_{\mathrm{JS}})y_i + \hat{B}_{\mathrm{JS}}x_i'b \\ &= (1 - \hat{B}_{\mathrm{JS}})(y_i - x_i'b) + x_i'b.\end{aligned}$$

Writing $\hat{\mu}_{\mathrm{JS},i}$ this way suggests that shrinking with $r > 0$ does not affect the r-dimensional regression space, but only shrinks toward 0 in the $k - r$ dimensional space orthogonal to it. Indeed, the problem can be "rotated" to an equivalent one in which the last r values of the residuals $y_i - x_i'b$ are all equal to 0, regardless of the value of y, for example, Stein [24]. The example just considered, with shrinkage toward zero, shows what happens to the residuals when shrinkage is toward a regression model.

Of course V needn't be 1.00, or even be known, provided there exists an independent Chi-square estimate of V. While that can be handled straightforwardly in the equal variance case [24], it will not be a central issue in any case if the degrees of freedom are substantial.

Using the JS estimator seems easy and powerful, but many complicating issues arise in practice:

1. What is the standard error of each individual estimate? One hopes the JS estimator for Hospital 10 improves $y_{10} = 2.05$ (with standard deviation = 1.00) by using the better estimate $\hat{\mu}_{10} = 0.64$. The sum of individual risks has decreased from 10 to 4.5 for all 10 hospitals, but this does not mean the variance for each individual estimate has dropped to 0.45. Furthermore, the JS estimator cannot even guarantee that every component (hospital) has a smaller risk (expected squared error) as a function of μ. Such an improvement is impossible because each individual y_i is an admissible estimate of its own μ_i, in one dimension. Rather, minimaxity of the JS estimator for sum of squared errors is accomplished by "balancing" or "trading off" component risks. Components with mean square errors that exceed V are guaranteed to have their risks more than offset by risk improvements on the remaining components. The minimaxity claim (improvement on the unshrunken vector of unbiased estimates) is for aggregate risk, and not for every component.
2. Why, even in this equal variance case, should the loss function be an unweighted sum of squares? In applications the loss function could require different relative weights to reflect unequal economic loss for the mean squared errors of different components (hospitals, here). That is, the appropriate loss function could be

$$L(\hat{\mu}, \mu) = \sum_{i=1}^{k} W_i(\mu_i - \hat{\mu}_i)^2$$

for some appropriate weights $W_1, \ldots, W_k > 0$.

Users of the James–Stein estimator typically assume that all W_i are equal in assessing its risk benefits. But would NY hospital administrators agree that hospital errors can be traded off with equal weights? Perhaps weights should differ for teaching hospitals, or for military hospitals, or for children's or other specialty hospitals, or for hospitals in areas far from medical centers, or for large hospitals. Getting agreement on that issue has arisen with various real shrinkage applications. Even if the administrators could agree on the values of the W_i, the James–Stein estimator would not dominate y when the W_i are sufficiently unequal. There is a way out that seems reassuring, at first, because a shrinkage estimator can be found to dominate y for any given weights W_i. But there is a rub. The dominating estimator for a set of weights depends on the specified

weights, and then it cannot be expected to dominate y for a different set of weights. Only the unshrunken estimator y can be guaranteed to be minimax independently of the weights W_i. Its risk, the minimax risk, is $V = \sum W_i$. More on this in Section 3.

3. Even with equal weights, $W_i \equiv 1$, another problem arising in practice and in the theory is that $\hat{B}_{\mathrm{JS}}$ can exceed 1. A (uniformly) better shrinkage constant uses $\min(1, \hat{B}_{\mathrm{JS}})$ instead and easily is seen to reduce the total risk. That change necessitates developing a new unbiased estimator of risk. This was made possible, and easy, by a simple calculus pioneered by Stein [25, 27] and independently by Berger's integration by parts technique [2].

   This truncated shrinkage estimator's improvement shows that the James–Stein estimator is inadmissible itself. The improved truncated estimator also is inadmissible, as it has a discontinuous derivative, while admissible estimators must have all their derivatives (as a function of the data). The search for admissible estimators began soon after the James–Stein estimator, for example, Stein [14] and Brown [4].

4. We already have noted that there is no agreed-upon way to estimate the component variances of the JS estimator. Correspondingly, there is no way to determine separate confidence intervals for each μ_i. Confidence ellipsoids, for example, Stein [26] and Brown [5], can be and have been developed for the equal variance setting. However, ellipsoids may be unattractive to a data analyst who has the alternative of estimating with y_i and using $V^{1/2}$ as the standard error, with a corresponding exact confidence interval for each component obtained via the Normal distribution. Unfortunately, only aggregates (ellipsoidal sets in this context) can provide uniformly better coverage if coverage must hold conditionally on the underlying μ for all μ, that is, with Model-I evaluations. There is no agreed upon componentwise procedure for standard errors and intervals for individual components μ_i simply because no such procedure is possible as a function of μ. This problem (and others too) can be rectified only via acceptance of a two-level, random effects model referred to here as Model-II.

5. The overriding difficulty for the JS estimator as a practical tool for data analysts is that, except for data produced by carefully designed experiments, real data rarely occur with equal variances $V_i = V$. Even the hospital data of Table 1 do not have exactly the same variances. The first author has participated in developing and in using shrinkage techniques for hospital profiling and for other applications (e.g., [7, 17]) without ever seeing hospital or medical data with equal variances, simply because hospital caseloads (numbers of patients) vary considerably. For this initial discussion to illustrate the JS estimator and related shrinkage procedures in the equal variances setting, we have picked 10 of the 31 hospitals (the 31 to be described later) that had similar variances. These 10 each have sample sizes within 15% of 550 patients.

2. THEORETICAL AND BAYESIAN DEVELOPMENTS FOR THE EQUAL VARIANCE CASE

This section reviews expansion of the assumptions of "Model-I" to a two-level model, "Model-II," which at Level-II includes a random effects model on the μ_i, with the Level-II parameters unknown but estimable from the data. Model-II and Stein's harmonic prior (SHP), to be introduced in this section, will be especially important as a basis for developing frequency procedures in the difficult unequal variances situation of Section 3. After briefly introducing the unequal variances case in this section, the equal variances setting is studied because of its relatively easy calculations. This enables Bayesian analysis that uses formal priors on the Level-II parameters that produce shrinkage estimators as posterior means. In the equal variance setting, many of these estimators have been proven to be minimax (some also are admissible) in the original Model-I sense of Stein, that is, for total square error loss and for every possible mean vector μ. The centerpiece of this section is Figure 2, which displays graphically certain famous distributions on the Level-II variance, "A" that are known to provide minimax shrinkage estimators. Of central importance is Stein's harmonic prior (SHP) on μ, which stems from imposing an improper flat prior on A and yields an admissible, minimax modification of the James–Stein estimator. This SHP shrinkage estimator leads to posterior interval estimates that meet confidence requirements for coverages in Model-II evaluations.

A generalization of the James–Stein estimator almost always is required in practice because the unequal variance situation arises, and also because data analysts often must provide interval estimates. Uniform risk dominance as a function of μ will be seen in

TABLE 2
Multilevel model layout

Level	Descriptive version	Inferential version
I	$y_i \mid \mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, V_i),\ i = 1, \ldots, k$	$y_i \mid \alpha \overset{\text{ind}}{\sim} \mathcal{N}(x_i'\beta, V_i + A)$
II	$\mu_i \mid \alpha = (\beta, A) \overset{\text{ind}}{\sim} \mathcal{N}(x_i'\beta, A)$	$\mu_i \mid y, \alpha \overset{\text{ind}}{\sim} \mathcal{N}((1 - B_i)y_i + B_i x_i'\beta, V_i(1 - B_i))$
III	$\alpha \sim \pi(\alpha)$	

Section 3 to require inappropriate (reversed) shrinkage patterns in practice. Of course, shrinkage methods are used commonly in applications, almost always being based on a two-level random effects model, the μ_i being random effects with their own distributions. Such models belong to frequentists and Bayesians alike, known as hierarchical models, multilevel models, empirical Bayes models and by other terms. Table 2 shows one such model with Normally distributed observations (Level-I), and Normally distributed random effects (Level-II). The two columns, that is, the Descriptive and the Inferential versions of the model, are equivalent in that both sides give rise to the same joint distributions of the data and the random effects, (y, μ), given the hyperparameter α that governs the joint distribution. These models allow "unequal variances" V_i, perhaps because $V_i = \sigma^2/n_i$ with different sample sizes. That anticipates Section 3, but in this "equal variances" Section we always assume $V_i \equiv V$.

In what follows, Model-I will refer to the distribution of $y|\mu$ at Level-I of Table 2 which treats μ as the unknown parameter, whereas Model-II will refer to the random effects model combining Model-I and the Level-II distribution of $\mu|\alpha$, which has unknown parameter $\alpha = (\beta, A)$. Model-III will refer to the fully Bayesian model embracing all Levels I, II and III for a single prior $\pi(\alpha)$ on α, and is used here primarily to construct Bayes rules to be evaluated via the assumptions of Model-I or Model-II, in the frequency sense for all y.

If the hyperparameter $\alpha = (\beta, A)$ were known in Model-II of Table 2, one would use the Level-II distribution of $\mu|y, \alpha$ in Table 2 to make inferences about each component value μ_i. For squared error loss the best estimator of μ_i then would be the posterior mean, which estimates μ_i by using the shrinkage factor

$$B_i = \frac{V_i}{V_i + A}$$

to compromise between the prior mean $x_i'\beta$ and the sample mean y_i.

While shrinkages needn't arise for many distributions that one could choose for Level-II, they do with the Normal distribution on μ_i because the Normal distribution on μ in Level-II is conjugate to the Normal Level-I likelihood. Conjugate priors at Level-II lead to linear posterior means and shrinkage coefficients for the Normal and for other exponential family models too; see, for example, Diaconis and Ylvisaker [8] and Morris and Lock [21]. They also are the "G_2 minimax" choice for Level-II [13, 18].

With $k > r + 2$ components, it is not required to assume $\alpha = (\beta, A)$ is known because information builds up through the k observations y_i, whose distributions are governed by their shared dependence on (β, A) via the likelihood function given by the right half of Level-I in Table 2. For the rest of this section we focus on the simplest case of the Table 2 model with $\beta = 0$ $(r = 0)$. Thus, $\alpha = A$ is the only unknown hyperparameter. With equal variances, studying the case $\beta = 0$ is much less restrictive than it might seem because use of the orthogonality trick described in Section 1 allows developments for $\beta = 0$ to be extended back to the case with β unknown.

Early work on the equal variance case strongly emphasized Model-I squared error evaluations made conditionally on μ. Even so, it was realized, for example, Stein [24], that if one also assumes Model-II, then it is easy to motivate shrinkage estimators and the JS estimator, since one can estimate A by considering the likelihood of A, or, equivalently, of $B_i \equiv B$. The likelihood of B follows from the marginal distribution of $y|B$ in the inferential column of Table 2 which has the form of a Gamma density, but conditioned on $B \leq 1$,

$$L(B) = B^{k/2} \exp(-BS/2).$$

Because of the equal variance assumption, $L(B)$ only depends on the 1-dimensional sufficient statistic for B in the model for $y|A$:

$$S = \sum_{i=1}^{k} y_i^2 / V.$$

The maximum likelihood estimate of B is $\hat{B} = k/S$. However, B (not A) enters linearly in $E[\mu_i|y]$, and by noting that

$$S|B \sim B^{-1}\chi^2_{(k)},$$

one sees that the James–Stein shrinkage estimate $\hat{B}_{\mathrm{JS}} = (k-2)/S$ is the best unbiased estimate of B. Both of these estimates lead to shrinkage or "empirical Bayes" estimators of μ_i via substituting k/S or $(k-2)/S$ for the shrinkage B, where B appears in

$$E[\mu_i|y, B] = (1-B)y_i.$$

Minimaxity of these and of other shrinkage estimators can be checked via Baranchik's minimax theorem, from his 1964 dissertation [1] under Stein. Assume the equal variance Normal setting of Table 2, $r = 0$, $k \geq 3$, and Model-I only. Suppose an estimator shrinks its k components toward 0 based on a shrinkage factor of the form

$$\hat{B}(S) = u(S)/S,$$

with $u(S)$ nondecreasing and with $0 \leq u(S) \leq 2(k-2)$. Then the estimator is minimax for total mean squared error risk under Model-I, that is, with risk at most kV for all μ. A similar but more general condition for minimaxity that lets $u(S)$ be decreasing also exists [10]. These minimaxity conditions easily extend to include shrinkage toward a fitted $r > 0$ dimensional subspace by making S be the residual sum of squares and by accounting for the loss of r degrees of freedom, so then $0 \leq u(S) \leq 2(k-r-2)$ is required.

2.1 Bayes and Formal Bayes Rules

The model of Table 2 can be expanded to Level-III to allow Bayesian and formal Bayesian inferences by assuming that α in general (in our simplified context, the unknown variance parameter A) has a proper or improper prior distribution. Shrinkage factors are then determined as integrals over the posterior distribution of B,

$$E[B|S] = \frac{\int_0^1 BL(B)\pi(B)\,\mathrm{d}B}{\int_0^1 L(B)\pi(B)\,\mathrm{d}B}$$

for some prior density π on B.

Two obvious families of priors arise in this context, to be charted in Figure 2:

1. *Scale-invariant priors on A.* Indexed by constants $c \geq 0$, these are improper (i.e., not finitely integrable) formal priors, with differential elements
$$A^{c/2}\,\mathrm{d}A/A, \quad A > 0.$$
As a distribution on B, this corresponds to
$$B^{-c/2-1}(1-B)^{c/2-1}\,\mathrm{d}B, \quad 0 < B < 1.$$
These have the form of Beta densities, but they do not integrate finitely. Only propriety of the posterior distribution is required, that is, after multiplication by $L(B)$, which imposes the additional restriction $0 < c < k$.
2. *Conjugate priors on B* take the form of the likelihood function $L(B)$, but with different values of k, S. We index this conjugate family by $k_0 > 2$ and by $S_0 \geq 0$, perhaps thinking of them as previous values of k and S. Posterior propriety now requires that k_0 satisfy $k_0 + k > 0$. The prior and posterior densities take the same form as $L(B)$, having differential element
$$B^{(k_0-2)/2}\exp(-BS_0/2)\,\mathrm{d}B/B, \quad 0 < B < 1.$$
If $S_0 > 0$, these are "truncated" $\chi^2_{(k_0-2)}$ distributions on $B \leq 1$, scaled by S_0. This second family involves proper priors if $k_0 > 2$, known as "Strawderman's priors" [28] when $S_0 = 0$. Strawderman showed (via Baranchik's theorem) that the posterior mean of B for these priors provides minimax and admissible shrinkage estimators if $k_0 \leq k-2$ (so $k \geq 5$ is required). These properties also hold if $S_0 > 0$. When $S_0 = 0$ B has a Beta$((k_0-2)/2, 1)$ distribution and
$$EB = (k_0-2)/k_0$$
a priori, again requiring $k_0 > 2$ for propriety. $EB \leq (k-4)/(k-2)$ is the upper limit for minimaxity, requiring $k \geq 5$. The special choice $k_0 = 4$ puts a Uniform$(0, 1)$ prior distribution on B and minimaxity then requires $k \geq 6$. Derived from proper priors, the posterior mean of μ, given the data y for any of these Strawderman priors, automatically qualifies as an admissible, minimax estimator in the Model-I sense for quadratic loss.

The densities of these two prior families can be combined by multiplication (and some reparametrization) to yield a 3-parameter family with densities on B of the form

$$\begin{aligned} &p(B|k_0, c, S_0) \\ &\quad \propto B^{(k_0-c)/2-1}(1-B)^{c/2-1} \\ &\qquad \cdot \exp(-BS_0/2)\,\mathrm{d}B, \quad 0 < B < 1. \end{aligned} \tag{1}$$

If $S_0 = 0$, this class of prior densities has the form

$$\mathrm{Beta}\left(\tfrac{1}{2}(u-k), \tfrac{1}{2}(k+k_0-u)\right) \tag{2}$$

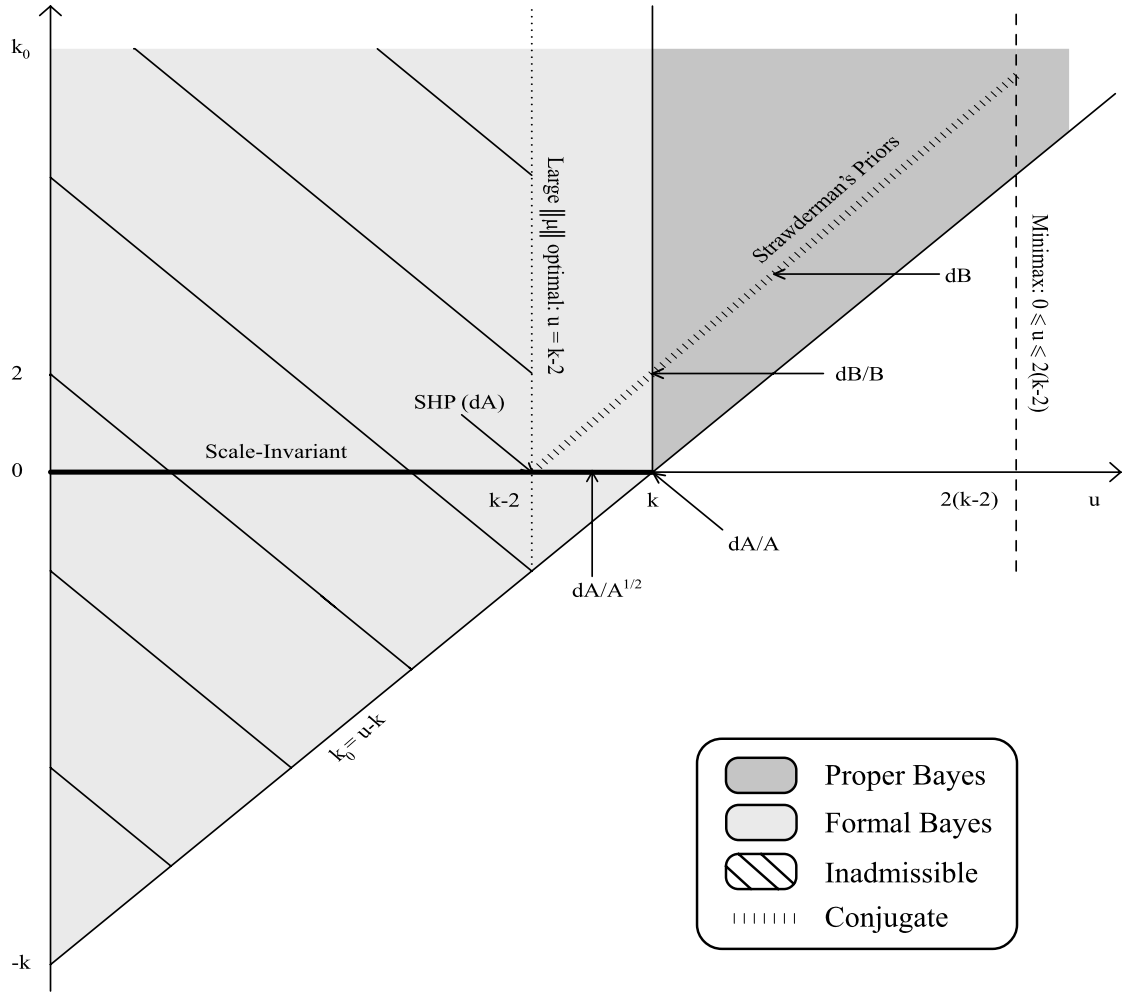

FIG. 2. *Classification of the proper and formal priors of the form given in (2). The u-axis determines limits on minimaxity, with smaller values providing less shrinkage. Larger k_0 indicates more prior information, and thus more shrinkage.*

with $u = k + k_0 - c$. They are proper only if $k_0 > c > 0$, that is, if $k_0 > u - k > 0$. The posterior density is proper if and only if $k_0 + k > c > 0$ since the exponential term that also appears in the posterior density, $\exp(-B(S + S_0)/2)$, is bounded in B so that term cannot affect posterior propriety.

Figure 2 shows the key regions for this formal Beta family (2) of prior densities (scaled for $k = 10$) in terms of the two parameters (u, k_0), ignoring the nearly irrelevant $S_0 = 0$. It emphasizes regions when minimaxity holds, $0 \le u \le 2(k-2)$. Instead of c, the horizontal axis uses $u = k + k_0 - c$, because u determines minimaxity. It can be seen that as $S \to \infty$, $S \times E[B|S] \to u$ for these priors, and Baranchik's theorem tells us that minimaxity for large S fails unless $0 \le u \le 2(k-2)$. This condition is necessary for minimaxity, but not sufficient.

Some explanation is in order, as follows in (a) through (h):

(a) Priors on $B = V/(V + A)$ that lead to minimax estimators are limited to $0 \le u \le 2(k-2)$.

(b) The posterior distribution is proper only if $k_0 > u - k$. The 45 degree line $k_0 = u - k$ in Figure 2 marks the (unattainable) lower bound for these priors.

(c) Proper priors require $k_0 > c$, so that $u > k$. Proper priors lie in the darkly shaded region to the right of the vertical line $u = k$. Improper priors are those with $u \le k$.

(d) The scale-invariant priors $A^{c/2}\,dA/A$ on $A > 0$ are on the horizontal axis $k_0 = 0$, so $c = k - u$ in these priors. Posterior propriety for these priors, as seen in (b), requires $0 \le u \le k$ so that shrinkage cannot extend all the way to the Baranchik limit $2(k-2)$. Scale invariant priors cannot be proper.

Viewed as distributions on μ, these scale-invariant priors have differential elements

$$\mathrm{d}\mu/\|\mu\|^u$$

[by integrating the $\mathcal{N}(0, AI_k)$ density with respect to $A^{c/2}\,\mathrm{d}A/A$, and using $c = k - u$]. If $u = 0$, that is, the prior located in Figure 2 at $(u, k_0) = (0, 0)$, we have the k-dimensional Lebesgue measure $\mathrm{d}\mu$, which leads to using y as the estimator of μ, that is, no shrinkage.

One never should use $(u, k_0) = (k, 0)$, although researchers sometimes make this mistake, thinking that the prior is vague because this is Jeffreys' form $\mathrm{d}A/A$ in other contexts. Actually, this prior forces $B = 1$ a posteriori, no matter what the magnitude of S might be. Obviously this full-shrinkage estimator cannot be minimax.

(e) The conjugate priors $B^{(k_0-4)/2}\,\mathrm{d}B$ on $0 < B < 1$ (setting $S_0 = 0$) form the upsloping line $k_0 = u - (k - 2)$. These have proper posteriors because this line lies above (and is parallel to) the line $k_0 = u - k$. They are proper if $u > k$, that is, $k_0 > 2$, being Strawderman's priors.

All these conjugate priors produce an easily calculated shrinkage factor (u need not be an integer in the Chi-squares) in this equal variances setting:

$$\hat{B} = E[B|S] = \frac{u}{S} \times \frac{P[\chi^2_{(u+2)} \le S + S_0]}{P[\chi^2_{(u)} \le S + S_0]}. \tag{3}$$

In this expression, $S\hat{B}$ is monotone increasing in S because the ratio of $\chi^2_{(u+2)}$ and $\chi^2_{(u)}$ densities is monotone increasing. Therefore, Baranchik's theorem applies and verifies minimaxity.

(f) The vertical line at $u = k - 2$ denotes priors that have the smallest Model-I risks as $\|\mu\| \to \infty$. This holds because all priors in Figure 2 have shrinkages $E[B|S]$ near to $\hat{B} = u/S$ for large S, and this must occur when $\|\mu\|$ is large.

On the other hand, the mean-squared-error risk for shrinkage estimators of the form $u/S = a\hat{B}_{\mathrm{JS}}$, with $a = u/(k - 2)$ for any $0 \le a \le 2$, is

$$E\left[\sum_{i=1}^{k}\big((1 - u/S)y_i - \mu_i\big)^2 \Big| \mu\right] = k - (k - 2)a(2 - a)E[B_{\mathrm{JS}}].$$

This risk is minimized uniformly at $a = 1$, showing that the James–Stein estimator is optimal among estimators of the form u/S. Combining these two facts shows that minimax priors with $u = k - 2$ lead to estimators with risk functions that, for large $\|\mu\|$, will be smaller than those in Figure 2 with $u \ne k - 2$.

(g) Admissibility of the resulting Bayes estimators of μ holds immediately for proper priors, so the priors in the rightmost wedge with $k < u \le 2(k - 2)$ provide admissible minimax estimators.

Improper priors may or may not produce admissible estimators. Various estimators based on priors with $k - 2 \le u \le k$ are admissible and minimax at least if k_0 isn't too small. The SHP prior, which corresponds to $(u, k_0) = (k - 2, 0)$, $\mathrm{d}A$ is an improper prior that does yield an admissible estimator. More on this later.

(h) Inadmissibility holds for many (perhaps all) of the priors with $u < k - 2$. That this holds is suggested by the fact that the risk of an estimator with $u < k - 2$ can be lowered for large $\|\mu\|$ by using a prior with $u = k - 2$ [as argued in (f) above]. Then it seems likely that such a prior can be found on the $u = k - 2$ vertical axis of Figure 2 that would increase shrinkage (shrinkage generally increases in the rightward direction on Figure 2) with lower risk everywhere as a function of $\|\mu\|$.

Early after it was recognized that the estimator y could be uniformly improved upon, numerous authors proposed priors captured by Figure 2, motivated by Bayesian and/or admissibility concerns. Many of these were scale-invariant priors with $k_0 = 0$, especially with $k - 2 \le u \le k - 1$, for example, Stein, K. Alam, T. Leonard, I. J. Good and D. Wallace, D. Rubin, D. V. Lindley and A. F. M. Smith. Others were proposed on the conjugacy line $k_0 = u - (k - 2)$, including $\mathrm{d}B/B$, that is, $(u, k_0) = (k - 1, 1)$, which has Jeffreys' form, and (being improper) falls at the edge of Strawderman's priors. Various authors since have repeated these and other suggestions, partly as "reference priors." Our hope is that these priors that decision theory has shown to lead to the best and most trustworthy estimators for the equal variances setting of Figure 2 are "transportable" to the unequal variances setting.

Charles Stein's choice is a prior on μ, not on A, "Stein's harmonic prior," SHP, corresponds to μ having a measure that stems from $c = 2$, A with a flat density. It is

$$p(\mu)\,\mathrm{d}\mu \propto \mathrm{d}\mu/\|\mu\|^{k-2}.$$

By (d), this corresponds to $(u, k_0) = (k - 2, 0)$ in Figure 2. The term "harmonic" refers to the fact that the Laplacian of the prior $\nabla^2 p(\mu)$ is uniformly equal to 0, except at the origin where it fails to exist. Technically, since $\nabla^2 p(0) = -\infty$, the prior is actually superharmonic (Laplacian less than or equal to 0), a term Stein himself employed when showing that the re-

sulting Bayes rule was both admissible and minimax by Model-I standards [27]. However, the term "harmonic" is simpler, nearly correct, and used by most researchers.

One motivation for the SHP prior stems from an easy calculation that shows the James–Stein shrinkage coefficient satisfies $E[B|S] = (k-2)/S = \hat{B}_{\mathrm{JS}}$ if one assumes the (absurd) prior that $A \sim \mathrm{Uniform}[-V, \infty)$ [19]. Of course, allowing $A < 0$ is illogical, and removing that part of the support for A gives $A \sim \mathrm{Uniform}[0, \infty)$, which yields the SHP.

A second motivation is that taking A uniform on $(0, \infty)$ lies uniquely in Figure 2 at the intersection of the scale-invariant priors ($k_0 = 0$) and the sloped line of conjugate priors [$k_0 = u - (k-2)$]. That is Stein's SHP. Indeed, the SHP sits on the "admissible boundary," being the scale-invariant admissible prior that shrinks least among the admissible ones. It is also optimal as $\|\mu\| \to \infty$ ($u = k-2$). Being formal Bayes but not proper Bayes, it provides little prior information about A. Its conjugacy makes its shrinkages easy to compute in the equal variance setting.

A third motivation, as will be seen, is that the aggregate conditional posterior risk $R^* = R^*(S) < kV$ for this prior, and, in turn, R^* exceeds the unbiased estimate of the aggregate risk $\hat{R}(S)$, not shown, on the SHP estimator; see Morris [16, 19]. More on this momentarily.

Using (3), the posterior mean of B resulting from the SHP prior is

$$\hat{B}_{\mathrm{SHP}} = E[B|y] = \frac{k-2}{S} \times \frac{P[\chi^2_{(k)} \le S]}{P[\chi^2_{(k-2)} \le S]}.$$

The posterior variance of B [16, 19] is, for $k \ge 3$,

$$v = \mathrm{var}(B|y) = \frac{2}{k-2}\hat{B}^2_{\mathrm{SPH}} - (\hat{B}_{\mathrm{JS}} - \hat{B}_{\mathrm{SHP}})\left(1 - \frac{k}{k-2}\hat{B}_{\mathrm{SHP}}\right).$$

For the $k = 10$ hospitals we have $\hat{B}_{\mathrm{SHP}} = 0.668 \times 0.829 = 0.571$ and $v = (0.218)^2$.

From the SHP posterior mean $\hat{B}_{\mathrm{SHP}}$ we obtain the formal Bayes rule of μ_i,

$$\hat{\mu}_{\mathrm{SHP},i} = E[\mu_i|y] = (1 - \hat{B}_{\mathrm{SHP}})y_i.$$

But what of interval estimates for μ_i? Our Model-III construction via SHP suggests use of posterior probability intervals. For the SHP these can easily be approximated after computing the posterior variance of μ_i, which for $r = 0$ is

$$s_i^2 = \mathrm{var}(\mu_i|y) = V(1 - \hat{B}_{\mathrm{SHP}}) + vy_i^2.$$

Figure 3 from Morris and Tang [22] shows coverage rates of μ_i for 2-sided intervals with nominal coverage 95%. Each interval is centered at its SHP shrinkage estimate and approximates each of the k posterior distributions as Normally distributed with interval widths determined by adding and subtracting $1.96s_i$.

The true coverages in Figure 3 for SHP are never less than 94.5% for any value of A for any of the three values of $k = 4, 10, 20$ shown. The coverage probabilities do not depend on i or on V, so this is tantamount to a proof that this procedure comes close to providing or exceeding the nominal coverage. Over-coverages rise noticeably above 95% as the between-groups variance A approaches 0, that is, as the shrinkage B approaches 1. One must keep in mind, however, that while these intervals are based on the

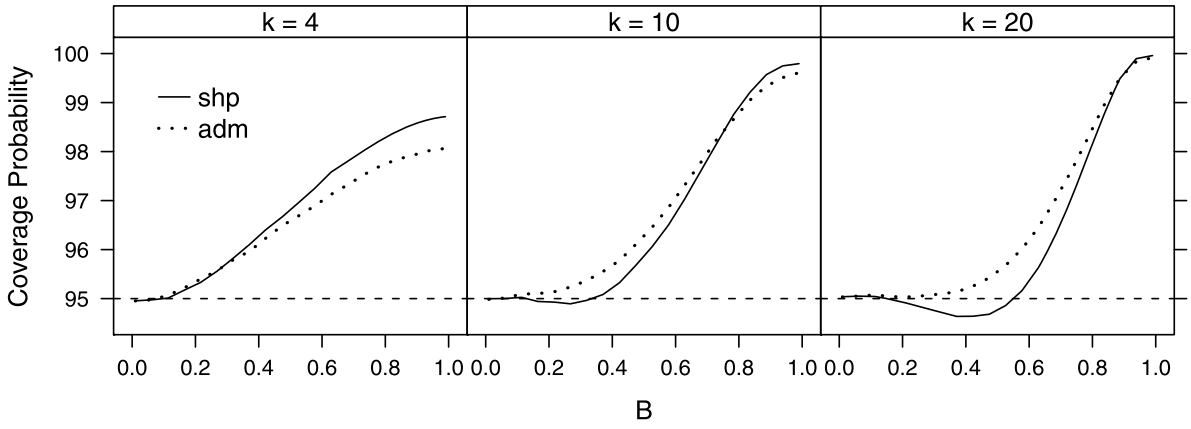

FIG. 3. *Exact coverage probabilities against true shrinkage factor $B = V/(V + A)$ for two equal variances rules, SHP (dark curve) and the ADM approximation to SHP (dotted curve), with nominal 95% coverages, for $r = 0$ and $k = 4, 10, 20$.*

posterior mean and variance, they are not the posterior intervals because such are not symmetric. For example, the under-coverage by 0.5% for SHP when $k = 20$ does not account for the posterior skewness of the distributions of μ_i, which is considerable when y_i is one of the extreme observations.

Intervals based on a different estimator, determined by an approximation technique called adjustment for density maximization (ADM) [20, 22], also are shown in Figure 3, having slightly better minimum coverage. These estimators are described in the next section.

Componentwise intervals better than those centered at y_i, that is, intervals that average being shorter than $2 \times 1.96\text{sd}_i$, do not exist for all μ, by Model-I standards. Such may exist for all A, when averaging over $\mu|A$ in Model-II. Indeed, note that s_i^2 also can be interpreted as the Bayes risk of the SHP rule $\hat{\mu}_{\text{SHP},i}$,

$$R_i^* = E[(\hat{\mu}_{\text{SHP},i} - \mu_i)^2|y] = s_i^2.$$

Let us contrast this with

$$\hat{R}_i = V(1 - 2\hat{B}) + y_i^2(\hat{B}^2 + 2v),$$

the unique unbiased estimate of the component risk of $\hat{\mu}_{\text{SHP},i}$. That is,

$$E\hat{R}_i = E[(\hat{\mu}_{\text{SHP},i} - \mu_i)^2|\mu]$$

is the Model-I component risk for any value of μ. Letting $R^* = \sum_{i=1}^k s_i^2$ and $\hat{R} = \sum_{i=1}^k \hat{R}_i$, by rearranging terms [16, 17] one sees that

$$\hat{R} < R^* < kV.$$

That $R^* < kV$ shows Model-I minimaxity of $\hat{\mu}_{\text{SHP}}$, since its risk is less than that of the minimax y. That $\hat{R} < R^*$ shows that the SHP prior is so vague that its Bayes risk is more conservative than its frequency-based unbiased estimate of risk. Averaging over both μ and y, the k componentwise risks

$$E[R_i^*|A] = E[(\hat{\mu}_i - \mu_i)^2|A]$$

are all the same. Thus, each is less than V for all $A \geq 0$. This establishes Model-II componentwise minimaxity, that is, improvement on y_i for all $A \geq 0$ and for every $i = 1, \ldots, k$.

Not only is the SHP rule componentwise minimax under Model-II evaluations, but its (approximate) coverage intervals are shorter on average than those accompanying the unbiased estimate y (since $Es_i < \sqrt{V}$ by Jensen's inequality). However, values of y exist for which some $s_i^2 > V$, although this happens with small probability.

TABLE 3
SHP estimates and posterior standard deviations of indices of success rates in the 10 *NY hospitals, and two estimates of the associated risk*

y_i	$\hat{\mu}_{\text{SHP},i}$	s_i	$R_i^* = s_i^2$	$\hat{R}_i$
−2.15	−0.92	0.81	0.649	1.803
−0.34	−0.15	0.66	0.435	−0.092
−0.08	−0.03	0.66	0.430	−0.138
0.01	0.00	0.66	0.429	−0.141
0.08	0.03	0.66	0.430	−0.138
0.57	0.24	0.67	0.445	−0.004
0.61	0.26	0.67	0.447	0.015
0.86	0.37	0.68	0.465	0.170
1.11	0.48	0.70	0.488	0.377
2.05	0.88	0.79	0.629	1.627

For the 10 hospitals we obtain $R^* = 4.85$ and $\hat{R} = 3.48$. Componentwise risks and other calculations are displayed in Table 3. Notice that some components have negative unbiased estimates of their mean-square-error, a not uncommon occurrence, and an undesirable feature of using this unbiased estimation approach for assessing component risks.

Unfortunately, real data rarely come with equal variances, designed experiments being the exception. Decision theorists have focused on this symmetric case because it is simple enough to enable exact (small sample) calculations. Decision theory has identified the SHP and other priors close to it that lead to shrinkage estimators with good frequency properties. Now the hope is that such priors are "transportable" to the unequal variances situation.

It should be clear that Model-I verifications are rarely appropriate for scientific applications, even when equal variances obtain. Acceptance of Model-II, and thus of evaluations that average over Level-II distributions (given the hyperparameters, e.g., A), has many advantages for applications. It makes assessing weights for the loss function become unimportant. Model-II allows estimators to exist that are minimax for every component for all A, not just when summed over all components. Confidence intervals exist that are on average shorter than standard intervals centered at y_i, and these also can have (nearly) uniformly higher coverages. The unequal variance setting gives further impetus to Model-II as a basis for evaluating the operating characteristics of shrinkage procedures, and also for constructing them from proper or improper priors that lead to good repeated sampling properties.

3. APPROACHES TO UNEQUAL VARIANCE DATA

In practice, equal variances are the exception rather than the rule. The variances for all 31 NY hospitals, not just the middle 10, differ by a factor of more than 20. Table 4 lists the data for these $k = 31$ NY hospitals and several shrinkage-related estimators, to be discussed further. The raw data contain the number of deaths d_i within a month of CABG surgeries for each hospital i, sorted by increasing caseload n_i. The indices for success rates are calculated as

$$y_i = C \times \big(\arcsin(1 - 2d_i/n_i) - \arcsin(1 - 2\overline{d}/\overline{n})\big),$$

a variance stabilizing transformation of the unbiased success rate estimates $\hat{p}_i = d_i/n_i$, assuming Binomial data, in which case the variance of the y_i is approximately $V_i = \overline{n}/n_i$ (with $\overline{n} = \frac{1}{k}\sum_{i=1}^{k} n_i$). The factor C is chosen so that the harmonic mean of the V_i, that is,

$$V_H = \frac{k}{\sum_{i=1}^{k} V_i^{-1}},$$

is equal to 1. Larger values of y_i correspond to higher success rates. The 10 hospitals used in the previous sections appear here as Hospitals 15–24, but in a different order.

The y_i cannot be nearly Normally distributed when n_i is small, for example, Hospitals 1 and 2, but we act here as if the y_i are Normal because that distribution is required for the estimators being considered. A more accurate model might approximate the data d_i as Poisson, as Christiansen and Morris [7] do for medical profiling. For the remainder of this section we also focus on shrinkage to 0 ($r = 0$), the approximate average of the y_i.

TABLE 4
NY hospital profiling data and shrinkages

i	y	sd	$\hat{B}_{\mathrm{HB}}$	$\hat{B}_{\mathrm{F}}$	$\hat{B}_{\mathrm{MLE}}$	$\hat{B}_{\mathrm{ADM}}$	$\hat{B}_{\mathrm{SHP}}$	$\sqrt{v}$	$\hat{\mu}_{\mathrm{SHP}}$	s_{SHP}	d	n
1	−2.07	2.78	0.079	0.947	0.952	0.922	0.926	0.047	−0.15	0.76	3	67
2	−0.22	2.76	0.081	0.946	0.952	0.921	0.925	0.047	−0.02	0.76	2	68
3	0.58	1.57	0.249	0.850	0.864	0.790	0.808	0.103	0.11	0.69	5	210
4	−1.87	1.42	0.305	0.823	0.839	0.754	0.777	0.115	−0.42	0.70	11	256
5	−0.74	1.39	0.318	0.817	0.833	0.746	0.770	0.118	−0.17	0.67	9	269
6	−1.97	1.37	0.327	0.812	0.829	0.741	0.766	0.119	−0.46	0.70	12	274
7	−1.90	1.36	0.332	0.810	0.827	0.738	0.763	0.120	−0.45	0.70	12	278
8	2.31	1.32	0.352	0.801	0.818	0.726	0.753	0.124	0.57	0.72	4	295
9	−0.14	1.22	0.413	0.774	0.794	0.694	0.725	0.133	−0.04	0.64	10	347
10	−1.21	1.22	0.413	0.774	0.794	0.694	0.725	0.133	−0.33	0.66	13	349
11	−1.43	1.20	0.427	0.769	0.788	0.687	0.719	0.134	−0.40	0.66	14	358
12	1.56	1.14	0.473	0.750	0.770	0.664	0.700	0.140	0.47	0.66	7	396
13	−0.00	1.10	0.508	0.736	0.758	0.648	0.686	0.144	−0.00	0.62	12	431
14	0.41	1.08	0.527	0.729	0.751	0.640	0.679	0.146	0.13	0.61	11	441
15	0.08	1.04	0.568	0.714	0.736	0.622	0.664	0.149	0.03	0.60	13	477
16	−2.15	1.03	0.579	0.710	0.733	0.618	0.660	0.150	−0.73	0.68	22	484
17	−0.34	1.02	0.590	0.706	0.729	0.613	0.656	0.151	−0.12	0.60	15	494
18	0.86	1.02	0.590	0.706	0.729	0.613	0.656	0.151	0.30	0.61	11	501
19	0.01	1.01	0.602	0.702	0.725	0.608	0.652	0.152	0.00	0.60	14	505
20	1.11	0.98	0.639	0.689	0.713	0.594	0.640	0.155	0.40	0.61	11	540
21	−0.08	0.96	0.666	0.680	0.704	0.584	0.631	0.157	−0.03	0.58	16	563
22	0.61	0.93	0.710	0.666	0.691	0.568	0.618	0.160	0.23	0.58	14	593
23	2.05	0.93	0.710	0.666	0.691	0.568	0.618	0.160	0.78	0.66	9	602
24	0.57	0.91	0.742	0.656	0.681	0.558	0.609	0.161	0.22	0.58	15	629
25	1.10	0.90	0.758	0.651	0.677	0.552	0.604	0.162	0.44	0.59	13	636
26	−2.42	0.84	0.870	0.619	0.646	0.518	0.575	0.167	−1.03	0.68	35	729
27	−0.38	0.78	1.000	0.584	0.611	0.481	0.542	0.171	−0.17	0.53	26	849
28	0.07	0.75	1.000	0.565	0.592	0.461	0.525	0.173	0.03	0.52	25	914
29	0.96	0.74	1.000	0.558	0.586	0.455	0.519	0.174	0.46	0.54	20	940
30	−0.21	0.66	1.000	0.501	0.529	0.399	0.469	0.177	−0.11	0.48	35	1193
31	1.14	0.62	1.000	0.470	0.498	0.369	0.442	0.178	0.64	0.51	27	1340

3.1 Minimaxity in Model-I

It may seem for unequal variances that the James–Stein estimator, which requires equal variances, can still be used. To do this, one would divide the values y_i by their standard errors $\mathrm{sd}_i = \sqrt{V_i}$ to create equal variances and apply James–Stein to y_i/sd_i. Then the shrinkage $\hat{B}_{\mathrm{JS}} = (k-2)/S = 0.697$, where $S = \sum y_i^2/V_i = 41.59$, emerges for estimating μ_i/sd_i. Transforming back to estimate μ_i yields a constant-shrinkage estimator

$$\hat{\mu}_{\mathrm{JS},i} = (1 - \hat{B}_{\mathrm{JS}})y_i.$$

This procedure is Model-I minimax if the loss function,

$$L(\hat{\mu}, \mu) = \sum_{i=1}^{k} W_i(\hat{\mu}_i - \mu_i)^2,$$

has weights $W_i = 1/V_i = n_i/\overline{n}$. However, if the loss function has equal weights $W_i \equiv 1$, then this estimator won't be minimax when the variances, equivalently the patient case-loads n_i, are substantially unequal, that is, it won't have uniformly lower mean squared error than y for all μ. Does any health leader exist with the insight to identify the proper weights W_i and the authority to enforce their use?

For unequal variances, component shrinkages would be expected to depend on i. How should these shrinkages be estimated, and by what criteria should the estimates be guided? Data analysts desire more shrinkage for larger V_i and less for smaller V_i, a pattern consistent with the law of large numbers, and with anticipated regression toward the mean, both of which suggest placing greater reliance on estimates y_i that are based on more data and that have smaller variances. Paradoxically, Model-I minimaxity in the unequal variance setting requires reversed shrinkages (more shrinkage for smaller V_i), as shown next.

Using an integration by parts technique pioneered by Stein [25] and Berger [2], Hudson [12] and Berger [2] independently developed a simple Model-I minimax shrinkage estimator for the sum of (unweighted) squared errors, that is, having risk less than $\sum V_i$, the risk of the unbiased estimate y, for all μ. Their estimator directly extends the James–Stein estimator to unequal variances by shrinking each y_i toward 0 using the shrinkage factor

$$\hat{B}_{\mathrm{HB},i} = \frac{(k-2)/V_i}{\sum_{j=1}^{k}(y_j/V_j)^2}.$$

More generally, this estimator can be adapted easily to provide a minimax estimator for any set of weights W_i in the loss function (by rescaling the y_i to $W_i^{1/2}y_i$, obtaining the shrinkage factors above, and then transforming back to the original scale). In the special case $W_i = 1/V_i$, this rescaling will produce the James–Stein estimator with its equal shrinkages $\hat{B}_{\mathrm{JS},i} \equiv \hat{B}_{\mathrm{JS}}$.

With equal weights $W_i \equiv 1$, the risk of this minimax estimator has a simple unbiased estimate:

$$\hat{R}_{\mathrm{HB}} = \sum_{i=1}^{k} V_i\bigl(1 - (k-2)\hat{B}_{\mathrm{HB},i}\bigr).$$

This is less than $\sum V_i$ for all values of y, because $\hat{B}_{\mathrm{HB},i} > 0$. It follows that the expectation of $\hat{R}_{\mathrm{HB}}$ given μ is less than $\sum V_i$, thereby proving the Hudson–Berger estimator uniformly dominates y and is minimax for an equally weighted loss function.

For the $k = 31$ hospitals the risk estimate of the Hudson–Berger rule is $\hat{R}_{\mathrm{HB}} = 31.25$. This is 36.3% smaller than the risk of the unbiased estimate's $\sum V_i = 49.06$. Slightly more improvement stems from using shrinkages $\min(1, \hat{B}_{\mathrm{HB},i})$. Five hospitals, Hospitals 27–31, have such $\hat{B}_{\mathrm{HB},i} > 1$, as shown in Table 4, and these shrinkages should be truncated at 1. However, these Model-I minimax shrinkage factors are smallest for the hospitals with the largest variances, even though the purpose of combining data in these applications is to borrow strength and thereby improve estimates for hospitals with less data.

Unfortunately, none of the 15 hospitals with the largest variances shrinks even as much as 2/3 of its standard error. By contrast, two of the six hospitals already with the most data and with the smallest variances (Hospitals 26–31) shrunk by about two of their own (small) standard errors, a dramatic adjustment for them. This minimax estimator would thrill the management of Hospital 26, whose negative performance estimate y_{26} (2.8 standard deviations below the mean) is shrunken upward by 2.5 standard deviations to make it nearly average. On the other hand, this minimax estimator shrinks Hospitals 27–31 all the way to 0 (the statewide average), so that Hospital 31 has its strong positive performance $y_{31} = 1.14$ (1.84 standard deviations above the mean) reduced by those 1.84 standard deviations so it also is estimated as average.

3.2 Exchangeability in Model-II

The culprit here is the Model-I minimax criterion, and not the mathematically elegant procedure derived to achieve Model-I minimaxity. With substantially unequal variances and summed equally-weighted squared error losses, achieving Model-I minimaxity (nearly) requires reversed shrinkages, that is, smaller shrinkages for those components with larger V_i. ("Nearly"

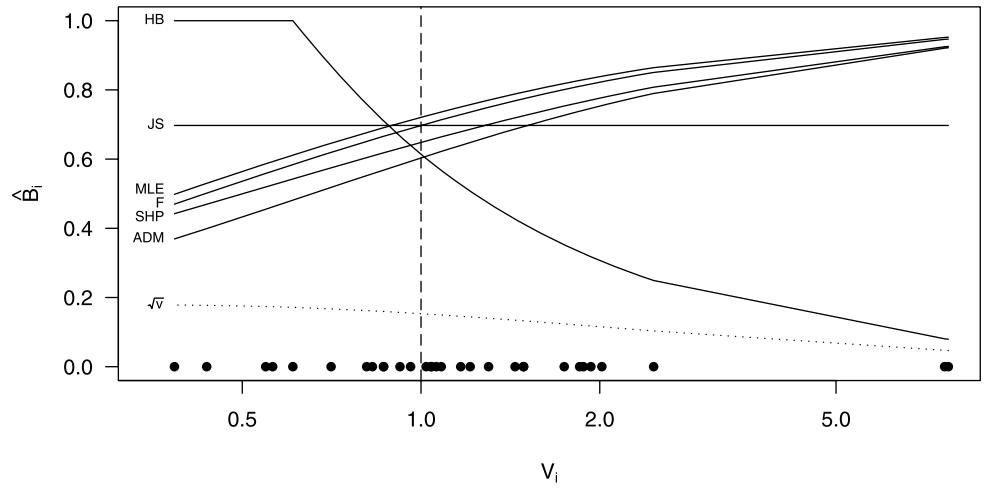

FIG. 4. *Shrinkage factors against V_i for various rules. The dots represent V_i for the* 31 *NY hospitals.*

acknowledges that one could drastically diminish all the larger shrinkages to eliminate the reversal, but then with minuscule resulting shrinkages and no practical benefit.) Meanwhile, procedures that do not suffer from reversed shrinkages abound in practice, by relying instead on exchangeability assumptions in multilevel models and on Bayesian and empirical Bayesian considerations.

Figure 4 shows the Hudson–Berger Model-I minimax shrinkage factors, labeled as "HB," plotted against the variances V_i. Note their reversed shrinkages that decrease as variances increase. The James–Stein shrinkage factors are constant at $\hat{B}_{\mathrm{JS}} = 0.697$, as shown by the horizontal line labeled "JS." Four other shrinkage rules will be introduced next, all motivated by Model-II considerations, so all with shrinkages that increase as variances increase.

Componentwise risks and interval coverages become more valuable when based on averages over both levels of Model-II. This requires accepting Level-II exchangeability for the random effects μ_i (or when $r > 0$, accepting exchangeability of the residuals $\mu_i - x_i'\beta$), given A. Shrinkages now may increase as the variances V_i increase. Exchangeability of μ (or of its residuals) replaces assessing weights for component losses in applications. As in the equal variances case, procedures that dominate on all k components become possible, as well as confidence intervals. With decision theoretic Model-II evaluations, componentwise dominance becomes the goal.

Most data analysts and modelers of real data are familiar with recognizing problems for which exchangeability assumptions are reasonable, for example, they make such judgements routinely for error terms when fitting regressions. Exchangeability considerations would stop anyone from combining estimates of butterfly populations and percentages of sports car sales to augment the estimation of the 31 NY hospital success rates. Model-I standards provide no guidance on this, in favor of requiring assessment of relative weights W_i for butterfly vs. hospital data.

With sufficiently disparate V_i, the minimax estimator of Hudson and Berger is not necessarily minimax for every component by Model-II evaluations. However, Model-II minimax shrinkage estimators do exist for any set of V_i. A recent such procedure by Brown, Nie and Xie [6] produces shrinkages that increase with V_i and with componentwise squared errors smaller than V_i for every i, for all $A \geq 0$, and for any variance pattern $V_1, \ldots, V_k$ for $k \geq 3$.

A popular Model-II shrinkage technique is based on the MLE of A. It provides relatively simple MLE estimates of the shrinkages $\hat{B}_{\mathrm{MLE},i} = V_i/(V_i + \hat{A}_{\mathrm{MLE}})$ and of the unknown means $\hat{\mu}_{\mathrm{MLE},i} = (1 - \hat{B}_{\mathrm{MLE},i})y_i$. It is often used to construct confidence intervals for the μ_i by estimating the conditional variance

$$\mathrm{var}(\mu_i | y, A) = (1 - B_i)V_i.$$

For $r = 0$, $\hat{A}_{\mathrm{MLE}}$ maximizes

$$L(A) = \sum_{i=1}^{k} \bigl(-S_i B_i + \log(B_i)\bigr)/2,$$

where $S_i = y_i^2/V_i$ and $B_i = V_i/(V_i + A)$. If $r > 0$ and Level-II in Table 2 specifies an unknown mean

$$E[\mu_i | \alpha] = x_i'\beta,$$

then restricted maximum likelihood (REML) should be used. This can be accomplished by analytically integrating out (not maximizing out) the r-dimensional β, assuming its prior density is flat in r dimensions, as in [22]. In this case the likelihood $L(A)$ above would be replaced by the resulting integral over β, and then maximization would lead to $\hat{A}_{\mathrm{REML}}$.

When $r > 0$, a larger value of k is required for any possibility of minimaxity, at least $k \geq 3 + r$, with $k \geq 5 + r$ needed for minimaxity of the MLE in the equal variance case. The MLE shrinkages are graphed in Figure 4 for the 31 hospitals on the curve labeled "MLE."

A flaw of the MLE is that $\hat{A}_{\mathrm{MLE}} = 0$ occurs commonly. This not only dictates full shrinkage, but also when $r = 0$ the conditional variance estimates $(1 - \hat{B}_{\mathrm{MLE},i})V_i$ are all equal to zero. In such cases using these for confidence intervals asserts that $\mu_i = 0$ with 100% confidence, a gross overstatement [22].

3.3 Construction at Level-III

Bayesian modeling extends Model-II to Model-III by constructing procedures from a single prior on the hyperparameters at Level-III. Bayes and formal Bayes procedures provide posterior means, variances and posterior distributions for the random effects μ_i, given the data. As such Model-III Bayesian procedures are widely used in applications, the question is: what are their frequency properties? The posterior moments and distributions may not be computable exactly, but they are estimable for any particular data set and prior via MCMC and other simulation techniques. Moreover, the fundamental theorem of decision theory tells us that Model-III constructions (Bayes and formal Bayes) are required for Model-II admissibility.

From the decision-theoretic perspective much more is yet to be learned, even for models as simple as the Normal distributions of Table 2 in Levels I–II. It still isn't known, even with $r = 0$, whether (formal) priors exist that provide Model-II minimax estimators of μ no matter how varied the V_i. Beyond that, only a little has been done in the unequal variance case to determine if posterior probability intervals for formal priors, perhaps computed to offer posterior coverages of 95%, actually cover μ_i for every i, $A \geq 0$ at that nominal 95% level.

3.3.1 *Stein's prior*: *Transported from the equal variance case*. For the family of priors discussed in the equal variances case in Section 2, Stein's SHP stands out as the prime candidate for minimaxity and for confidence intervals in the unequal variances setting, assuming Model-II evaluations. Unfortunately, no general theorems about these properties have been proved for the SHP, formal mathematical proofs being hindered by the complexity of the posterior moments and intervals. However, particular investigations with the SHP have been encouraging.

Indeed, for any shrinkage estimator $\hat{\mu}_i = (1 - \hat{B}_i)y_i$ with $0 < \hat{B}_i < 1$, the difference between the component risks of y_i and $\hat{\mu}_i$ conditioned on A and y,

$$\begin{aligned} r_i &= E[(y_i - \mu_i)^2 | A, y] - E[(\hat{\mu}_i - \mu_i)^2 | A, y] \\ &= B_i^2 y_i^2 - (B_i - \hat{B}_i)^2 y_i^2 \\ &= (2B_i - \hat{B}_i)\hat{B}_i y_i^2, \end{aligned}$$

is positive for any value of $A < V_i$, which, when integrating over y, shows that the Model-II risk of $\hat{\mu}_i$ is less than V_i for any $A < V_i$. Also, SHP will dominate the unbiased estimate y_i when A becomes large enough, since the componentwise Model-II risk converges to that of equal variances as A tends to infinity.

The estimator $E[\mu_i | y]$ for any prior on A, for each i and set of variances V_i, involves computing $E[B_i | y]$. For the SHP, with $L(A)$ being the Model-II likelihood of A, this is

$$E[B_i | y] = \hat{B}_{\mathrm{SHP},i} = \frac{\int_0^\infty V_i/(V_i + A)L(A)\,\mathrm{d}A}{\int_0^\infty L(A)\,\mathrm{d}A},$$

and the resulting estimate of μ_i is

$$\hat{\mu}_{\mathrm{SHP},i} = (1 - \hat{B}_{\mathrm{SHP},i})y_i.$$

As with the equal variance case, the posterior variances $s_i^2 = \mathrm{var}(\mu_i | y)$ for any prior are given by

$$s_i^2 = V_i(1 - E[B_i | y]) + v_i y_i^2,$$

where v_i is the posterior variance of B_i. For SHP this is

$$\begin{aligned} v_i &= \mathrm{var}(B_i | y) \\ &= \frac{\int_0^\infty V_i^2/(V_i + A)^2 L(A)\,\mathrm{d}A}{\int_0^\infty L(A)\,\mathrm{d}A} - \hat{B}_{\mathrm{SHP},i}^2. \end{aligned}$$

The SHP shrinkage estimates $\hat{B}_{\mathrm{SHP},i}$ for the hospital data are plotted in Figure 4 on the curve labeled "SHP." The associated posterior standard deviations $\sqrt{v_i}$ are given by the dotted curve labeled "$\sqrt{v}$." Figure 5 displays a stochastic estimate of the relative Model-II risk improvement of SHP over the unbiased estimate y_i,

$$\frac{V_i - E[(\hat{\mu}_{\mathrm{SHP},i} - \mu_i)^2 | A]}{V_i} = \frac{E[r_i | A]}{V_i}$$

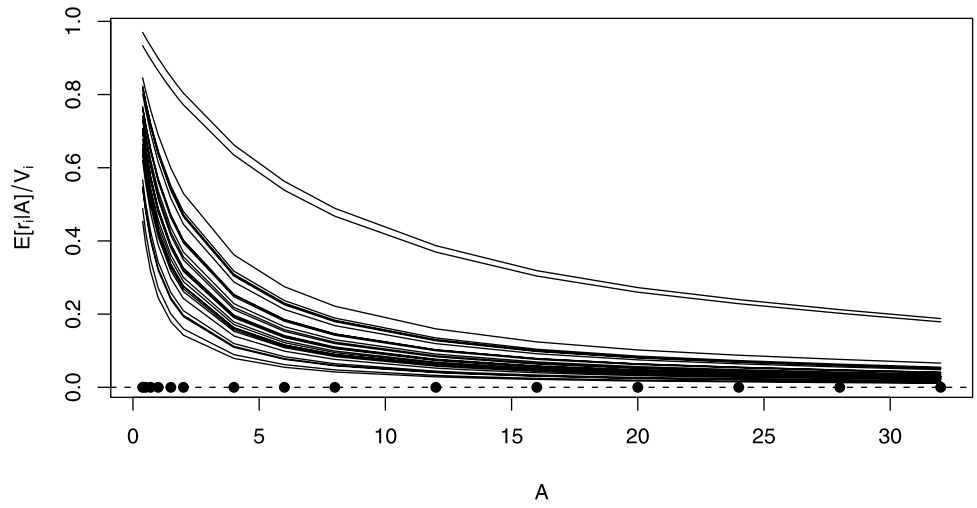

FIG. 5. *Stochastic estimate of SHP's Model-II componentwise relative risk improvement for* 31 *variances as in the hospitals, as a function of A. The dots represent the values of A at which the simulations were performed* (20,000 *replicates of y for each A*).

for $k = 31$ and for the variance pattern of the 31 hospitals $V_1, \ldots, V_{31}$ as a function of A. This was done by simulating 20,000 replicates of y at 15 different values of A, and averaging the 20,000 values of r_i / V_i at each A. Different curves plot the risk improvement for different components i. All the curves are positive and strictly decreasing. The curves are ordered according to their V_i values, the largest (V_1) providing the top curve. Thus, for this variance pattern, at least and seemingly generally, the greatest shrinkage benefit accrues to the components with the greatest uncertainty.

The graph's monotonicity suggests that the minimum Model-II risk improvement for each component occurs as A approaches infinity. That corresponds to the limiting equal variance case. Interestingly, despite their stochastic nature, the curves do not cross each other. These results, although only for one data set, give hope for establishing componentwise Model-II risk dominance for all A of the SHP shrinkage procedure over the unbiased estimate y.

For equal variances, Figure 3 showed that $\hat{\mu}_{\mathrm{SHP},i} \pm 1.96 s_{\mathrm{SHP},i}$ produces minimum coverage of μ_i very close to 95%. Figure 6 investigates the corresponding coverage properties for the unequal variances in the pattern of the 31 NY hospitals. For each y and A of the previous simulation, the coverage probability

$$P(\mu_i \in \{\hat{\mu}_{\mathrm{SHP},i} \pm 1.96 s_{\mathrm{SHP},i}\} | y, A)$$

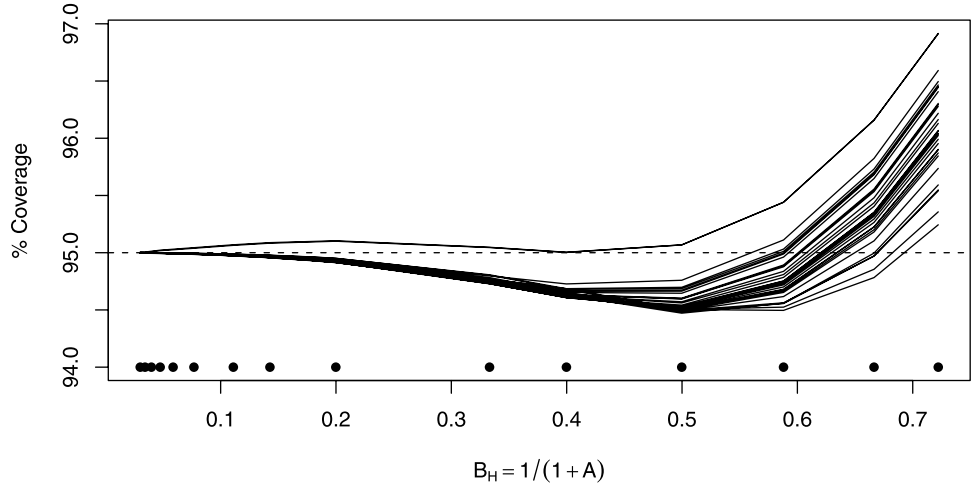

FIG. 6. *Simulation of SHP's* 95% *Normal interval Model-II componentwise coverage probabilities for each of the* 31 *hospital variances as a function of* B_H. *The dots represent the values of* B_H *at which the simulations were performed.*

of μ_i by the "SHP Normal" interval given y and A is analytically computed from

$$\mu_i|y, A \overset{\text{ind}}{\sim} \mathcal{N}[(1-B_i)y_i, V_i(1-B_i)],$$

then averaged over the 20,000 values of y for each A. Thus, Figure 6 displays the coverage probabilities for each Hospital i using Model-II of Table 2 as a function of the harmonic mean B_H of the shrinkage factors,

$$B_H = \frac{k}{\sum_{i=1}^{k} B_i^{-1}} = \frac{V_H}{V_H + A} = \frac{1}{1+A},$$

a monotone decreasing function of A (recall that the 31 CABG indices have been scaled to have $V_H = 1$).

All but two of the 31 curves exhibit a pattern similar to that of equal variances in Figure 3 when $k = 20$: exactly 95% coverage for B_H close to 0, a minimum with 0.5% under-coverage near $B_H = 0.6$, and over-coverage for B_H close to 1. The curves are nonintersecting and increasing with V_i for the 4 highest values of B_H, but cross each other repeatedly for $B_H < 0.5$, presumably because of simulation inaccuracy. The two nearly superimposed highest curves which never (or barely) overcover μ_i correspond to the two hospitals with the highest variances, Hospitals 1–2, these variances being nearly 8 times the size of the 31 variances' harmonic mean. In all cases the coverage probabilities are never below 94.5%.

Figure 7 compares SHP and unshrunken estimates, and their standard deviations for the data with the 31 NY hospitals. The absolute value of the rules, $|\hat{\mu}_{\text{SHP},i}|$ (circle) and $|y_i|$ (+/−), are plotted above the x-axis, and the negative standard deviations, $-s_{\text{SHP},i}$ and $-\text{sd}_i$, are plotted below. "Plus" signs indicate that the estimates were positive, for example, Hospitals 3 and 8, whereas "minus" signs indicate that the estimates were negative, for example, Hospitals 1–2. It appears for these data that all the SHP coverage intervals will be shorter than those of the unbiased estimate, although this need not always hold for all data sets y, as discussed earlier for the equal variances in Section 2.

3.3.2 *Posterior mean versus posterior mode*: *The ADM technique*. Deriving the SHP rule for unequal variances requires numerical computation of $k+1$ integrals (including the common denominator in $\hat{B}_{\text{SHP},i}$). ADM (adjustment for density maximization, Morris [20]) is used here for shrinkage estimation to provide a relatively simple approximation to the SHP, as in Morris and Tang [22]. To explain the ADM, the MLE provides a simple shrinkage formula from the mode of the likelihood $L(A)$ that is equivalent to the posterior mode of A for the SHP. However, the mode of a right-skewed distribution like that of A underestimates the mean. Furthermore, the mean $E[B_i|y]$ is needed, not the mode. The ADM provides a better approximation than the MLE for shrinkage factors while still requiring only two derivatives to approximate the posterior distributions of $B_i|y$. The ADM can be used with various priors in Figure 2, but here we apply it to approximate the posterior distribution of each shrinkage $B_i|y$ when the SHP is the chosen prior distribution for A.

For shrinkage estimation, ADM approximates the distribution of each $B_i = V_i/(V_i + A)$ by a Beta distribution. Because shrinkage coefficients lie in $[0, 1]$, and these coefficients linearly determine the Level-II distributions (Table 2), two-parameter Beta distributions

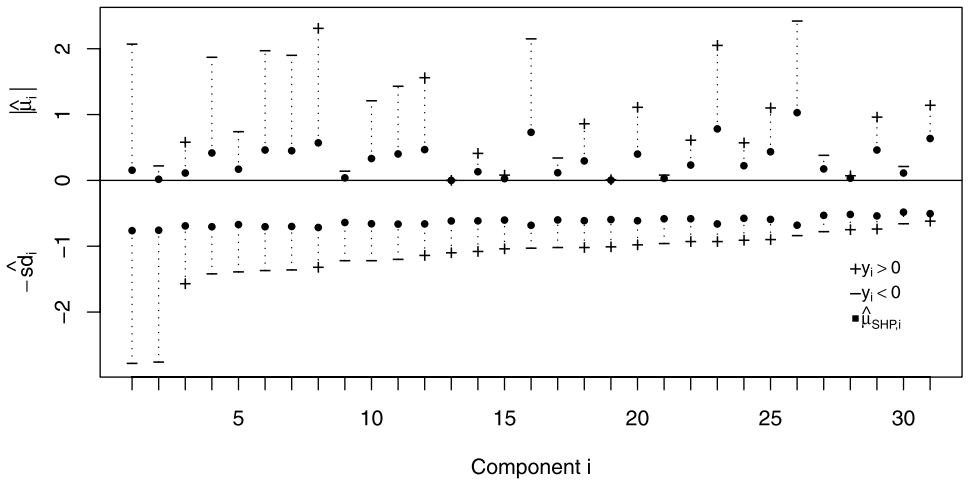

FIG. 7. *SHP (black circles). Absolute values of unshrunken and SHP estimates with signs indicated by (+/−) top half. Standard deviations (bottom half) for SHP are always closer to* 0 *than* V_i. *"Plus" signs indicate positive estimates* $y_i \geq 0$.

are the natural choice for shrinkage approximations, and not the Normal distribution (the distribution for which MLE and the posterior mean would coincide). When the prior on A is taken to be the SHP, and with Beta distribution approximations to $B_i = V_i/(V_i + A)$, the ADM "adjustment" simply amounts to maximizing $A \cdot L(A)$, rather than $L(A)$, for each i and V_i. Note that the maximum always occurs with $A \geq 0$. Calling the maximizing value $\hat{A}_{\mathrm{ADM}}$, then $E[B_i|y]$ is approximated by $\hat{B}_{\mathrm{ADM},i} = V_i/(V_i + \hat{A}_{\mathrm{ADM}})$. This ADM approach has been used before for shrinkage estimation, for example, by Christiansen and Morris [7], Li and Lahiri [15] and Morris and Tang [22].

For the 31 hospitals, $\hat{A}_{\mathrm{ADM}} = 0.657$, so $E[B_i|y]$ is approximated by $\hat{B}_i = V_i/(V_i + 0.657)$. The variances of B_i could be obtained from the second derivative of $\log(A \cdot L(A))$ at the adjusted mode, $\hat{A}_{\mathrm{ADM}} = 0.657$. The ADM shrinkages are graphed in Figure 4 on the curve labeled "ADM." They are more conservative than those of the MLE, and indeed follow the SHP curve closely for all but the smallest variances V_i.

As was seen before in Figure 3, standard errors and interval estimates with the SHP coverages as approximated by ADM are never perceptibly below 95%, for equal variances and $k = 4, 10, 20$. The ADM is readily applicable to approximate posterior point and interval estimates for other priors on A in the unequal variance case. Further, Model-II evaluations of ADM include investigations by Morris–Tang [22] for Normal distributions, Everson–Morris [11] for multivariate Normal data, and Christiansen–Morris [7] for Poisson data. Evidence therein with special cases and/or with special data sets has been quite encouraging, with no negative experiences thus far.

3.4 Potential of the Multilevel Model: A Useful Rule of Thumb

For equal variances, good shrinkage rules such as James–Stein or SHP are simple enough to calculate that they can be implemented immediately in practice. For unequal variances the calculations are much more involved and easily accessed software may be unavailable or need to be mastered. Researchers justifiably may ask how much they stand to gain by fitting a hierarchical model before actually fitting it, their alternatives being to use unbiased estimates $\hat{\mu}_i = y_i$ or the fully shrunken estimates, here $\hat{\mu}_i = 0$ (for $r = 0$), or when $r > 0$ to shrink all the way to a grand mean or to a linear regression estimate.

A helpful feature of using MLE or ADM methods to fit shrinkages, perhaps with a model like that of Table 2, is that a simple point estimate $\hat{A}$ of A suffices to estimate all shrinkage factors B_i, and consequently also all means μ_i. Moreover, an estimate $\hat{A}$ of A leads to a simple estimate $\hat{B}_H$ of the harmonic mean of the shrinkage factors B_H through the identity

$$B_H = V_H/(V_H + A). \tag{4}$$

Analogously to its equal variance counterpart B, the harmonic mean shrinkage $0 \leq B_H \leq 1$ provides a useful summary for gauging the benefits of fitting a shrinkage model. Values of B_H close to 0 suggest that there will be relatively little shrinkage overall, in which case a researcher might be justified to use the unbiased estimates y_i. Or, values of B_H close to 1 might justify using the fully shrunken regression estimates $x_i'b$,

$$b = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

where $X' = [x_1, \ldots, x_k]$ and $V = \mathrm{diag}(V_1, \ldots, V_k)$. Values of B_H near $1/2$ give the strongest case for estimating shrinkages.

Letting $S = \sum_{i=1}^{k}(y_i - x_i'b)^2/V_i$, when $r \geq 0$ and the variances are equal, $B_H = B$ and we have

$$E[S|A] = (k-r)/B \quad \text{and} \quad E[(k-r-2)/S|A] = B,$$

which leads to the James–Stein estimator. When the variances are unequal, it is easily seen for $r = 0$ that

$$E[S|A] = \sum_{i=1}^{k}(V_i + A)/V_i = k/B_H.$$

Taken together these facts suggest a simple point estimate for B_H,

$$\hat{B}_H = \frac{k-r-2}{S} = \frac{k-r-2}{k-r} \times \frac{1}{\hat{\sigma}^2},$$

where

$$\hat{\sigma}^2 = \frac{1}{k-r} \times \sum_{i=1}^{k} \frac{(y_i - x_i'b)^2}{V_i}$$

is the mean square error from a (weighted linear) regression output. Note that one can easily rearrange (4) to solve for

$$\hat{A} = V_H(1 - \hat{B}_H)/\hat{B}_H.$$

This estimate, in turn, can be used to provide simple estimates of each individual shrinkage factors B_i by $V_i/(V_i + \hat{A})$. Even if $\hat{B}_H$ is small, having this rough estimate of every B_i is useful in case there are a few $\hat{B}_i$ that are appreciably bigger than 0.

These estimates of B_i are plotted as the fourth and final Model-II rule in Figure 4, labeled "F," giving a curve that is almost identical to the MLE shrinkages.

Data analysts can use this easy "rule-of-thumb" that can be based on regression outputs for anticipating individual and overall shrinkages, without computing more precise shrinkage estimates. For the 31 hospitals $S = 41.59$ and $\hat{B}_H = 0.697$, suggesting that a good Model-II rule would outperform both the individual estimates y_i and the fully shrunken estimates, alike.

4. SUMMARY AND CONCLUSIONS

We have reviewed a special and relatively simple class of hierarchical models, models for Normal distributions that have received significant attention from a nonasymptotic (in k) decision-theoretic perspective. Early equal-variance Model-I shrinkage estimators, evaluated by a (unweighted) sum of squared errors criterion, were found that provided Model-I minimaxity and even admissibility. That opened exciting new vistas. However, the great preponderance of applications (even when Normal distributions apply) arise with unequal variances, and there Model-II evaluations are seen to be much more appropriate. Model-II evaluations are both less and more general than Model-I, less because they average over the Level-II parameters, and more general by not requiring judgements about appropriate weights for component losses, and also by empowering interval estimation. A Level-II exchangeability assumption, for example, as in Table 2, enables componentwise Model-II dominance to be possible.

Many more investigations are needed in the Model-II setting for small and moderate numbers k of random effects $\mu = (\mu_1, \ldots, \mu_k)'$. Does Stein's harmonic prior (SHP) transport to the unequal variance case, for example, by offering Model-II componentwise minimaxity, conditionally on all hyperparameters, especially on all $A \geq 0$? Our experience suggests that this is entirely possible for both the equal and the unequal variances settings, but there are no formal proofs yet. Does the full posterior distribution, geared to offer 95% posterior probability of coverage for fixed data with the SHP prior, provide intervals that cover at least 95% of the time? Showing this with Model-II would require at least 95% coverage for every fixed value of $A \geq 0$ that holds for every component (e.g., for every hospital), after averaging over both levels of Model-II. If intervals cover less than 95% of the cases, how close does the minimum coverage come to 95%? How well and when do relatively simple methods for estimating shrinkages work, like MLE and ADM methods? What Level-III priors lead to Model-II dominance by providing componentwise minimaxity and confidence intervals that are shorter for every component? Do SHP intervals cover every μ_i more often for every i, A than do the standard (unshrunken) confidence intervals used by data analysts?

These theoretical questions about operating characteristics under Model-II evaluations can be asked for other yet more complicated models, especially for other distributions at Level-I and at Level-II. Shrinkage estimators arise when fitting generalized linear multilevel models to data that follow exponential families at Level-I, if conjugate distributions are used for the Level-II random effects. That is, just as Normal conjugate distributions are used at Level-II in Table 2, Gamma distributions are conjugate when Level-I specifies Poisson likelihoods, and Betas are conjugate for Binomial likelihoods. The advantage of conjugate distributions at Level-II is that shrinkage factors arise in conditional means, given the observations. Crucially, conjugate distributions are relatively robust, having the virtue of being "G_2 minimax" among all possible Level-II distributions (priors) in the sense of Jackson et al. [13, 18]. This helps make shrinkage estimators simple and robust. Shrinkage factors also provide useful summaries, so can serve a purpose like R^2 does with OLS regressions.

We have argued that Model-II and it's exchangeability assumptions are more appropriate than Model-I for developing and evaluating shrinkage estimators. This holds especially for applications in which improvements would be expected to hold for every μ_i. Hospital directors might agree to having their own hospital's performance be estimated by combining information from other hospitals, but not unless each was assured that doing so would make their own hospital's estimate more accurate.

This paper argues especially that evaluations of shrinkage methods for unequal variance data have received too little attention, relative to the large literature on the Normal equal variances case. It is time to change that.

REFERENCES

[1] BARANCHIK, A. (1964). Multiple regression and estimation of the mean of a multivariate normal population Technical Report 51, Dept. Statistics, Stanford Univ.

[2] BERGER, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* **4** 223–226. MR0397940

[3] BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. ed. Springer, New York. MR0804611

[4] BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136. MR0216647

[5] BROWN, L. D. (2009). Personal communication.
[6] BROWN, L. D., NIE, H. and XIE, X. (2011). Ensemble minimax estimation for multivariate normal means. *Ann. Statist.* To appear.
[7] CHRISTIANSEN, C. L. and MORRIS, C. N. (1997). Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.* **92** 618–632. MR1467853
[8] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238
[9] EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.
[10] EFRON, B. and MORRIS, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **4** 11–21. MR0403001
[11] EVERSON, P. J. and MORRIS, C. N. (2000). Inference for multivariate normal hierarchical models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 399–412. MR1749547
[12] HUDSON, H. M. (1974). Empirical Bayes estimation Technical Report 58, Dept. Statistics, Stanford Univ.
[13] JACKSON, D. A., O'DONOVAN, T. M., ZIMMER, W. J. and DEELY, J. J. (1970). $\mathcal{G}_2$-minimax estimators in the exponential family. *Biometrika* **57** 439–443. MR0270491
[14] 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, CA. MR0133191
[15] LI, H. and LAHIRI, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *J. Multivariate Anal.* **101** 882–892. MR2584906
[16] MORRIS, C. (1977). Interval estimation for empirical Bayes generalizations of Stein's estimator. The Rand Paper Series, The Rand Corporation.
[17] MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849
[18] MORRIS, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *Ann. Statist.* **11** 515–529. MR0696064
[19] MORRIS, C. N. (1983). Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis, and Robustness* (*Madison, Wis.*, 1981). *Publ. Math. Res. Center Univ. Wisconsin* **48** 25–50. Academic Press, Orlando, FL. MR0772762
[20] MORRIS, C. N. (1988). Approximating posterior distributions and posterior moments. In *Bayesian Statistics* 3 (*Valencia*, 1987) 327–344. Oxford Univ. Press, New York. MR1008054
[21] MORRIS, C. N. and LOCK, K. F. (2009). Unifying the named natural exponential families and their relatives. *Amer. Statist.* **63** 247–253. MR2750349
[22] MORRIS, C. and TANG, R. (2011). Estimating random effects via adjustment for density maximization. *Statist. Sci.* **26** 271–287.
[23] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. 3rd Berkeley Sympos. Math. Statist. Probab.* **I** 197–206. Univ. California Press, Berkeley. MR0084922
[24] STEIN, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Research Papers in Statistics* (*Festschrift J. Neyman*) 351–366. Wiley, London. MR0210232
[25] STEIN, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of Prague Symposium on Asymptotic Statistics* (*Charles Univ., Prague*, 1973) **II** 345–381. Charles Univ., Prague. MR0381062
[26] STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **24** 265–296. MR0148184
[27] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098
[28] STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939