Analytic Robust Inference for Small Area Means Using Student t Prior Distributions

by

Richard B. Evans[*]

Iowa State University

Ames, IA

Abstract

Hierarchical models for L studies, domains or experiments often assume that the study means have a common normal population distribution. However, modeling normal sampling distributions with a normal population distribution may overstate the level of exchangeability of the studies and cause "borrowing of strength" among dissimilar

Key Words and Phrases: Heavy tails, empirical Bayes, complex analysis, robust inference, small area estimation

[*] email address: revans@iastate.edu. Part of this work was done while the author was visiting the National Center for Toxicological Research. The author thanks Jeff Zimmerman for discussions of the swine farm size example, and Joe Sedransk for discussions about heavy tailed models.

domains. When there is uncertainty about the similarity of the domains, using heavy tailed population distributions, in particular t distributions, provide some protection from combining dissimilar studies, domains or experiments (Gelman, Carlin, Stern, and Rubin, 1995, O'Hagan, 1988). We also use t population distributions, but use an analytic method instead of approximations or Monte Carlo methods to provide posterior inference. We expand the current analytic inferential methodology to the situation when when the sampling variance and the prior scale are unknown. In the case that the prior parameters are unknown the analytic method may be modified to provide a parametric empirical Bayes method for inference about the study means. The analytic method circumvents Markov chain Monte Carlo convergence problems and permits a more direct model sensitivity analysis. Using examples and analytic results we demonstrate the characteristics of posterior means and variances under t distribution priors, and suggest that for applied data analysis problems the Cauchy prior is a reasonable population distribution. The examples also suggest that the prior variance should be estimated from the data rather than assigned an arbitrary constant. Finally we demonstrate the method using real data in a small area estimation example.

1 Introduction

It is sometimes desired to pool data from multiple sources to strengthen inference for the mean of a particular study, experiment, or domain of interest. For example, sample surveys may obtain a few observations from small domains (e.g., counties), but inference is required for characteristics of the small domain (Datta and Lahiri, 1995). Methods for combining information in small area estimation may not account for the uncertainty associated with combining possibly dissimilar domains. Another example (Evans and Sedransk, 2001) is 6 clinical trials that investigate the efficacy of aspirin therapy after myocardial infarction. The sample size for each trial is large, and for all the trials mortality is not significantly different between the aspirin and placebo group. For the largest trial (N=4524), greater mortality is associated with aspirin use. Inference for each trial could sharpend if the data were combined. However, some caution must be exercised to guard against combining dissimilar trials.

A traditional approach for combining information across studies is with a normal hierarchical model (Gelman, Carlin, Stern, and Rubin, 1995). Let y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, L$ be data from L studies. Often the sample mean in considered normally distributed and the sampling distribution is

$$\overline{y}_i | \theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \frac{\sigma_i^2}{n_i}\right), \; j = 1, \ldots, n_i, \; i = 1, \ldots, L, \tag{1}$$

where $\overline{y}_i = \sum_j^{n_i} y_{ij}/n_i$, $i = 1, \ldots, L$. It is also common to replace σ_i^2 in (1) with the sample variance. To combine information across the studies they are considered exchangeable, with

$$\theta_i | \eta, \delta^2 \overset{ind}{\sim} N\left(\eta, \delta^2\right), \; i = 1, \ldots, L, \tag{2}$$

and a diffuse prior for η. We are primarily interested in inference about the θ_i, and if σ_i^2 and δ^2 are known, then using (1) and (2),

$$E\left(\theta_i|y\right) = \lambda_i \overline{y}_i + (1-\lambda_i)\overline{\overline{y}},$$

where $y = \{y_{ij} : j = 1, \ldots, n_i,\ i = 1, \ldots, L\}$, $\lambda_i = \delta^2/(\delta^2 + \sigma_i^2/n_i)$, and $\overline{\overline{y}} = (\sum_{i=1}^{L} n_i/\sigma_i^2)^{-1} \sum_{i=1}^{L} \overline{y}_i(n_i/\sigma_i^2)$. However, (2) may be overly optimistic in the sense that it assumes that the studies are exchangeable. If λ_i is not 1, then the posterior mean for each study i "borrows strength" from the other studies through $\overline{\overline{y}}$, even if the studies are not similar.

When the studies are partially exchangeable a general form of the prior distribution is

$$\theta_i|\eta_k, \delta_k^2 \overset{ind}{\sim} N\left(\eta_k, \delta_k^2\right),\ i \in S_k,\ k = 1, \ldots, G, \tag{3}$$

where the S_k are a collection of mutually exclusive and exhaustive subsets of $\{i : i = 1, \ldots, L\}$. For example, if $G = 1$ then (3) is the same as (2), and if $G = L$ then none of the studies are modeled as exchangeable and no "borrowing strength" occurs. Prior distribution (3) combines the studies in sets of similar studies. The problem is that the composition of the S_k may be unknown, and then robust inference for the θ_i is desirable to guard against "borrowing strength" across dissimilar studies.

One class of robust methods are partition models (George, 1986, Hartigan, 1990, Malec and Sedransk, 1993, Evans and Sedransk, 2001) which model the uncertainty about the S_k. A partition g is a decomposition of the set $\{i : i = 1, \ldots, L\}$ into mutually exclusive and exhaustive subsets $S_k\left(g\right)$. For example, if $L = 6$ then the partition $\{(1,2,3,4,5,6)\}$ corresponds to (2) and $\{(1,2,6),(3,4,5)\}$ models studies 1, 2 and 6 as exchangeable independently from studies 3, 4 and 5, which are also modeled as exchangeable. For $L = 6$ there are 203 partitions. Let $d(g)$ be the number of studies in $S_k(g)$ and let

$$\theta_i|\eta_k(g), \delta_k^2(g) \overset{ind}{\sim} N\left(\eta_k(g), \delta_k^2(g)\right), \; i \in S_k(g), \; k = 1, \ldots, d(g),$$

be the prior distribution of the θ_i conditional on g. Let the $\eta_k(g)$ have diffuse prior distributions, g have prior distribution $p(g)$, and assume that the variance components are known. Then

$$E(\theta_i|y) = \sum_g E(\theta_i|y,g)\, p(g|y),$$

where $p(g|y)$ is the posterior probability of partition g. The conditional expected values exhibit "borrowing of strength" within subsets defined by the $S_k(g)$, but not between the subsets. This is a model averaging approach where the conditional (on g) expected values are averaged over the posterior probability of the partitions. Partitions that are congruent with the data have larger posterior probability so that information from similar studies are combined with higher probability. Some partition models have good asymptotic properties (Evans and Sedransk, 2001). However, the number of partitions grows faster than 2^L (Hartigan, 1990) and the large number of partitions may make computation unwieldy. The number of partitions may be reduced by assigning prior probability 0 to some partitions, or using model space reduction methods such as Occam's Window (Madigan and Raftery, 1994).

Another robust approach for inference about the θ_i is to use (1) and a heavy tailed prior distribution for the θ_i. Dawid (1973) showed that under some regularity conditions, if a prior distribution has heavier tails than the likelihood then discordant prior information is rejected from inference. O'Hagan (1988) investigated two heavy tailed models. In this paper we consider one of the models, which uses (1), and the t distribution

$$p(\theta_i|\eta, \delta^2) \propto \delta^\beta \{\beta\delta^2 + (\theta_i - \eta)^2\}^{-\frac{\beta+1}{2}}, i = 1, \ldots, L. \tag{4}$$

Note that δ^2 is not the prior variance, which is $\delta^2\beta/(\beta - 2)$ for $\beta > 2$. O'Hagan (1988) calls δ^{-2} the prior "strength." We follow a similar convention and call δ^2 the prior "scale."

O'Hagan (1988) uses (4) with $n_i = 1$, $\sigma_i^2 = 1$, $\beta = 5$, $\delta^2 = 1/4$, $L = 3$, data y_1, y_2, and y_3, corresponding to the three studies, and a diffuse prior on η. Using numerical results he shows that the posterior means and standard deviations of θ_1 and θ_2 reject y_3 as y_3 becomes large, and the posterior means and standard deviations for θ_1 and θ_2 are consistent with pooling data from studies one and two only, that is, the third study is ignored. Also, the posterior mean and standard deviation of θ_3 reject y_1 and y_2 as y_3 becomes large. For $L > 3$ outlying data are rejected from inference about experiments in the main body of data and are also not pooled among themselves (O'Hagan, 1988). For example, if $L = 6$ and y_5 and y_6 are outlying but grouped, then y_5 and y_6 do not influence posterior inference for θ_i, $i < 5$ and y_5 (y_6)

does not influence inference about θ_6 (θ_5). Therefore (4) is most applicable when the studies under consideration are assumed to be exchangeable but wish to be modeled with some caution to guard against relatively few outliers.

Several authors have investigated analytic approaches to inference using models with normal sampling distributions and heavy tailed priors. Generally, previous work has either used Monte Carlo methods including MCMC, a known variance, or an approximation. Angers (1992) uses t distributions with odd degrees of freedom and known sampling variation. The means (e.g.; the θ_i) are integrated analytically, and the hyperparameters of the t distributions are integrated using numerical methods. Pericchi and Smith (1992) also assume known sampling variation and use t distribution priors for inference about a single mean. They develop approximations to the first and second posterior moments of the mean using the Laplace method. Datta and Lahiri (1995), in the context of small area estimation, use a multivariate normal sampling distribution with known covariance matrix and a Cauchy prior distribution on the small area means. They provide asymptotic results ((i.e.; when an observation goes to infinity) that are consistent with the rejection properties of heavy tailed priors. Spiegelhalter (1985) uses an integration method from complex analysis to provide inference for parameters with Cauchy distributions using non-informative prior distributions. We use the same integration method in our work.

In this paper we describe an analytic method to provide inference for the θ_i using

(1), (4) with odd degrees of freedom and

$$p(\sigma_i^2) \propto \sigma_i^{-2},\ i = 1, \ldots, L, \tag{5}$$

when η and δ^2 are known. In applied problems η and δ^2 are usually unknown, and δ^2 may require a more careful treatment than assigning it an arbitrary fixed value. We describe an empirical Bayes (EB) method (Carlin and Louis, 2000) that may be used for inference about the θ_i when η and δ^2 are unknown. The approach is to use (1), (4) and (5) and analytically integrate the θ_i and the σ_i^2, giving the marginal distribution of the data $m(y|\eta, \delta^2)$. A simple numerical "peak finding" method is used to determine parametric empirical Bayes estimates η_{EB} and δ^2_{EB} that maximize $m(y|\eta, \delta^2)$ with respect to η and δ^2. Using (4) the EB estimates are substituted into the prior giving the EB prior

$$p_{EB}(\theta_i) \propto \delta^{\beta}_{EB}\{\beta\delta^2_{EB} + (\theta_i - \eta_{EB})^2\}^{-\frac{\beta+1}{2}}, i = 1, \ldots, L. \tag{6}$$

Finally, using (1), (5) and (6), we provide inference for the θ_i, again using the analytic method that is described below. We will show that the inferences using the EB method demonstrate the outlier rejection properties consistent with heavy tailed prior distribution models.

An alternative to the analytic and EB methods described in this paper is to use (1), (4), and (5) and a fully Bayes approach by assigning prior distributions to η and δ^2 and then use Markov chain Monte Carlo methods to provide inference for the

θ_i. The EB approach is often preferable because simulations using the fully Bayes approach suggest that successful convergence of the Markov chain representing the target distribution for δ^2 is dependent on the severity and number of the outlying studies. If the data are such that iterates of δ^2 with values close to zero are sampled regularly from the target distribution of δ^2 then convergence of the chains for the θ_i is extremely slow. This happens for some robust models because the heavy tailed prior rejects the outlying studies and inference for δ^2 uses the set of similar studies with sampling means that are relatively close together.

The next section describes the technique to obtain the marginal distribution of the data and posterior moments of the θ_i when the prior parameters are known. A simple example is provided to demonstrate the method. Section 3 presents the results of numerical experiments designed to describe the properties of the first two posterior moments of the θ_i for select values of prior parameters. The parametric empirical Bayes estimators for η and δ^2 are developed in section 4. Section 5 is an comparison of state level average swine farm size using two models and a real data set, the 1995 National Animal Health Monitoring System (NAHMS), general swine farm report survey (Losinger, W.C., Bush, E. J., Hill, G. W., Smith, M. A., Garber, L. P., Rodriguez, J. M., and Kane, G., 1998). The models are a normal hierarchical model (that assumes the states are exchangeable) and a t prior distribution model that protects inference for the θ_i against states that have possibly outlying average swine farm sizes. Finally, section 6 is the discussion of the paper.

2 Inference for the θ_i with known η and δ^2

In this section we describe the method to integrate the θ_i, provide a simple example to demonstrate the method and show the analytic formula for the first two moments of the posterior distribution of θ_i for a simple case.

We begin by considering a single study and give a description of the method of integration for an individual θ_i, and then apply that method to integrate all the θ_i. The integration uses a result from complex analysis (described below) which requires the calculation of the residues of the joint distribution of the data and θ_i. Residues are defined as the coefficient of the first negative power term of the Laurent series for the joint distribution, and can be calculated using derivatives (with respect to θ_i) that are evaluated at the complex poles (singularities) of the joint distribution. The first step is to integrate σ_i^2 so that the joint distribution of the data and θ_i is a product of two t distributions. It is straightforward to determine the complex poles of the t distributions and then calculate the residues that are associated with the poles.

Using (1) and (5), let $\prod_{j=1}^{n_i} g(y_{ij}, \sigma_i^2|\theta_i)$ be the joint distribution of the data (for domain i) and σ_i^2, then

$$f(y_i|\theta_i) = \int_0^\infty \prod_{j=1}^{n_i} g\big(y_{ij}, \sigma_i^2|\theta_i\big)\, d\sigma_i^2 \propto t_{n_i-1}\left(y_i, \theta_i, s_i^2/n_i\right), i = 1, \ldots, L, \tag{7}$$

where $y_i = \{y_{ij} : j = 1, \ldots, n_i\}$, $t_{n_i-1}(y_i, \theta_i, s_i^2/n_i)$ is a t distribution with $n_i - 1$

degrees of freedom, mean θ_i, scale s_i^2/n_i, and $s_i^2 = \frac{1}{n_i - 1}\sum_j (y_{ij} - \overline{y}_i)^2$ (Gelman, Carlin, Stern, and Rubin, 1995). Note that for small area estimation problems n_i will be small, so that $f(y_i|\theta_i)$ will have small degrees of freedom. For outlier rejection to occur (4) should have heavier tails than (7), that is, $\beta < n_i - 1$. Using (4) and (7) the marginal distribution of the data and θ_i is proportional to the product of two t distributions,

$$f\left(y_i|\theta_i\right)p(\theta_i) \propto \frac{\left(\frac{s_i}{\sqrt{n_i}}\right)^{\alpha_i}\delta^\beta}{\left\{\alpha_i \frac{s_i^2}{n_i} + (\theta_i - \overline{y}_i)^2\right\}^{\frac{\alpha_i+1}{2}}\left\{\beta\delta^2 + (\theta_i - \eta)^2\right\}^{\frac{\beta+1}{2}}}, \tag{8}$$

where $\alpha_i = n_i - 1$. The t distributions are rational functions of θ_i and have no real poles, so that $f\left(y_i|\theta_i\right)p(\theta_i)$ has no real poles (e.g.; real roots of the denominator). However, t distributions can be factored over the complex plane to determine their complex poles. For the likelihood,

$$\begin{aligned} t_{\alpha_i}\left(y_i, \theta_i, s_i^2/n_i\right) &\propto \left(\frac{s_i}{\sqrt{n_i}}\right)^{\alpha_i}\left\{\alpha_i\frac{s_i^2}{n_i} + (\theta_i - \overline{y}_i)^2\right\}^{-\frac{\alpha_i+1}{2}} \\ &= \left(\frac{s_i}{\sqrt{n_i}}\right)^{\alpha_i}\left\{\left[\theta_i - (\overline{y}_i - \sqrt{\alpha_i\frac{s_i^2}{n_i}I})\right]\left[\theta_i - (\overline{y}_i + \sqrt{\alpha_i\frac{s_i^2}{n_i}I})\right]\right\}^{-\frac{\alpha_i+1}{2}} \end{aligned} \tag{9}$$

where $\alpha_i = n_i - 1$ and the poles are $\theta_i = \overline{y}_i \pm \sqrt{\alpha_i\frac{s_i^2}{n_i}I}$ with $I = \sqrt{-1}$. The positive complex pole is called the upper half plane pole, and it's order defined as $(\alpha_i + 1)/2$. The prior distribution (4) is factored as

$$t_\beta\left(\eta, \delta^2\right) \propto \delta^\beta\left\{\left[\theta_i - (\eta - \sqrt{\beta\delta^2 I})\right]\left[\theta_i - (\eta + \sqrt{\beta\delta^2 I})\right]\right\}^{-\frac{\beta+1}{2}}, \tag{10}$$

with poles $\theta_i = \eta \pm \sqrt{\beta\delta^2}I$ that have order $(\beta+1)/2$. Note that if $\alpha_i > \beta$, then we may experience the rejection of η in the posterior mean of θ_i, described in Dawid (1973). Also, we require both α_i and β to be odd integers.

To calculate the posterior distribution of θ_i from (8), the normalizing constant is calculated using the formula

$$m(y_i) = \int_{-\infty}^{\infty} f\left(y_i|\theta_i\right) p(\theta_i) d\theta_i = 2\pi I \left\{\text{sum of the upper half plane residues}\right\}, \quad (11)$$

where $y_i = \{y_{ij} : j = 1, \ldots, n_i\}$ (Kreyszig, 1988).

To calculate (11) it is necessary to determine the upper half plane residues of $f\left(y_i|\theta_i\right) p(\theta_i)$ at the two upper half plane poles. Residues are calculated at a pole by removing the pole from $f\left(y_i|\theta_i\right) p(\theta_i)$ (by multiplying by the appropriate term), taking one fewer derivatives than the order of the pole, and then substituting the pole into the result:

$$R_{i1} = \frac{1}{c_i!}\frac{d^{c_i}}{d\theta_i}\left\{ f\left(y_i|\theta_i\right) p(\theta_i) \left[\theta_i - (\overline{y}_i + \sqrt{\alpha_i \frac{s_i^2}{n_i}} I)\right]^{(c_i+1)} \right\}\Bigg|_{(\overline{y}_i + \sqrt{\alpha_i \frac{s_i^2}{n_i}} I)} \quad (12)$$

and

$$R_{i2} = \frac{1}{c_0!}\frac{d^{c_0}}{d\theta_i}\left\{ f\left(y_i|\theta_i\right) p(\theta_i) \left[\theta_i - (\eta + \sqrt{\beta\delta^2} I)\right]^{(c_0+1)} \right\}\Bigg|_{(\eta + \sqrt{\beta\delta^2} I)}, \quad (13)$$

where $c_i = n_i/2 - 1$ and $c_0 = (\beta+1)/2 - 1$. The terms in square brackets remove the poles from $f\left(y_i|\theta_i\right) p(\theta_i)$ so the residues are not singular. If c_i or c_0 are 0 then there is no derivative and the upper half plane poles are substituted into (12) and (13) for θ_i.

Using (11), (12) and (13) the marginal distribution of the data is

$$m(y_i) \propto 2\pi I \left\{ R_{i1} + R_{i2} \right\}, \tag{14}$$

and using (14) the posterior distribution of θ_i is

$$p\left(\theta_i|y_i\right) = \frac{f\left(y_i|\theta_i\right)p(\theta_i)}{m(y_i)}.$$

The k^{th} posterior moment ($k = 2, 3, ...$) is calculated with

$$\begin{aligned} E\left\{\left[\theta_i - E\left(\theta_i|y_i\right)\right]^k |y_i\right\} &= \frac{1}{m(y_i)}\int_{-\infty}^{\infty}\left[\theta_i - E\left(\theta_i|y_i\right)\right]^k f\left(y_i|\theta_i\right)p(\theta_i)d\theta_i,\ k = 2,3,4,\ldots \\ &= \frac{2\pi I}{m(y_i)}\left\{\text{sum of the upper half plane residues}\right\} \\ &= \frac{\left\{R^*_{i1} + R^*_{i2}\right\}}{m(y_i)}, \qquad (15) \end{aligned}$$

where

$$R^*_{i1} = \frac{1}{c_1!}\frac{d^{c_i}}{d\theta_i}\left.\left\{\left[\theta_i - E\left(\theta_i|y\right)\right]^k f\left(y_i|\theta_i\right)p(\theta_j)\left[\theta_i - (\overline{y}_i + \sqrt{\alpha_i\frac{s_i^2}{n_i}I})\right]^{(c_i+1)}\right\}\right|_{(\overline{y}_i+\sqrt{\alpha_i\frac{s_i^2}{ni}I})},$$

and

$$R^*_{i2} = \frac{1}{c_0!}\frac{d^{c_0}}{d\theta_i}\left.\left\{\left[\theta_i - E\left(\theta_i|y\right)\right]^k f\left(y_i|\theta_i\right)p(\theta_i)\left[\theta_i - (\eta + \sqrt{\beta\delta^2 I})\right]^{(c_0+1)}\right\}\right|_{(\eta+\sqrt{\beta\delta^2 I})},$$

are the upper half plane residues of $\left[\theta_i - E\left(\theta_i|y\right)\right]^k f\left(y_i|\theta_i\right)p(\theta_i)$. Because $\left[\theta_i - E\left(\theta_i|y\right)\right]^k$ is multiplied in the numerator of (8), the poles are unchanged (unless $k+2$ is greater than the order of θ_i in the denominator, in which case the residue method does not apply). The first posterior moment is (15) with θ_i substituted for $\left[\theta_i - E\left(\theta_i|y_i\right)\right]^k$.

Example 1 *Calculation of residues and posterior moments*

In this example we demonstrate (11) by providing inference for a single study mean θ_i when η and δ^2 are known. Let $\alpha_i = 3$, $\beta = 1$ and $s_i^2/4 = \tau^2$. The first step is to use (11), (12) and(13) to calculate the normalizing constant $m(y_i)$ where y_i is the data set. Next, (15) is used to calculate the posterior expected value and posterior variance of θ_i. Using (8), (9) and (10)

$$f\left(y_i|\theta_i\right)p(\theta_i) \propto$$

$$\frac{\tau^3\delta}{\left[\theta_i-(\overline{y}_i+\sqrt{3}\tau I)\right]^2\left[\theta_i-(\overline{y}_i-\sqrt{3}\tau I)\right]^2\left[\theta_i-(\eta+\sqrt{\delta^2}I)\right]\left[\theta_i-(\eta-\sqrt{\delta^2}I)\right]},$$

which has a upper half plane pole of order 2 at $\theta_i = \overline{y}_i + \sqrt{3}\tau I$, and another of order 1 at $\theta_i = \eta + \sqrt{\delta^2}I$. The corresponding residues are

$$R_1 = \frac{d}{d\theta_i}\left\{\frac{\tau^3\delta}{\left[\theta_i-(\overline{y}_i-\sqrt{3}\tau I)\right]^2\left[\theta_i-(\eta+\sqrt{\delta^2}I)\right]\left[\theta_i-(\eta-\sqrt{\delta^2}I)\right]}\right\}\Bigg|_{(\overline{y}_i+\sqrt{3}\tau I)},$$

and

$$R_2 = \left\{\frac{\tau^3\delta}{\left[\theta_i-(\overline{y}_i+\sqrt{3}\tau I)\right]^2\left[\theta_i-(\overline{y}_i-\sqrt{3}\tau I)\right]^2\left[\theta_i-(\eta-\sqrt{\delta^2}I)\right]}\right\}\Bigg|_{(\eta+\sqrt{\delta^2}I)}.$$

Evaluating R_1 requires substantial algebraic manipulation. When the poles are of higher order it may be useful to use symbolic manipulation software. Finally,

$$m(y_i) \propto 2\pi I\left\{R_1+R_2\right\} = \frac{2\pi\sqrt{3}\left(\delta^3+\delta x^2+6\sqrt{3}\tau^3+15\delta\tau^2+4\sqrt{3}\delta^2\tau\right)}{18\left(x^2+3\tau^2+2\sqrt{3}\tau\delta+\delta^2\right)},$$

where $x^2 = \left(\overline{y}_i - \theta_i\right)^2$.

The moments of the posterior distribution of θ_i are calculated using the same technique of integration because they add $\left[\theta_i - E\left(\theta_i|y_i\right)\right]^k$ to the numerator of the residues which does not affect the poles. However, (11) is only valid when the degree of the polynomial (with respect to θ_i) in the denominator is at least 2 more than the degree of the numerator. For the expectation of θ_i use (11),

$$E\left(\theta_i|y_i\right) = \frac{\int_{-\infty}^{\infty} \theta_i f\left(y_i|\theta_i\right) p(\theta_i) d\theta_i}{m(y_i)},$$

where the of the numerator of the expected value is evaluated using the residues

$$R_1^* = \frac{d}{d\theta_i}\left\{\frac{\theta_i \tau^3 \delta}{\left[\theta_i - (\overline{y}_i - \sqrt{3}\tau I)\right]^2 \left[\theta_i - (\eta + \sqrt{\delta^2} I)\right] \left[\theta_i - (\eta - \sqrt{\delta^2} I)\right]}\right\}\Bigg|_{(\overline{y}_i + \sqrt{3}\tau I)}$$

and

$$R_2^* = \left\{\frac{\theta_i \tau^3 \delta}{\left[\theta_i - (\overline{y}_i + \sqrt{3}\tau I)\right]^2 \left[\theta_i - (\overline{y}_i - \sqrt{3}\tau I)\right]^2 \left[\theta_i - (\eta - \sqrt{\delta^2} I)\right]}\right\}\Bigg|_{(\eta + \sqrt{\delta^2} I)}$$

Note that the residues only change with a θ_i in their numerators. Then

$$\begin{aligned} E\left(\theta_i|y_i\right) &= \left\{R_1^* + R_2^*\right\} / \left\{R_1 + R_2\right\} \\ &= \frac{\left[\delta(9\tau^2 + x_i^2) + 4\sqrt{3}\tau\delta^2 + \delta^3\right]\overline{y}_i + 6\left[\delta\tau^2 + \sqrt{3}\tau^3\right]\eta}{\left[\delta(9\tau^2 + x_i^2) + 4\sqrt{3}\tau\delta^2 + \delta^3\right] + 6\left[\delta\tau^2 + \sqrt{3}\tau^3\right]}, \end{aligned} \tag{16}$$

and

$$V\left(\theta|y\right) = 3\frac{\sqrt{3}\tau^2\left(\sqrt{3}\delta + 6\tau\right)\left(\delta^2 + x_i^2 + 2\sqrt{3}\tau\delta + 3\tau^2\right)^2}{\left(12\tau\delta^2 + \sqrt{3}\delta^3 + \sqrt{3}\delta x_i^2 + 18\tau^3 + 15\sqrt{3}\tau^2\delta\right)^2}, \tag{17}$$

where $x_i^2 = \left(\overline{y}_i - \eta\right)^2$.

The posterior expected value of θ_i is a convex combination of the prior mean and $\overline{y}_i$. The weights are a function of the sampling and prior scale and the distance between the sample mean and the prior mean. The coefficient of $\overline{y}_i$ is a function of x_i so as the distance between the means increase, more weight is on the sample mean, which is expected because the likelihood has lighter tails than the prior.

3 Numerical experiments

This section describes some of the properties of the posterior expected value and posterior variation of $\theta_i = \theta$ (a single experiment) when η and δ^2 are known, using graphs with overlaying curves generated from models with different values of the prior parameters. The model for each graph uses (15) with $\overline{y} = 1$ and $s^2/n = 1/n$, and a Cauchy prior distribution (i.e.; (4) with $\beta = 1$). There are two reasons to focus on examples that use Cauchy prior distributions. First, domains in small area estimation problems have small sample sizes so that the sampling distribution has low degrees of freedom. In order to maintain the outlier rejection property the prior distribution must have smaller degrees of freedom than the sampling distribution, and there may be only a few t distributions with sufficiently small degrees of freedom. The second reason (demonstrated in the graphs) is that the amount of outlier rejection is often modest even when the sampling distribution has substantially lighter tails than the prior. The sample size determines the sampling degrees of freedom (for (7)) so that the only way to achieve the largest difference in tail weight (and the maximum protection against outliers) is to use $\beta = 1$.

Using (15), Figure 1 is a plot of $E(\theta|y)$ for four models corresponding to $\alpha = n-1 = 3, 5, 19, 49$ and $\delta^2 = 1$, as a function of η. For each of the models, as the distance between η and $\overline{y}$ increases, η is rejected from the posterior expected value. Rejection of prior information is most pronounced for the the curve with the largest degrees of freedom, $\alpha = 49$, (the dash-dot curve) but when α is greater than 19, the gains in rejection are minimized (solid curve). Thus when a Cauchy prior distribution is used sampling distribution with tails lighter than 19 degrees of freedom may not offer substantial gains in rejection. The dashed curve represents the $\alpha = 3$ case. Although rejection of η occurs, it does so only for extreme outliers and the rejection is gradual. The sampling standard error is a decreasing function of the degrees of freedom (α) so that there is also less borrowing of strength as α increases and the sampling standard error becomes small relative to the prior variance.

Figure 2 is a plot of $E(\theta|y)$ for $\alpha = 19$, and $\eta = 1.05, 1.15, 2$ as a function of the square root of the prior scale, δ. This plot shows the effect of δ on the posterior mean. The choices of η represent three amounts of discrepancy with respect to $\overline{y} = 1$ and $s^2/n = 1/20$. The prior mean $\eta = 1.05$ is congruent with the data for any value of δ, but $\eta = 1.15$ may conflict with the data, and $\eta = 2$ is an outlier unless δ is large. For each curve, $E(\theta|y) = \eta$ for very small δ and then have a moderate decline to $\overline{y} = 1$, except for the dashed curve ($\eta = 2$) which has a steep initial decline. In the example the posterior mean (16) suggests that large δ downweight the contribution of η to the posterior mean. This is reflected in Figure 2, where for the most discrepant prior mean, $\eta = 2$, rejection begins at very small δ, but then has a very gradual decline to $E(\theta|y) = \overline{y} = 1$. The dashed curve corresponding to the "marginal outlier," $\eta = 1.15$, (which is the most interesting case from a practical point of view because it is a suspected outlier) is influenced by δ over a fairly wide range of δ. Thus when δ is unknown, a coherent treatment is required instead of arbitrarily assigning a fixed value of δ, such as $\delta = 1$.

Using (15), Figure 3 is a plot of $V(\theta|y)$ for four models corresponding to $\alpha = 3, 5, 19, 49$ and $\delta^2 = 1$, as a function of η. For all the curves, the posterior variance is the smallest when $\overline{y} = 1$ and η agree. As $|\overline{y} - \eta|$ increases the posterior variation also increases until the discordant prior information is rejected and then the posterior variance decreases. The dashed curve corresponds to $\alpha = 3$ and is (17). Although for $\alpha = 3$ the posterior expected value rejects the discordant η, the variance, using (17), is asymptotically $3/4\left(1 + \sqrt{3}\right)$. This suggests that the heavy tailed method works best when the study sample size permits at least 5 degrees of freedom. For $\alpha = 5, 19, 49$ the closed form formulae for the posterior variances do not lend themselves to easy interpretation.

Figure 4 is a plot of $V(\theta|y)$ for $\alpha = 19$, and $\eta = 1.05, 1.2, 1.7, 2$ as a function of δ. This plot shows the effect of the square root of the prior scale on the posterior variation. The four curves show different patterns. The dash-dot curve ($\eta = 1.05$)

and the dashed curve ($\eta = 1.2$) both have monotonic behavior. Although $\eta = 1.2$ is relatively far from $\overline{y} = 1$ it's corresponding curve is similar to the one corresponding to $\eta = 1.05$, the prior mean that is close to the data. The dotted curve ($\eta = 1.7$) and the solid curve ($\eta = 1.7$) first increases and then decreases as the conflict with the prior is resolved. All have posterior variance equal to the sampling variance ($1/20 \times 19/17$) for large δ.

4 Unknown η and δ^2

In this section we describe a method to provide inference for the θ_i when η and δ^2 are unknown. The method is parametric empirical Bayes (Carlin and Louis, 2000). The joint distribution of the data and the θ_i is the product of the L pairs of t distributions (8), with upper half plane poles $\theta_i = \overline{y}_i + \sqrt{\alpha_i \frac{s_i^2}{n_i} I}$, and $\theta_i = \eta + \sqrt{\beta \delta^2 I}$, $i = 1, \ldots, L$. Using (5),(7) and (14),

$$\begin{aligned} m\left(y\right) & \propto \prod_{i=1}^{L} \int_{-\infty}^{\infty} f\left(y_i | \theta_i\right) p(\theta_i) d\theta_i \\ & = \prod_{i=1}^{L} 2\pi I \left\{R_{i1} + R_{i2}\right\}, \end{aligned} \tag{18}$$

where $y = \{y_{ij} : j = 1, \ldots, n_i,\ i = 1, \ldots, L\}$.

The next step is to maximize (18) over η and δ^2 to produce the EB estimates η_{EB} and δ^2_{EB}. If the maximizer for δ^2 is negative, then by definition $\delta^2_{EB} = 0$. The EB estimates reflect the rejection property. For example, if there are several outlying studies with relatively large sample means, then η_{EB} will be smaller than under a normal - normal model.

Inference for the θ_i follows directly from the method for known η and δ^2; use (15) with η_{EB} and δ^2_{EB} substituted into the equation for η and δ^2. If $\delta^2_{EB} = 0$ then using (4) the residues are not defined and the residue integration method is not applicable. However, that case suggests that there are no outlying studies, that is, the variation of the study means are within the sampling variance and the studies are estimating a common population mean.

5 NAHMS example

In this section we analyze a data from the 1995 National Animal Health Monitoring System (NAHMS), general swine farm report survey (Losinger, W.C., Bush, E. J., Hill, G. W., Smith, M. A., Garber, L. P., Rodriguez, J. M., and Kane, G., 1998) using a normal hierarchical model and a heavy tailed model and then compare the posterior inferences. This example demonstrates that in the presence of a possible outlier, inference using the robust model provides greater shrinkage of posterior means from similar studies than under a normal hierarchical model. Intuitively, this is because in the normal model the outlier increases the population variance, so that there is less shrinkage to the population mean. In the robust model, the outlier is rejected, so the population variance is smaller.

NAHMS is a United States Department of Agriculture sponsored program to collect, analyze and disseminate information about animal (primarily production animal) health, management, and productivity. Data are collected with a series of surveys and biological samples. The swine surveys are limited to swine farms and to states that produce the majority of swine. In 1995, 418 swine farms from 16 states (these states accounted for 91% of the total US hog and pig production in 1995) were surveyed with questionnaires. The number of farms surveyed in each state is roughly proportional to the number of swine farms in the state. The 1995 general swine farm report survey has 204 questions. Additional topical surveys with disease specific questions and biological sample collection were also used. One survey question asks for the swine farm sizes (total number of hogs and pigs on June 1, 1995). For this example we provide inference for the mean of the logarithm transformed swine farm size for each state. Some states have only a few sampled farms.

The sample data are summarized in Table 1. The distribution of farm size by state is skewed because most swine producing states have some large “mega producers” with tens of thousands of swine. The first step is to use the natural logarithm transformation to reduce skewness, and then we provide inference for average of the logarithm transformed farm size. The first three columns of Table 1 list the state names, number of farms surveyed in each state, and the median farm size. Column 4 gives the sample average and standard error of the transformed data that we will be using in this example.

Using column 4 of Table 1, The states appear to have about the same average transformed farm size, with the exception of North Carolina. In this example, North Carolina is the possible "outlier." Georgia, Tennessee and Kentucky have the smallest sample sizes, so we expect that they will be affected by outliers under the normal model, but have a conservative mean under the robust model.

5.1 The normal hierarchical model

Let y_{ij}, $j = 1, \ldots, n_i$ $i = 1 \ldots 16$, be the transformed farm sizes, and n_i be number of farms surveyed by state (from Table 1). Let

$$y_{ij}|\theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \frac{\sigma_i^2}{n_i}\right), \; j = 1, \ldots, n_i, \; i = 1, \ldots, 16. \tag{19}$$

We wish to contrast the robust method with a method that assumes the θ_i are from a common distribution without rejection properties. Let

$$\theta_i|\eta, \delta^2 \overset{ind}{\sim} N\left(\eta, \delta^2\right), \; i = 1, \ldots, 16. \tag{20}$$

We assume proper, non-informative prior distributions for the nuisance parameters.

5.2 The heavy tailed prior model

The heavy tailed model is (5), (6) and (7) with $\beta = 1$. The Cauchy prior is used for two reasons; the Cauchy prior offers the most outlier protection, and because some of the states have few farms sampled which bounds the degrees of freedom of the prior distribution. If a state has an odd number of sampled farms then the sampling distribution (7) will have an even degree of freedom. The residues require odd degree of freedom, so we conservatively use an even sample size of $n_i - 1$. For example, using Table 1, Indiana has 17 observations, and the sampling distribution for Indiana, (7), is assigned 15 degrees of freedom instead of 16 degrees of freedom. This adjustment has no practical effect on inference except for very small sample sizes, and for this example there were no differences in inference between adjusting the odd sample sizes with $n_i - 1$ or with $n_i + 1$.

Maximum likelihood estimates for η and δ^2 are obtained from (18) using an iterative maximization algorithm. The EB estimates are $\eta_{EB} = 7.218$ and $\delta^2_{EB} = 0.003$.

5.3 Results

Inference for the normal model (i.e.; (19), (20)) is obtained using Markov chain Monte Carlo. Ten thousand iterates were used (one of every 10 sampled from a chain of length 100,000) after a burn-in of 5000 iterates. Standard convergence tests verified that the chains converged to the target distributions.

An attempt was made to use MCMC with a heavy tailed model ((19), with (20) replaced with (4) and $\beta = 1$ and non-informative priors on the hyperparameters). However, convergence of the chain corresponding to δ^2 was extremely difficult to obtain because the iterates are often close to zero. This is expected because the target distribution of δ^2 reflects the population variation of the states with North Carolina and other outliers either rejected or downweighted, so that the mode of the distribution is close to zero.

The results for the example are in Table 2, which contains the posterior expectations and posterior standard deviations for θ_i, $i = 1, \ldots, 16$, under both models. The posterior expectation for North Carolina, under the robust model (9.167) is much closer sample median (9.222) than is the posterior expectation under the normal model (8.172). To roughly convert the log means to the original scale for average farm sizes, $\exp(9.167) = 9,576$ and $\exp(8.172) = 3,540$ which corresponds to a 37% decrease in farm size. Typically one wishes to protect studies against the undue influence of outliers, but in this example, the robust method protects the "outlier" from excessive shrinkage.

For the three states with the smallest sample sizes, Georgia, Kentucky and Tennessee, the difference between the sample mean and the posterior expected values are greater for the heavy tailed model than for the normal model. That is, more borrowing strength occurred.

Finally, the posterior standard deviations are all smaller under the heavy tailed model, but empirical Bayes estimates don't reflect the variability of η and δ^2.

6 Discussion

We have proposed a method to use t distributions for robust inference for practical data analysis problems. The advantages of the method are that it circumvents possible MCMC convergence problems and that the empirical Bayes method is relatively easy to implement. Also, the method permits inference for means when the sampling variance and prior scale are unknown. In some cases the posterior mean and variance are in a form that is easily interpretable.

The graphs in Section 3 suggest that the prior degrees of freedom influence inference (for small degrees of freedom). The prior degree of freedom will rarely be known a priori, and there are several approaches to address this issue. First, one can select a degree of freedom using a principle, such as the conservative approach using the Cauchy prior. Second, a sensitivity analysis may be used to assess the impact of the degrees of freedom on posterior inference and a choice made after evaluating the sensitivity analysis. Finally, averaging the posterior inferences over the degrees of freedom is an approach consistent with a fully Bayes procedure. Gelman, Carlin, Stern, and Rubin (1995) give an example of 8 coaching programs to improve PSAT scores. The sample sizes are large enough to assume a normal likelihood (with known variance), and several t distributions are used for robust inference about the mean scores for the coaching programs. They demonstrate a sensitivity analysis for the prior degrees of freedom, concluding that prior degrees of freedom have little impact on posterior inference for the coaching program means (Gelman et al., 1995, p. 357). However, they caution that extreme posterior tail probabilities may be affected by prior degrees of freedom. For their example, the normal likelihood (a t distribution with infinite degrees of freedom) permits a wide range of possible prior distributions, in contrast to the NAHMS example which only permits priors with a few degrees of freedom. They also describe how to average over the degrees of freedom, which would be a logical step if the sensitivity analysis indicated that the choices of prior degrees of freedom impact posterior and there are no logical choice of prior degrees of freedom.

References

Angers, J. F. (1992), “Use of the Student-t prior for the estimation of normal means: A computational approach”, Bayesian statistics 4. Proceedings of the Fourth Valencia International Meeting, 567-575

Carlin, Bradley P. , and Louis, Thomas A. (2000), “Bayes and empirical Bayes methods for data analysis, (Second edition)”, Chapman & Hall Ltd (London; New York)

Datta, G. S. , and Lahiri, P. (1995), “Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers”, Journal of Multivariate Analysis, 54, 310-328

Dawid, A. P. (1973), “Posterior expectations for large observations”, Biometrika, 60, 664-667

Evans, R. and Sedransk, J. (2001), “Bayesian inference when the pooling of data is uncertain,” to appear in Biometrika.

Gelman, A. , Carlin, J. B. , Stern, H. S. , and Rubin, D. B. (1995), “Bayesian data analysis”, Chapman & Hall Ltd (London; New York)

George, Edward I. (1986), “Combining minimax shrinkage estimators”, Journal of the American Statistical Association, 81, 437-445

Hartigan, J. A. (1990), “Partition models”, Communications in Statistics, Part A – Theory and Methods, 19, 2745-2756

Kreyszig, E. (1988), “Advanced Engineering Mathematics, (Sixth edition),” John Wiley and Sons (New York)

Losinger, W.C., Bush, E. J., Hill, G. W., Smith, M. A., Garber, L. P., Rodriguez, J. M., and Kane, G. (1998), “Design and implementation of the United States National Animal Health Monitoring System 1995 National Swine Study,” Preventive Veterinary Medicine, 34, 147-159

Madigan, David , and Raftery, Adrian E. (1994), “Model selection and accounting for model uncertainty in graphical models using Occam”s window”, Journal of the American Statistical Association, 89, 1535-1546

Malec, Donald , and Sedransk, J. (1992), “Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain”, Biometrika, 79, 593-601

O’Hagan, A. (1988), “Modelling with heavy tails”, Bayesian statistics 3, 345-359

Pericchi, L. R. , and Smith, A. F. M. (1992), “Exact and approximate posterior moments for a normal location parameter”, Journal of the Royal Statistical Society, Series B, Methodological, 54, 793-804

Spiegelhalter, D. J. (1985), “Exact Bayesian inference on the parameters of a Cauchy distribution with vague prior information”, Bayesian statistics 2, 743-749

Table 1. NAHMS data summary

State	sample size	median farm size	average ln farm size (se)
Georgia	6	2571	7.813 (0.438)
Illinois	34	2021	7.574 (0.184)
Indiana	17	1679	7.430 (0.300)
Iowa	79	1121	7.200 (0.121)
Kansas	26	780	7.026 (0.210)
Kentucky	10	1341	7.344 (0.339)
Michigan	19	1820	7.589 (0.246)
Minnesota	70	1111	7.294 (0.128)
Missouri	24	1087	7.254 (0.2129)
Nebraska	29	880	7.040 (0.199)
North Carolina	24	12246	9.222 (0.219)
Ohio	27	771	6.836 (0.206)
Pennsylvania	12	1356	7.069 (0.309)
South Dakota	12	921	7.207 (0.309)
Tennessee	9	850	6.908 (0.357)
Wisconsin	20	972	6.822 (0.240)

Table 2. Posterior inferences for two models

	Heavy tailed model		Normal hierarchical model	
State	$E(\theta\|y)$	$\sqrt{V(\theta\|y)}$	$E(\theta\|y)$	$\sqrt{V(\theta\|y)}$
Georgia	7.284	0.201	7.411	0.353
Illinois	7.326	0.148	7.465	0.195
Indiana	7.251	0.114	7.365	0.180
Iowa	7.213	0.064	7.206	0.103
Kansas	7.181	0.103	7.100	0.170
Kentucky	7.232	0.119	7.300	0.248
Michigan	7.292	0.147	7.422	0.255
Minnesota	7.285	0.070	7.278	0.125
Missouri	7.224	0.090	7.251	0.194
Nebraska	7.182	0.099	7.117	0.185
North Carolina	9.167	0.234	8.172	0.635
Ohio	7.114	0.155	6.958	0.181
Pennsylvania	7.199	0.115	7.135	0.183
South Dakota	7.216	0.109	7.234	0.248
Tennessee	7.181	0.140	7.042	0.205
Wisconsin	7.131	0.156	6.963	0.194

Figure 1: Plots of $E(\theta|y)$ as a function of the prior mean η. The curves correspond to sampling distributions with different degrees of freedom. The dashed line corresponds to 3 df, the dotted line corresponds to 5 df, the solid line corresponds to 19 df, and the dash-dot line corresponds to 49 df. In each case the prior distribution is Cauchy. The sampling variance is 1/n and the prior strength is 1.

Figure 2: Plots of $E(\theta|y)$ as a function of the prior strength δ. All curves have a sampling distribution with 19 df and a Cauchy prior distribution. The solid curve corresponds to $\eta = 1.05$, the dotted curve corresponds to $\eta = 1.15$, and the dashed curve corresponds to $\eta = 2$.

Figure 3: Plots of $V(\theta|y)$ as a function of the prior mean η. The curves correspond to sampling distributions with different degrees of freedom. The dashed line corresponds to 3 df, the dotted line corresponds to 5 df, the solid line corresponds to 19 df, and the dash-dot line corresponds to 49 df. In each case the prior distribution is Cauchy. The sampling variance is 1/n and the prior strength is 1.

Figure 4: Plots of $V(\theta|y)$ as a function of the prior strength δ. All curves have a sampling distribution with 19 df and a Cauchy prior distribution. The dash-dot curve corresponds to $\eta = 1.05$, the dashed curve corresponds to $\eta = 1.3$, the dotted curve corresponds to $\eta = 1.7$ and the solid curve corresponds to $\eta = 2$.