1-1-2008

Empirical Bayes and Hierarchical Bayes Estimation of Skew Normal Populations

Naveen K. Bansal
Marquette University, naveen.bansal@marquette.edu

Mehdi Maadooliat
Marquette University, mehdi.maadooliat@marquette.edu

Xiaowei Wang
Marquette University

Accepted version. *Communications in Statistics - Theory and Methods*, Vol. 37, No. 7 (2008): 1024-1037. DOI.

Empirical Bayes and Hierarchical Bayes Estimation of Skew Normal Populations

Naveen K. Bansal, Mehdi Maadooliat and Xiaowei Wang

Department of Mathematics, Statistics, and Computer Science, P.O. Box 1881, Marquette University, Milwaukee, WI 53201, U.S.A.
naveen.bansal@marquette.edu

ABSTRACT

We develop an empirical and hierarchical Bayesian methodologies for the skew normal populations through the EM algorithm and the Gibbs sampler. A general concept of skewness to the normal distribution is considered throughout. Motivations are given for considering the skew normal population in applications, and an example is presented to demonstrate why the skew normal distribution is more applicable than the normal distribution for certain applications.

MSC: 62Exx, 62F15, 62P10
Keywords: Skew normal distribution; Gibbs sampler; EM algorithm.

1. Introduction

When testing for a treatment effect in a pre-post study, generally paired t-test is used under the assumption that the shift from pre to post follows a normal distribution. In other words, it assumes that $d = Y - X$, where X and Y are the pre and post variables, follows normal distribution. The paired t-test works fairly well if $d \sim p(\frac{x-\xi}{\tau})$, where $p(\cdot)$ is a symmetric density with no heavy tail. However, the assumption that the shift from pre to post treatment is symmetric around some ξ may not be realistic in many applications. For example, when studying the effect of a particular treatment, it is possible that the effect only applies to a minority (or a majority) of the population, while rest of the population are not affected by the treatment. In such cases, the distribution of the shift would be asymmetric. There are many distributions that can be considered to model this asymmetry. One distribution that has received a lot of attention recently is the skew normal distribution. The basic idea is that if $f(\cdot)$ is a symmetric density,

then asymmetry around some ξ can be introduced by considering a density of the form

$$\frac{2}{\tau} f(\frac{x-\xi}{\tau}) G(\lambda \frac{x-\xi}{\tau}), \tag{1.1}$$

where $G(\cdot)$ is the cumulative distribution function corresponding to some symmetric density g, as long as $\lambda \neq 0$. This idea was first introduced by Azzalini (1985) for the normal density. If f and g both are $N(0,1)$ density, then (1.1) is called the skew-normal distribution $SN(\xi, \tau^2, \lambda)$. One of the benefits of this distribution is that the skewness can be introduced by a single parameter λ. For more details on $SN(\xi, \tau^2, \lambda)$, readers are referred to Azzalini(1985, 1986), Branco and Day(2001), and a collection of papers in Genton (2004). For some extensions to $SN(\xi, \tau^2, \lambda)$, see Azzalini and Dalla Valle (1996), Arellano-Valle, Gomez and Quintana (2003), and Branco and Day (2001).

In this paper, we discuss the empirical Bayes and Hierarchical Bayes methodologies for the skew-normal populations. As a motivation, we will refer to the example of pre to post treatment effect throughout the paper. In section 2, we present some motivations behind the skew-normal distribution, and give some preliminary results. In section 3, empirical Bayes methodology, and in section 4, hierarchical Bayesian methodology are presented.

2. Motivation and some Preliminary Results

The purpose of this section is to present some motivations behind the skewness parameter λ. To make the motivation clear, we assume $\xi = 0$, and $\tau = 1$ in this section. Note that the results of this section can easily translated to the general case by the transformation $X = \xi + \tau S$, where $S \sim SN(0, 1, \lambda)$.

There are basically three different representation results of the skew-normal distribution.

1. If (X, Y) is a bivariate normal random vector with $E(X) = E(Y) = 0$, $Var(X) = Var(Y) = 1$, and $Corr(X, Y) = \delta$, then the conditional distribution of Y given $X > 0$ is $SN(0, 1, \lambda(\delta))$, where $\lambda(\delta) = \delta/\sqrt{1-\delta^2}$.

2. If Y_0 and Y_1 are independent $N(0,1)$ random variates, and $\delta \in (-1, 1)$, then $S = \delta|Y_0| + \sqrt{(1-\delta^2)}Y_1$ follows $SN(0, 1, \lambda(\delta))$.

3. If $Y \sim N(0,1)$, and if conditionally on $Y = y$,

$$Q_Y = \begin{cases} +1 \text{ with probability } \Phi(\lambda y) \\ -1 \text{ with probability } 1 - \Phi(\lambda y) \end{cases} \tag{2.1}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0,1)$, then $S = Q_Y Y$ follows $SN(0,1,\lambda)$.

The first representation is due to Arnold et al.(1993), while the other two are due to Azzalini(1985, 1986). In this paper, we use a modified version of the third representation, which has been also termed as generalized skew normal distribution by some authors (Loperfido, 2004). In this modified version, the distribution function $\Phi(\cdot)$ in (2.1) is replaced by a general distribution function $G(\cdot)$ of a symmetric random variable. Thus, if $Y \sim N(0,1)$, and if conditionally on $Y = y$,

$$Q_Y = \begin{cases} +1 \text{ with probability } G(\lambda y) \\ -1 \text{ with probability } 1 - G(\lambda y) \end{cases} \tag{2.2}$$

where $G(\cdot)$ is a distribution function of a symmetric random variable, then it is easy to see that the probability density of $S = Q_Y Y$ is

$$h(x;\lambda) = 2\phi(x)G(\lambda x), \quad -\infty < x < \infty, \tag{2.3}$$

where $\phi(x)$ is the $N(0,1)$ density. Note that general $G(\cdot)$ brings flexibility to the skewness in the normal distribution. One of the $G(\cdot)$ that we will focus in this paper is

$$G(x) = \frac{e^x}{1+e^x}, \quad -\infty < x < \infty. \tag{2.4}$$

Since G is only used to introduce skewness in the normal distribution, we still call the density (2.3) as skew-normal instead of generalized skew-normal, but denote its distribution by $SN(0,1,\lambda,G)$ to emphasize its dependence on G. The corresponding distribution with the location ξ and the scale τ (having density (1.1)) will be denoted by $SN(\xi,\tau^2,\lambda,G)$. G will be assumed to be known throughout.

We now present some propositions for the motivation behind the parameter λ. Proposition 2.2 is due to Vidal et al. (2004). Proofs of the Propositions 2.1 and 2.3 will be given in the Appendix. The notation $H(x;\lambda)$ will be used to denote the cdf of S.

Proposition 2.1. *The family of cdfs* $\{H(x;\lambda),\ \lambda \in \mathbb{R}\}$ *is stochastically increasing with* $\lim_{\lambda\to\infty} H(x;\lambda)$ *as the right half normal cdf* $[2\Phi(x) - 1]I(x > 0)$, *and* $\lim_{\lambda\to-\infty} H(x;\lambda)$ *as the left half normal cdf* $2\Phi(x)I(x < 0)$.

Proposition 2.2. *The L_1-distance between $H(x;\lambda)$ and the normal cdf $\Phi(x)$ is given by*

$$L_1(\Phi, H) = \sup_{A\in\mathcal{B}} |P(A|H) - P(A|\Phi)| = 2\int_0^\infty \phi(x)G(|\lambda|x)dx - \frac{1}{2}$$

where $\mathcal{B}$ denotes the Borel set of subsets of $\mathbb{R}$, and $P(A|\cdot)$ denotes the probability of A under H or Φ.

Proposition 2.3. *The Kullback-Liblier distance between $H(\cdot;\lambda)$ and $\Phi(\cdot)$ is given by*

$$KL(\Phi, H) = -\log 2 - \int_0^\infty \log(G(|\lambda|x)(1 - G(|\lambda|x)\phi(x)dx$$

Propositions 2.2 and 2.3 imply that the distance between the $N(0,1)$ and the skew-normal $SN(0,1,\lambda,G)$ is 0 if and only if $\lambda = 0$, and the distance increases as λ moves away from 0 toward $\pm\infty$. The Kullback-Liblier distance in fact increases to ∞ as $|\lambda| \to \infty$. This implies that a higher value of λ indicate a significant departure from normality. Proposition 2.1 also shows that non-parametric inference such as sign-test or sign-rank test can be used to make inference about the parameter λ (Lehmann, 1986). Also note that for $\lambda > 0$ ($\lambda < 0$), $P(S > 0) > 1/2$ ($< 1/2$), and it increases to 1 (0) as λ increases (decreases) to $+\infty$ ($-\infty$). For the pre-post treatment studies, these propositions show that the parameter λ brings extra flexibility to model treatment effect when the shift from pre to post need not be symmetric. Note that a high positive value of λ implies that a high majority of the population show an improvement; while a low negative value of λ implies that only a few minority of the population show an improvement.

3. Empirical Bayes Estimation

Let μ_i $(i = 1, ..., n)$ be the true unobservable score of the ith subject of the sample, and let X_i $(i = 1, ..., n)$ be the corresponding observable score. In a pre-post treatment study, μ_i can be considered as the true improvement while X_i as an observed improvement at a particular instance for the ith subject. It would be appropriate to model the true improvement with a skew normal distribution, i.e., $\mu_i \sim SN(\xi, \tau^2, \lambda, G)$, where the parameters ξ and τ would reflect the shift and spread in the improvement respectively, while λ would reflect how skew is the improvement. The observed variable X_i conditionally on μ_i can be assumed to

follow $N(\mu_i, \sigma^2)$, for some $\sigma > 0$. Thus the model can be viewed as $X_i = \mu_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$, $\mu_i \sim SN(\xi, \tau^2, \lambda, G)$, and where ε_i and μ_i are independent. The structure of the model is thus hierarchical with $X_i|\mu_i \sim N(\mu_i, \sigma^2)$, and $\mu_i \sim SN(\xi, \tau^2, \lambda, G)$. We will assume that σ^2 is known until the end of the section where we discuss how to handle the case of unknown σ^2. The reason for assuming σ^2 known is that when $G = \Phi$, the model is unidentifiable when σ^2 is unknown. To see this note that marginally, $X_i \sim SN(\xi, \tau^2 + \sigma^2, \lambda\tau/\sqrt{\sigma^2(1+\lambda^2)+\tau^2})$, see Azzalini (1985). Thus when σ^2 is unknown, four parameters ξ, τ^2, λ, and σ^2 produce only a three-dimensional parameter space. It is difficult to see for any other G if the model is identifiable when σ^2 is unknown, but we conjecture that this will be true for any G other than Φ.

We now present the empirical Bayesian methodology that requires the posterior distribution of μ_i, and the estimation of $\boldsymbol{\theta} = (\xi, \tau, \lambda)^T$.

3.1. Posterior Distribution of μ_i

Denoting $\mathbf{x} = (x_1, x_2, ..., x_n)^T$, and the density of μ_i by $\pi(\mu_i|\xi, \tau, \lambda)$, the posterior of μ_i is

$$\pi(\mu_i|\mathbf{x}, \boldsymbol{\theta}) \propto \frac{1}{\sigma}\phi(\frac{x_i - \mu_i}{\sigma})\pi(\mu_i|\xi, \tau, \lambda). \tag{3.1}$$

Here, the skew-normal density $\pi(\mu_i|\xi, \tau, \lambda)$ is given by

$$\pi(\mu_i|\xi, \tau, \lambda) = \frac{2}{\tau}\phi(\frac{\mu_i - \xi}{\tau})G(\lambda\frac{\mu_i - \xi}{\tau}). \tag{3.2}$$

Combining (3.1) and (3.2) yields

$$\pi(\mu_i|\mathbf{x}, \boldsymbol{\theta}) \propto \frac{1}{\sigma_*}\phi(\frac{\mu_i - \mu_*(x_i)}{\sigma_*})G(\lambda\frac{\mu_i - \xi}{\tau}), \tag{3.3}$$

where

$$\mu_*(x_i) = \frac{\tau^2}{\tau^2 + \sigma^2}x_i + \frac{\sigma^2}{\tau^2 + \sigma^2}\xi, \quad \text{and} \quad \sigma_*^2 = (\frac{1}{\sigma^2} + \frac{1}{\tau^2})^{-1}. \tag{3.4}$$

Note that the posterior density (3.3) is not a skew-normal density.

If $G = \Phi$, then a closed form solution of the posterior density (3.3) is available, and is given by

$$\pi(\mu_i|\mathbf{x}, \boldsymbol{\theta}) = [\Phi(\frac{\lambda(\mu_*(x_i) - \xi}{\sqrt{\tau^2 + \lambda^2\sigma_*^2}})]^{-1}\phi(\frac{\mu_i - \mu_*(x_i)}{\sigma_*})\Phi(\lambda\frac{\mu_i - \xi}{\tau}). \tag{3.5}$$

This follows from the identity $E[\Phi(hU + k)] = \Phi(k/\sqrt{1 + h^2})$, where $U \sim N(0, 1)$, (Azzalini, 1985).

3.2. Estimation of $\boldsymbol{\theta}$

We now derive the maximum likelihood estimate of $\boldsymbol{\theta}$ from the marginal distribution of $X = (X_1, X_2, ..., X_n)^T$. Following the EM-algorithm, by considering $\mu_i, i = 1, ..., n$ as missing values (see, Carlin and Louis, 1998), the likelihood of the complete data is given by

$$L_c \propto \frac{1}{\sigma^n \tau^n} \prod_{i=1}^{n} \phi(\frac{x_i - \mu_i}{\sigma}) \phi(\frac{\mu_i - \xi}{\tau}) G(\lambda \frac{\mu_i - \xi}{\tau}). \tag{3.6}$$

Denoting the $\hat{\boldsymbol{\theta}}^{(k-1)}$ as the estimated value at the $(k-1)^{th}$ iteration, the E-step yields

$$\begin{aligned} & E[\log L_c | \mathbf{x}; \hat{\boldsymbol{\theta}}^{(k-1)}] \\ & \quad = const. - \frac{n}{2} \log \tau^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} E[(\mu_i - x_i)^2 | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] \\ & \qquad - \frac{1}{2\tau^2} \sum_{i=1}^{n} E[(\mu_i - \xi)^2 | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] + \sum_{i=1}^{n} E[\log G(\lambda \frac{\mu_i - \xi}{\tau}) | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] \end{aligned} \tag{3.7}$$

The posterior expectation above is with respect to the posterior distribution (3.3) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}^{(k-1)}$. Now, the estimate $\hat{\boldsymbol{\theta}}^{(k)}$ for the M-Step is obtained by maximizing (3.7) with respect to τ^2, ξ, and λ. Differentiating (3.7) with respect to τ^2, ξ, and λ and equating to zero yield the following set of equations.

$$\tau^2 = \frac{1}{n} \sum_{i=1}^{n} E[(\mu_i - \xi)^2 | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] \tag{3.8}$$

$$\xi = \frac{1}{n} \sum_{i=1}^{n} E[\mu_i | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] - \tau\lambda \frac{1}{n} \sum_{i=1}^{n} E[p(\lambda \frac{\mu_i - \xi}{\tau}) | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] \tag{3.9}$$

$$\frac{1}{n} \sum_{i=1}^{n} E[p(\lambda \frac{\mu_i - \xi}{\tau})(\mu_i - \xi) | x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] = 0, \tag{3.10}$$

where $p(\cdot) = G'(\cdot)/G(\cdot)$. $(\hat{\tau}^{(k)}, \hat{\xi}^{(k)}, \hat{\lambda}^{(k)})$ are the solution for (τ, ξ, λ) of the above equations. Note that $\hat{\lambda}^{(k)} = 0$ is always a feasible solution, but it need not yield

the maximum likelihood. This happens because $\lambda = 0$ is an inflection point of the likelihood function. This phenomenon has been reported by several authors in the context of maximum likelihood method for the standard skew normal problem (Azzalini, 1985; Dalla Valle, 2004). Thus if $\hat{\lambda}^{(k)} = 0$ is a suspect, the initial guess of $\hat{\lambda}^{(0)}$ close to 0 should be avoided. We recommend the initial guess to be the method of moments estimates. The method of moments estimates can be exactly computed when $G(\cdot) = \Phi(\cdot)$. In this case, as we noted earlier, the marginal distribution of X_i is $SN(\xi, \tau^2+\sigma^2, \lambda\tau/\sqrt{\sigma^2(1+\lambda^2)+\tau^2})$. The first three moments of this distribution are given by $\mu_1' = \xi + \sqrt{\frac{2}{\pi}}\delta\tau,\ \mu_2' = \xi^2 + 2\sqrt{\frac{2}{\pi}}\xi\delta\tau + \tau^2 + \sigma^2,\ \mu_3' = \xi^3 + 3\sqrt{\frac{2}{\pi}}\xi^2\delta\tau + 3\xi(\tau^2+\sigma^2) + \sqrt{\frac{2}{\pi}}\delta\tau(3\sigma^2 + (3-\delta^2)\tau^2)$, where $\delta = \lambda/\sqrt{1+\lambda^2}$. The method of moments from these can be easily obtained. Even when $G(\cdot) \neq \Phi(\cdot)$, it would be reasonable to use these estimates as the initial guess in the EM algorithm.

The posterior expectations in (3.8) - (3.10) can be computed by generating normal random variates as follows. From (3.3), for any measurable function $q(\cdot)$,

$$
\begin{aligned}
&E[q(\mu_i)|x_i; \hat{\boldsymbol{\theta}}^{(k-1)}] \\
&\quad = \frac{1}{c_i^{(k-1)}} \int_{-\infty}^{\infty} q(\mu_i) \frac{1}{\hat{\sigma}_*^{(k-1)}} \phi\Big(\frac{\mu_i - \hat{\mu}_*^{(k-1)}(x_i)}{\hat{\sigma}_*^{(k-1)}}\Big) G\Big(\hat{\lambda}^{(k-1)} \frac{\mu_i - \hat{\xi}^{(k-1)}}{\hat{\tau}^{(k-1)}}\Big) d\mu_i \\
&\quad = \frac{1}{c_i^{(k-1)}} E\Big[q(N_{i(k-1)}) G\Big(\hat{\lambda}^{(k-1)} \frac{N_{i(k-1)} - \xi^{(k-1)}}{\hat{\tau}^{(k-1)}}\Big)\Big] \qquad (3.11)
\end{aligned}
$$

where

$$
\begin{aligned}
c_i^{(k-1)} &= E\Big[G\Big(\hat{\lambda}^{(k-1)} \frac{N_{i(k-1)} - \xi^{(k-1)}}{\hat{\tau}^{(k-1)}}\Big)\Big], && (3.12) \\
N_{i(k-1)} &\sim N(\hat{\mu}_*^{(k-1)}(x_i),\ \hat{\sigma}_*^{(k-1)2}), && (3.13) \\
\hat{\mu}_*^{(k-1)}(x_i) &= \frac{\hat{\tau}^{(k-1)2}}{\hat{\tau}^{(k-1)2} + \sigma^2} x_i + \frac{\sigma^2}{\hat{\tau}^{(k-1)2} + \sigma^2} \xi^{(k-1)}, && (3.14) \\
\sigma_*^{(k-1)2} &= \Big(\frac{1}{\sigma^2} + \frac{1}{\hat{\tau}^{(k-1)2}}\Big)^{-1} && (3.15)
\end{aligned}
$$

Thus, if $N_{i(k-1)}$ is generated with M copies $\{N_{i(k-1)}^{(j)}, j = 1, ..., M\}$ for suffi-

ciently large M, then

$$E[q(\mu_i)|x_i;\hat{\boldsymbol{\theta}}^{(k-1)}] \approx \sum_{j=1}^{M} w_{ij}^{(k-1)} q(N_{i(k-1)}^{(j)}), \tag{3.16}$$

where

$$w_{ij}^{(k-1)} = \frac{G(\hat{\lambda}^{(k-1)} \frac{N_{i(k-1)}^{(j)} - \xi^{(k-1)}}{\tau^{(k-1)}})}{\sum_{j=1}^{M} G(\hat{\lambda}^{(k-1)} \frac{N_{i(k-1)}^{(j)} - \xi^{(k-1)}}{\tau^{(k-1)}})}. \tag{3.17}$$

Note that $\{w_{ij}^{(k-1)}, j = 1, ..., M\}$ can be considered as weights given to the normal variates to adjust the skewness. Here, $\{N_{i(k-1)}^{(j)}, j = 1, ..., M\}$ need not be generated separately for each i and k; only one set of $N(0,1)$ variates $\{z_j, j = 1, ..., M\}$ need to be generated. $N_{i(k-1)}^{(j)}$ can be then taken as

$$N_{i(k-1)}^{(j)} = \frac{\hat{\tau}^{(k-1)2}}{\hat{\tau}^{(k-1)2} + \sigma^2} x_i + \frac{\sigma^2}{\hat{\tau}^{(k-1)2} + \sigma^2} \hat{\xi}^{(k-1)} + (\frac{1}{\sigma^2} + \frac{1}{\hat{\tau}^{(k-1)2}})^{-1/2} z_j.$$

Thus, based on (3.17), from (3.8)-(3.10), the solution for the updated $(\tau^{2(k)}, \xi^{(k)}, \lambda^{(k)})$ of the EM-algorithm can be computed as a solution of

$$\hat{\tau}^{2(k)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} w_{ij}^{(k-1)} (N_{i(k-1)}^{(j)} - \hat{\xi}^{(k)})^2 \tag{3.18}$$

$$\hat{\xi}^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} w_{ij}^{(k-1)} N_{i(k-1)}^{(j)} - \hat{\tau}^{(k)} \hat{\lambda}^{(k)} \sum_{i=1}^{n} \sum_{j=1}^{M} w_{ij}^{(k-1)} p(\hat{\lambda}^{(k)} \frac{N_{i(k-1)}^{(j)} - \xi^{(k)}}{\tau^{(k)}}) \tag{3.19}$$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} w_{ij}^{(k-1)} (N_{i(k-1)}^{(j)} - \hat{\xi}^{(k)}) p(\hat{\lambda}^{(k)} \frac{N_{i(k-1)}^{(j)} - \xi^{(k)}}{\tau^{(k)}}) = 0 \tag{3.20}$$

Numerical solution of the above equations can be obtained using iterative algorithms such as Newton-Raphson method.

3.3. Posterior Inference

Posterior inference based on empirical Bayesian methodology involves the estimation of $E[\mu_i|x_i], i = 1, ..., n+1$ or other posterior quantities, where μ_{n+1} and x_{n+1} are associated with a possible new observation. Using the empirical Bayes estimate of $\hat{\theta}$, from (3.3), the estimate of $E[\mu_i|\mathbf{x}]$ is given by

$$\hat{E}[\mu_i|x_i] = c_i^{-1} \int_{-\infty}^{\infty} \mu_i \frac{1}{\hat{\sigma}^*} \phi(\frac{\mu_i - \hat{\mu}_*(x_i)}{\hat{\sigma}^*}) G(\hat{\lambda}\frac{\mu_i - \hat{\xi}}{\hat{\tau}}) d\mu_i,$$

where

$$\hat{\mu}_*(x_i) = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2} x_i + \frac{\sigma^2}{\hat{\tau}^2 + \sigma^2} \hat{\xi} \tag{3.21}$$

$$\hat{\sigma}_* = (\frac{1}{\sigma^2} + \frac{1}{\hat{\tau}^2})^{-1/2}, \text{ and} \tag{3.22}$$

$$c_i = \int_{-\infty}^{\infty} \frac{1}{\hat{\sigma}^*} \phi(\frac{\mu_i - \hat{\mu}_*(x_i)}{\hat{\sigma}^*}) G(\hat{\lambda}\frac{\mu_i - \hat{\xi}}{\hat{\tau}}) d\mu_i. \tag{3.23}$$

where the estimates $(\hat{\xi}, \hat{\tau}, \hat{\lambda})$ are obtained through the EM algorithm as described above. It should be also noted that $\hat{E}[\mu_i|x_i]$ can be approximated by simulated $N(0,1)$ variates $\{z_j, j = 1, ..., M\}$ as

$$\hat{E}[\mu_i|x_i] \approx \sum_{j=1}^{M} w_{ij} \hat{N}_{ij} \tag{3.24}$$

where $\hat{N}_{ij} = \hat{\mu}_*(x_i) + \hat{\sigma}_* z_j$, and $w_{ij} = G(\hat{\lambda}\frac{\hat{N}_{ij} - \hat{\xi}}{\hat{\tau}}) / \sum_{j=1}^{M} G(\hat{\lambda}\frac{\hat{N}_{ij} - \hat{\xi}}{\hat{\tau}})$.

Note that when $\hat{\lambda} = 0$, the $\hat{E}[\mu_i|x_i]$ is same as the typical empirical Bayes rule; when $\hat{\lambda} > 0$, higher weights are assigned for high values of $\hat{N}_{ij}$ and lower weights for low values of $\hat{N}_{ij}$, and the reverse happens when $\hat{\lambda} < 0$.

Although, in the empirical Bayesian methodology, the main interest is about the posterior inference on $\mu_i, i = 1, ..., n$; however, in some applications such as in a pre-post treatment study, the estimate of the distribution $SN(\xi, \tau^2, \lambda, G)$ or perhaps the estimates of the quantities like $P(\mu_i > 0)$ can be useful to see the effect of the treatment. $SN(\xi, \tau^2, \lambda, G)$ can be estimated by $SN(\hat{\xi}, \hat{\tau}^2, \hat{\lambda}, G)$, and $P(\mu_i > 0)$ by $\hat{P}(\mu > 0) = 2 \int_0^{\infty} \frac{1}{\hat{\sigma}} \phi(\frac{x - \hat{\xi}}{\hat{\sigma}}) G(\hat{\lambda}\frac{x - \hat{\xi}}{\hat{\sigma}}) dx$.

3.4. Unknown σ^2 case

When σ^2 is unknown, as it will be the case in practice, an estimate of it is needed. In the presence of replications such estimate is possible. Suppose, $\{X_{ij}, j = 1, 2, ..., m_i\}$ are the repeated observations for each $i = 1, 2, ..., n$, then $X_{ij} = \mu_i + \varepsilon_{ij}$, where μ_i and ε_{ij} are independent, $\mu_i \sim SN(\xi, \tau^2, \lambda, G)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. Then $(\bar{X}_i = \frac{1}{n}\sum_{j=1}^{m_i} X_{ij}, i = 1, ..., n)$ and $S^2 = (\sum m_i - n)^{-1} \sum_{i=1}^{n}(m_i - 1)S_i^2$, where $S_i^2 = (m_i - 1)^{-1}\sum_{j=1}^{m_i}(X_{ij} - \bar{X}_i)^2$, are sufficient statistics. It can be seen that $\bar{X}_i, i = 1, ..., n$ and S^2 are independent, $\bar{X}_i|\mu_i \sim N(\mu_i, \sigma^2/m_i)$, and $(\sum m_i - n)S^2 \sim \sigma^2\chi^2_{(\sum m_i - n)}$. As suggested by Berger (1985) for the normal populations, the method discussed as above can be used by replacing σ^2 by s^2/n, where s^2 is a realization of S^2. A direct empirical Bayes estimator of σ^2 can also be obtained along the lines of EM algorithm as described above by replacing σ^2 by σ^2/n in (3.6), and by multiplying the density of S^2 in it.

4. Hierarchical Bayesian Methodology

Consider the following hierarchical structure.

$$x_i|\mu_i \sim N(\mu_i, \sigma^2); \quad \mu_i \sim SN(\xi, \tau^2, \lambda, G); \ (\xi, \tau, \lambda) \sim \pi(\xi, \tau, \lambda) \tag{4.1}$$

where $\pi(\xi, \tau, \lambda)$ is a given prior distribution. We assume that σ^2 is known. The unknown σ^2 case will be discussed at the end of this section. Of particular interests are the posterior distributions of μ_i and the posterior distribution of ξ and λ. In a pre-post treatment study, the posterior of ξ and λ would reflect the nature of the treatment effects. The posterior of ξ would reflect the overall treatment effect; while the posterior of λ would reflect, through the skewness, the extent of majority (or minority) enjoying the effect of the treatment. The direct derivation of the marginal posterior of μ_i, ξ, and λ would be complicated; however, the Gibbs sampler can be used to generate the samples of the posterior. In order to implement the Gibbs sampler, we use the data augmentation technique (Tanner, 1996). Note that from the third representation result of the skew normal distribution as discussed in section 2, $\mu_i = \xi + \tau q_i z_i$, where $z_i \sim N(0, 1)$, and where conditionally on z_i, $q_i = 1$ with probability $G(\lambda z_i)$, and $q_i = -1$ with probability $1 - G(\lambda z_i) = G(-\lambda z_i)$. Thus, from (4.1), the joint posterior density of $(\xi, \tau, \mathbf{q}, \mathbf{z})$,

where $\mathbf{q} = (q_1, q_2, ..., q_n)^T$ and $\mathbf{z} = (z_1, z_2, ..., z_n)^T$, is

$$\pi(\xi, \tau, \mathbf{q}, \mathbf{z}) \propto \prod_{i=1}^{n} \frac{1}{\sigma} \phi(\frac{x_i - \xi - \tau q_i z_i}{\sigma}) \phi(z_i) [G(\lambda z_i)]^{\frac{1+q_i}{2}} [G(-\lambda z_i)]^{\frac{1-q_i}{2}} \pi(\xi, \tau, \lambda). \tag{4.2}$$

If ξ, τ^2, and λ are apriori independent, i.e., $\pi(\xi, \tau, \lambda) = \pi_1(\xi)\pi_2(\tau)\pi_3(\lambda)$ for some priors π_1, π_2, and π_3, then the full posterior conditionals are given by

$$\pi_1(\xi|\tau, \lambda, \mathbf{z}, \mathbf{q}; \mathbf{x}) \propto \phi(\sqrt{n}\frac{\xi - \bar{x} - \tau\overline{qz}}{\sigma})\pi_1(\xi)$$

$$\pi_2(\tau|\lambda, \mathbf{z}, \mathbf{q}; \mathbf{x}) \propto \prod_{i=1}^{n} \phi(\frac{x_i - \xi - \tau q_i z_i}{\sigma})\pi_2(\tau)$$

$$\pi_3(\lambda|\xi, \tau, \lambda; \mathbf{x}) \propto \prod_{i=1}^{n} [G(\lambda z_i)]^{\frac{1+q_i}{2}} [G(-\lambda z_i)]^{\frac{1-q_i}{2}} \pi_3(\lambda)$$

$$\pi_{4i}(z_i|\xi, \tau, \lambda, \mathbf{z}_{(-i)}, \mathbf{q}; \mathbf{x}) \propto \phi(\sqrt{1 + \frac{\tau^2}{\sigma^2}}(z_i - \frac{\tau q_i (x_i - \xi)}{\sigma^2 + \tau^2}))[G(\lambda z_i)]^{\frac{1+q_i}{2}} [G(-\lambda z_i)]^{\frac{1-q_i}{2}}$$

$$\pi_{5i}(q_i = 1|\xi, \tau, \lambda, \mathbf{z}, \mathbf{q}_{(-i)}; \mathbf{x}) = d(\xi, \lambda, \tau, z_i; \mathbf{x}) \exp(z_i \tau (x_i - \xi)/\sigma^2) G(\lambda z_i)$$

$$\pi_{5i}(q_i = -1|\xi, \tau, \lambda, \mathbf{z}, \mathbf{q}_{(-i)}; \mathbf{x}) = d(\xi, \lambda, \tau, z_i; \mathbf{x}) \exp(-z_i \tau (x_i - \xi)/\sigma^2) G(-\lambda z_i),$$

where $\overline{qz} = n^{-1} \sum q_i z_i$, $d(\xi, \lambda, \tau, z_i; \mathbf{x}) = [\exp(z_i\tau(x_i - \xi)/\sigma^2)G(\lambda z_i) + \exp(-z_i\tau(x_i - \xi)/\sigma^2)G(-\lambda z_i)]^{-1}$, and notations $\mathbf{z}_{(-i)}$, and $\mathbf{q}_{(-i)}$ respectively are used to denote all $z_j, j = 1, ..., n$ except z_i, and all $q_j, j = 1, ..., n$ except q_i. Note that if $\pi_2(\xi)$ is normal or $\pi_2(\xi) \propto const.$, then the conditional posterior of ξ is normal; and if $\pi_2(\tau) \propto const.$, then the conditional posterior of τ is truncated normal.

4.1. Unknown σ^2 case

Under the repeated measurements $\{X_{ij}, j = 1, 2, ..., m_i\}, i = 1, 2, ..., n$, using the data augmentation technique as above, it can be seen that the posterior of $(\xi, \tau, \lambda, \sigma^2, \mathbf{q}, \mathbf{z})$ is given by

$$\pi(\xi, \tau, \sigma^2, \mathbf{q}, \mathbf{z}) \propto \prod_{i=1}^{n} \frac{1}{\sigma^{m_i}} \phi(\sqrt{m_i}\frac{\bar{x}_i - \xi - \tau q_i z_i}{\sigma}) \exp(\frac{(\sum m_i - n)S^2}{\sigma^2}) \phi(z_i)[G(\lambda z_i)]^{\frac{1+q_i}{2}} [G(-\lambda z_i)]^{\frac{1-q_i}{2}} \pi(\xi, \tau, \sigma^2, \lambda)$$

Gibbs sampler can then be performed by obtaining the full conditional posterior in the same manner as described above.

5. An Example

As an example, we consider the data from a pre-post study on 13 neurosurgical patients who underwent thalamic chronic electrode implants as a treatment for dyskinesia and chronic pain (Bhatnagar and Mandybur, 2005). The purpose of the study was to study if the electrical stimulation of a specific thalamic nucleus such as the ventrolateral nucleus (VL) has any effect on the patients' language and cognitive processing. The hypothesis of the study was that the stimulation of the VL improves the language and speech impairment. The subjects were assessed before and after the stimulation on a detailed neurolinguistic test battery which included subjects recalling and naming pictures of slides shown for four seconds. One of the observed variables was the number of seconds it took to recall and naming the pictures. The results of this variable are presented in the table below.

Table 1

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13
Before	25.00	15.02	25.50	19.30	22.49	18.90	24.91	23.17	18.13	26.70	25.30	18.80	26.18
After	17.82	14.00	18.23	16.30	15.93	14.50	19.25	18.00	16.30	19.70	18.50	14.00	11.90

The improvement after the stimulation, i.e., the differences between before and after are (in seconds) {7.18, 1.02, 7.27, 3.00, 6.56, 4.40, 5.66, 5.17, 1.83, 7.00, 6.80, 4.80, 14.28}. Note that one patient had a significant improvement (14.28 seconds); however, it cannot be considered as an outlier since this kind of improvement is anticipated in some patients. We should also point out that removing this patient did not significantly alter the results as far as skewness is concerned.

The empirical Bayes methodology and Hierarchical Bayes analysis developed in section 3 and section 4 were performed on the differences. The distribution function $G(x) = \exp(x)/(1 + \exp(x))$ was used throughout, and σ^2 which can be termed as the within subject variability was assumed to be 1 after the consultation with a researcher of the study (a very high within subject variability was not expected). As initial values of the EM-algorithm, we used the method of moments estimates based on $G(\cdot) = \Phi(\cdot)$. All the calculations of the empirical Bayes methodology was done using Mathematica 5.0 (Wolfram Research Inc.), and Hierarchical Bayesian analysis was done using WinBugs 1.4.1.

Method of moments estimates of (ξ, τ, λ) are $(1.35, 5.33, 2.25)$. Using these estimates as the the initial estimates for the EM steps, we obtained the empirical Bayes estimates of (ξ, τ, λ) from (3.18) - (3.20) using $M = 150$, which are given by

$(2.42, 4.47, 5.48)$. The EM algorithm converged after 10 iterations. We did noticed that the solution to (3.18) - (3.20) is sensitive to the initial values. High initial value of λ yielded diverging sequence as it has been observed by several authors for the maximum likelihood estimates (see, for example, Dalla Valle(2004)). Note that the estimate of λ is very high, indicating a strong evidence of skewness.

The empirical Bayes estimate of μ_{13} corresponding to $x_{13} = 14.28$, from (3.24) using $M = 150$, is 13.66 with a 95% confidence interval $(11.30, 16.01)$. The empirical Bayes estimate of μ_7, corresponding to $x_7 = 5.66$, is 5.51 with a 95% confidence interval $(3.28, 7.73)$. The confidence intervals here are approximate based on the approximation $(\hat{E}[\mu_i|x_i] - 1.96\hat{V}[\mu_i|x_i],\ \hat{E}[\mu_i|x_i] + 1.96\hat{V}[\mu_i|x_i])$, where $\hat{E}[\mu_i|x_i]$ is given by (3.24) and $\hat{V}[\mu_i|x_i]$ is computed similarly as $\hat{V}[\mu_i|x_i] = \sum_{j=1}^{M} w_{ij}(\hat{N}_{ij} - \hat{E}[\mu_i|x_i])^2$, where $w_{ij} = G(\hat{\lambda}\frac{\hat{N}_{ij}-\hat{\xi}}{\hat{\tau}})/\sum_{j=1}^{M} G(\hat{\lambda}\frac{\hat{N}_{ij}-\hat{\xi}}{\hat{\tau}})$. The exact confidence region can be computed using, for example, the highest posterior density region from (3.3); see, Maritz and Lwin (1989), chapter 6.

We, now, contrast the above estimates with the empirical Bayes estimator under normal distribution, i.e., when $\lambda = 0$. Under normal distribution, $\tilde{E}[\mu_i|x_i] = \frac{\tilde{\tau}^2}{\tilde{\tau}^2+\sigma^2}x_i + \frac{\sigma^2}{\tilde{\tau}^2+\sigma^2}\tilde{\xi}$, and $\tilde{V}[\mu_i|x_i] = (\frac{1}{\sigma^2}+\frac{1}{\tilde{\tau}^2})^{-1}$, where $\tilde{\xi} = \bar{x}$, and $\tilde{\tau}^2 = \max(n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sigma^2,\ 0)$ (Berger, 1985). The empirical Bayes estimate of μ_{13} under normal distribution is 13.49 with a 95% confidence interval $(11.62, 15.35)$, while the empirical Bayes estimate of μ_7 under normal distribution is 5.67 with a 95% confidence interval $(3.80, 7.54)$. Note that the empirical Bayes estimates under the normal distribution underestimates the μ_{13}, while it overestimates the μ_7. This is perhaps due to the fact that the empirical estimates under the normal distribution is always shrinking the observations towards the mean $\bar{x}$ at the same rate.

As we pointed out in section 3 that the distribution of μ_i, $SN(\xi, \tau, \lambda, G)$, with ξ, τ, and λ replaced by its estimates, by itself, can shed some light on the nature of the improvement from pre to post. Fig 1 shows the probability density function of $SN(\hat{\xi}, \hat{\tau}, \hat{\lambda}, G)$, where $\hat{\xi}, \hat{\tau}$, and $\hat{\lambda}$ are the empirical Bayes estimates. $P(\mu_i > 0)$ can be estimated from $SN(\hat{\xi}, \hat{\tau}, \hat{\lambda}, G)$, which is 0.9999. This value indicates the proportion of subjects having positive improvement.

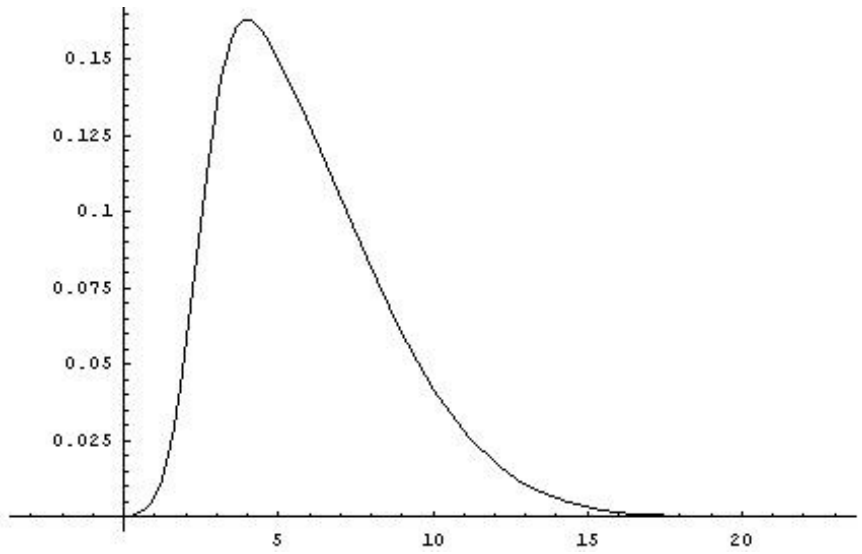

Fig 1

Hierarchical Bayesian analysis of section 4 was performed by WinBugs. We considered the following prior on (ξ, λ, τ): $N(0, 10^6)$ for ξ, $N(0, 100)$ for λ, and $U(0, 100)$ for τ with the assumption that ξ, λ, and τ are independent. Gibbs sampler produced samples of all the augmented data $\mathbf{z}$ and $\mathbf{q}$, and the parameters (ξ, τ^2, λ) and all $\mu_i s$. We only report the posterior posterior distributions of ξ, λ, μ_{13}, and μ_7. 20,000 samples were generated with every 10th observation recorded. Two chains were constructed in WinBugs. Both chains conveged after the burnout period of 300. See, Fig 2 for the posterior density of ξ, λ, μ_7, and μ_{13}. The mean of the posterior of λ was 7.12 indicating a strong evidence of skewness, while the mean of the posterior of ξ was 2.896. The posterior of μ_7 had a mean of 5.548 with 95% confidence limit of (3.686, 7.542), while the posterior of μ_{13} had a mean of 13.72 with 95% confidence limits of (11.82, 15.66).

The hierarchical Bayes estimates are close to the empirical Bayes estimates. Considering the difficulties is maximizing the marginal likelihood, it may be advantageous to use the hierarchical Bayes methodology than the empirical Bayes methodology. Also, note that the hierarchical Bayes confidence intervals are shorter than the empirical Bayes confidence intervals. This may be perhaps since we did not use the highest posterior density confidence intervals in the empirical Bayes methodology.

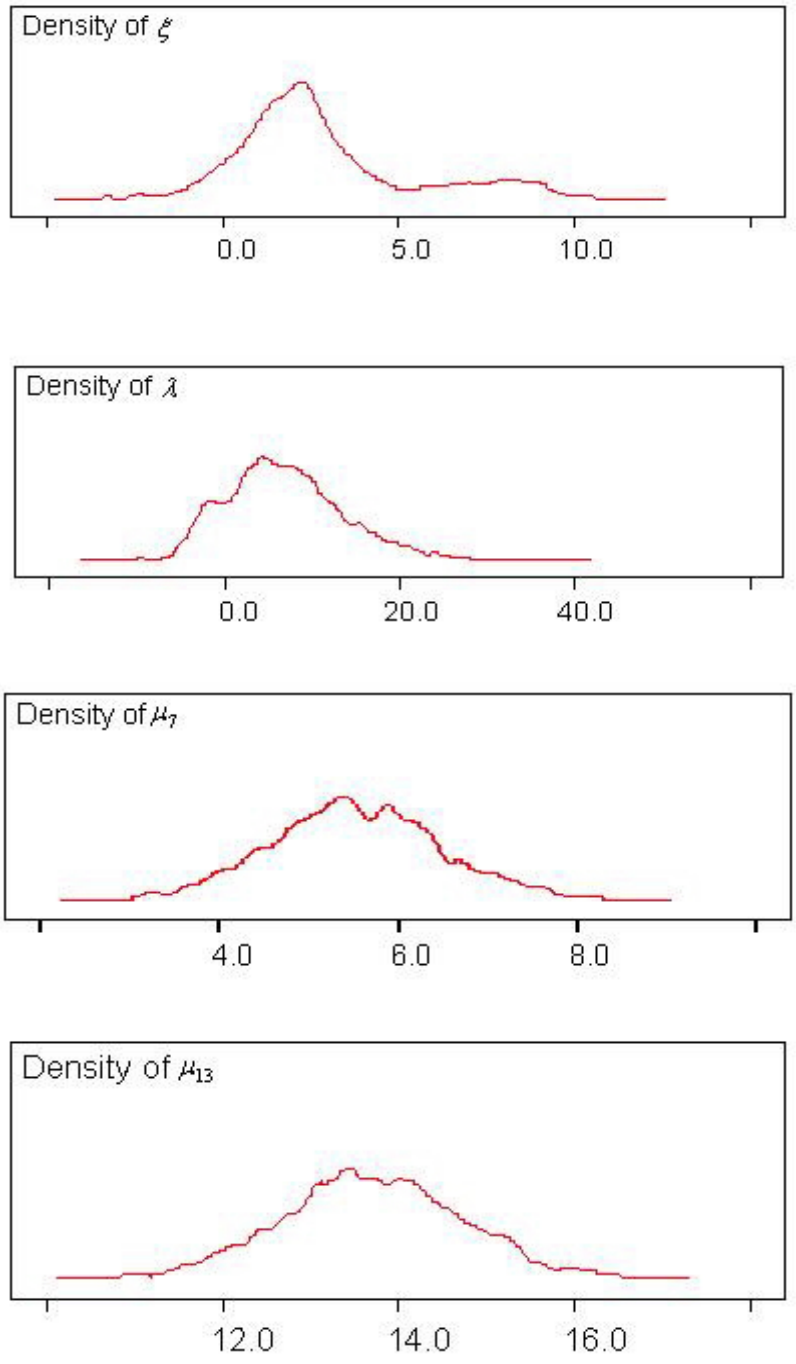
Density of ξ
0.0
5.0
10.0
Density of λ
0.0
20.0
40.0
Density of μ_7
4.0
6.0
8.0
Density of μ_{13}
12.0
14.0
16.0

Fig 2

6. Appendix

6.1. Proof of Proposition 2.1

It is clear that for any $x < 0$,

$$H(x;\lambda) = 2\int_{-\infty}^{x} G(\lambda y)\phi(y)dy$$

decreases as λ increases, and $\lim_{\lambda\to-\infty} H(x;\lambda) = 2\Phi(x)$ and $\lim_{\lambda\to\infty} H(x;\lambda) = 0$. For any $x > 0$, since $\phi(-y) = \phi(y)$, and $G(-\lambda y) = 1 - G(\lambda y)$,

$$\begin{aligned} H(x;\lambda) &= 2\int_{-\infty}^{0} G(\lambda y)\phi(y)dy + 2\int_{0}^{\infty} G(\lambda y)\phi(y)dy \\ &= 2\int [1 - G(\lambda y)]\phi(y)dy + 2\int_{0}^{\infty} G(\lambda y)\phi(y)dy \\ &= 1 - 2\int_{x}^{\infty} G(\lambda y)\phi(y)dy, \end{aligned}$$

which decreases as λ increases, and $\lim_{\lambda\to-\infty} H(x;\lambda) = 1$ and $\lim_{\lambda\to\infty} H(x;\lambda) = 2\Phi(x) - 1$.

6.2. Proof of Proposition 2.3

The Kullback-Liblier distance between $H(\cdot;\lambda)$ and $\Phi(\cdot)$ is given by

$$KL(\Phi, H) = -\int_{-\infty}^{\infty} \log(2G(\lambda x))\phi(x)dx$$

Using the symmetry of $\phi(x)$, and since $G(-\lambda x) = 1 - G(\lambda x)$, it can be seen that

$$KL(\Phi, H) = -\log 2 - \int_{0}^{\infty} \log(G(|\lambda|x)(1 - G(|\lambda|x)\phi(x)dx$$

REFERENCES

Arellano-Valle, R. B., Gomez, H. W., and Quintana, F. A. (2004). "A new class of skew-normal distributions," *Comm. Statist. Theory Methods*, **33**, No. 7, 1465-1480.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A., and Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrica,* **58**, 471- 478.

Azzalini, A. (1985). "A class of distributions which includes the normal ones," *Scand. J. Statist.*, **12**, 171-178.

Azzalini, A. (1986). "Further results on a class of distributions which includes the normal ones," *Statistica*, **46**, 199-208.

Azzalini, A., and Dalla Valle, A. (1996). "The multivariate skew-normal distribution," *Biometrika* **83**, 715-726.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer-Verlag.

Bhatnagar, S. C., and Mandybur, G. T. (2005). "Effects on intralaminar thalamic stimulation on language functions," *Brain and Language*, **92**, 1-11.

Branco, M. D., and Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.*, **79**, 99-113.

Carlin, B. P., and Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*. Monographs on Statistics and Applied Probability, 69. London: Chapman & Hall.

Dalla Valle, A. (2004). "The skew-normal distributions." In *Skew-Elliptical Distributions and their Applications, A Journey Beyond Normality*, M. G. Genton (ed.), Boca Raton, FL: Chapman & Hall/CRC.

Genton, M. G. (2004). Editor. *Skew-Elliptical Distributions and their Applications, A Journey Beyond Normality*. Boca Raton, FL: Chapman & Hall/CRC.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition. New York: Wiley.

Loperfido, N. M. R. (2004) "Generalized skew-normal distributions." In *Skew-Elliptical Distributions and their Applications, A Journey Beyond Normality*, M. G. Genton (ed.), Boca Raton, FL: Chapman & Hall/CRC.

Maritz, J. S., and Lwin, T. (1989). *Empirical Bayes Methods*, 2nd edition. Monographs on Statistics and Applied Probability, 35. London: Chapman & Hall.

Tanner, M.A. (1996). *Tools for Statistical Inference* (3rd ed.), New York: Springer.

Vidal, I., Iglesias, Branco, M. D., and Arellano-Valle, R. B. "Bayesian sensitivity analysis and model comparison for skew elliptical models," Preprint.