UNIVERSITY OF SOUTHAMPTON

Bayesian Modelling with Skew-Elliptical Distributions

by

High Seng CHAI

Thesis submitted for the degree of Doctor of Philosophy

in the

Faculty of Engineering, Science and Mathematics
School of Mathematics

November 2004

UNIVERSITY OF SOUTHAMPTON
FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF MATHEMATICS

Doctor of Philosophy

Bayesian Modeling with Skew-Elliptical Distributions

by

High Seng CHAI

ABSTRACT

The dissertation is devoted to modeling with a new class of multivariate skew elliptical distributions. This family of distributions extends the elliptical ones by the addition of a vector of shape parameters. It contains the multivariate skew normal, skew Student's t and skew Cauchy as special cases.

Detailed exploration is confined to the case of the univariate skew normal distribution. In particular, salient properties of the density are studied and comparisons are drawn with alternative skew normal proposals. Applications considered include linear regression, variance components and survival models. Bayesian analysis with these models are shown to be easily accomplished through the use of the Gibbs sampler. The latter proves very straightforward to specify distributionally and to implement computationally. Numerical examples show that skew normal modeling is a viable competitor to the celebrated normal theory methods.

Contents

List of Figures

List of Tables

Acknowledgements

I owe a debt of gratitude to those who helped make this thesis possible. Dr. Sujit Sahu deserves special thanks for his time, insight and support in supervising this research. The thesis had benefited greatly from his invaluable guidance and suggestions. My girlfriend Xinyun Zhu patiently read a preliminary version of the manuscript and made many constructive and detailed comments for its improvement. I would like to express my deep appreciation for her understanding and encouragement during all this time. I also wish to record my thanks to staffs and many students in the Faculty, working with them gave me a broader vision of the field which is hopefully incorporated in this thesis. My special thanks goes to Ralph Manson for efficient computational help in *LaTex*. The financial support from the Development Trust Bursary and the Faculty through a doctoral fellowship is gratefully acknowledged. Finally, I am grateful to my parents for their love and support during the preparation of this thesis.

Chapter 1

Introduction

1.1 Aims and motivations

Statistical analysis on the treatment of continuous observations within a parametric approach is usually proceeded by assuming:

(i) simplicity of the structure for the mean of the data,

(ii) constancy of error variability, and

(iii) normality of error distributions.

The requirement of assumptions (i) and (ii) aims both to allow an efficient analysis and to achieve ease of understanding. A typical example of (i) is the assumption of additivity. Assumption (iii) is mainly driven by the formal properties of the normal distribution, in particular its analytical beauty and also the simplicity when dealing with fundamental operations like marginalization, conditioning and linear combinations. The other reason of imposing the normality assumption is that the outcomes of the experiment are usually expected to obey the central limit theorem, thereby resulting in approximately normally distributed observations.

In general terms, there are two ways of dealing with data which do not satisfy the above assumptions. The first one is to develop new methods of analysis with assumptions which fit the data in its original scale. The second most commonly adopted approach, however, is to bend the data in order that

assumptions (i), (ii) and (iii) are approximately satisfied by making a monotonic non-linear transformation. The customary purposes of transformation are threefold, but the primary motivation of transformation has tended to be on obtaining normality so as to exploit the unrivalled mathematical tractability of the normal distribution. Nevertheless, there have often been doubts, reservations and criticisms about the use of transformation for normality for two major reasons. Firstly, in multivariate setting, transformations are usually carried out on each component separately. Thus the appropriateness of the joint normality assumption is highly questionable. Secondly, the requirement for variance stability or simplicity in the mean surface often demands a transformation which is different from that for achieving normality. Therefore it seems too demanding to accomplish three goals simultaneously by means of transformation alone.

Should there be a conflict between the requirements for normality and for model simplicity (e.g. improving additivity and homoscedasticity), it is best to pay most attention to the latter to allow for ease of description and interpretation. Hence less restrictive families of distributions that can accommodate asymmetry and non-normal peakedness and allow a continuous departure from normality to non-normality can be valuable in analyzing non-Gaussian data. The aim of this dissertation is to present an extended version of the skew-elliptical distribution and to assess its potential in practical implementations. This new class of multidimensional distributions has reasonable flexibility in distributional shape. Its ability to account for practical values of skewness and kurtosis means that more weight can be given to the considerations of assumptions (i) and (ii) in real data fitting. For simplicity of exposition, detailed development is confined to the skew-normal case in its univariate settings.

1.2 Skewed distributions

Rapid advances in computing technology in conjunction with the development of Markov chain Monte Carlo simulation methods render data fitting using increasingly flexible models a real practical possibility. This ability naturally leads us to wish to posit and develop more realistic distributions for statistical

analysis. Modeling with what are known as skewed distributions has sparked considerable attention since the publication of the pioneering paper by Azzalini (1985). The families are useful for analyzing unimodal empirical data with possible skewness present. As a consequence, data modeling can now be performed without the need for ad-hoc transformations to symmetry. In other words, the skewed distributions offer an appealing alternative to symmetric distributions (e.g the normal, t and logistic distributions) frequently employed in linear models.

One such class of skewed distributions was proposed by Sahu, Dey and Branco in 2003. They introduced asymmetry into elliptical distributions through simple transformation and conditioning techniques. Their setup indicates that skewness in the case of univariate distributions is regulated solely by a single parameter. It is instructive to extend the generating mechanism. In this thesis, we instead employ a vector of 'skewness' parameters to control the shape of the resultant univariate densities. Additional flexibility in modeling kurtosis present in the data may be anticipated from this generalization.

1.3 Overview of Bayesian inference

Data analysis techniques used within this research are based on the Bayesian paradigm. A Bayesian approach to data analysis typically involves the following four steps.

1. Specify a parametric model that more or less describes the phenomena underlying the collected data.

2. Formulate a prior distribution for the unknown model parameters. This density reflects our genuine beliefs about the parameters before seeing the data.

3. Obtain the posterior distribution of the model parameters via Bayes theorem. Currently, all knowledge about the parameters available from the prior and the data is represented in the posterior density.

4. Draw inferences from this updated information.

In essence, Bayesian inference is the science of making conclusions about a random process using information observed from that process. Utilizing probability as the fundamental measure for all forms of uncertainty is one of the attractive features of Bayesian methods.

The form of the posterior distribution can be rather complex. Subsequent technical difficulties in carrying out the requisite calculation for inference have long served as an impediment to the implementation of Bayesian statistics. Although many techniques have been developed, this thesis provides an approximate solution for such calculations through a sampling-based approach called the Gibbs sampler.

1.4 Outline of thesis

The rest of the document is set out as follows. Chapter 2 covers the relevant background information on the Bayesian approach to statistics. This includes the basic concepts, some model selection methods, and Markov chain Monte Carlo computational techniques.

Chapter 3 introduces the new skew elliptical distributions, with special emphasis on the univariate skew normal case. Specifically, mathematical expressions for the moments of the skew normal density are derived, the main properties of the density are stated, and comparisons are drawn with alternative proposals.

Chapters 4 – 6 focus on data modeling aspects of the skew normal distribution. We consider linear regression modeling under skew normal errors in Chapter 4. Variance components models based on skew normal random effects are studied in Chapter 5. Lastly, Chapter 6 examines the potential applications of the univariate skew normal density in survival analysis.

The closing chapter, Chapter 7, contains final thoughts on the completed work and some recommendations for future development. A short appendix provides some descriptions of two freely available software packages – `BUGS` and `CODA`.

Chapter 2

The Bayesian approach to statistical inference

2.1 Introduction

The main purpose of statistical inference is to make conclusions or inferences about a population from a sample drawn from that population. In the Bayesian approach, inferences are based on the conditional probability distribution of all unobserved quantities about which we wish to learn, given the observed sample. The most fundamental characteristic of Bayesian methods is that all unobserved quantities whether they are observable, e.g. future observations, or unobservable, e.g. model parameters, are treated as random variables. Thus Bayesian statistical conclusions about these unobserved quantities are made in terms of probability statements. The explicit use of probability to quantify uncertainty in inferences is one of the primary motivations for using Bayesian methods.

Our intention in this chapter is to provide a review of several elementary concepts and methods involved in the Bayesian approach to inference. These methods will be subsequently used in later chapters of this thesis. The remainder of this chapter is organized as follows. We begin by giving an overview of the fundamentals of Bayesian statistical analysis in Section 2.2. Some Bayesian ways of making prediction and model selection are discussed in Sections 2.3 and 2.4, respectively. The chapter then proceeds with Section 2.5 on two common

approaches to the associated problem of prior selection. We shall focus on the issues of summarizing posterior information in Section 2.6. Section 2.7 describes some techniques devoted to performing Bayesian inference using Markov Chain Monte Carlo simulation.

2.2 The posterior distributions

The usual starting point of statistical inference is the assumption that data are modeled from a distribution which is exchangeable. The random quantities $Y_1, \cdots, Y_n$ are said to be exchangeable if their joint probability density $p(y_1, \cdots, y_n)$ is invariant under permutation of the subscripts. A special case of exchangeability is independently and identically distributed (iid) sequences. Throughout we should regard our observed data $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ as iid given θ, the unobserved model parameter.

In its basic form, a full Bayesian model is constructed by using two ingredients: a prior density $p(\theta)$ and a sampling density $p(\mathbf{y}|\theta)$. The first ingredient tells us what is known about the distribution of θ before observing $\mathbf{y}$. This information is based on previous experience and understanding about similar experimentation. The second ingredient determines how the probabilities of different values of $\mathbf{Y}$ are distributed conditionally on θ. Thinking of $p(\mathbf{y}|\theta)$ as a function of θ gives the likelihood function for θ, which represents the information about θ coming from the observed data. Hence it is only sensible that all inferences are based on the updated distribution of θ by incorporating both historic and data information. This distribution is called the posterior distribution of θ denoted henceforth by $p(\theta|\mathbf{y})$.

The Bayes theorem expresses the updated probability statement about the unobserved θ as

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta}. \tag{2.1}$$

Notice that the left hand side of (2.1) is a density for θ and any factor in the right hand side which does not depend on θ can be considered as constant. Therefore a more compact form of the Bayes theorem is

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta). \tag{2.2}$$

In words, the posterior density is proportional to the likelihood multiplied by the prior density. As described above, Bayesian inference is entirely based on the posterior distribution which contains all current information about θ. This means that the data $\mathbf{y}$ affect the posterior inference only through the likelihood function $p(\mathbf{y}|\theta)$. More formally, Bayesian inference procedures obey the so-called likelihood principle, which states that all relevant information brought by a given sample of data is entirely contained in the likelihood function. Furthermore, the same inference should be made from two different sampling experiments if their likelihood functions are proportional to each other (as functions of θ).

The Bayes theorem provides an easy mechanism by which sequential analysis can be performed. For example, suppose that after observing $\mathbf{y}$ a new set of independent observations $\mathbf{y}'$ also becomes available. For the initial data set, prior knowledge is modified via (2.2) to obtain

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta).$$

This then serves as the new prior information before observing $\mathbf{y}'$. Thus the entire calculation of the posterior distribution need not be redone. In other words,

$$\begin{aligned} p(\theta|\mathbf{y},\mathbf{y}') &\propto p(\theta|\mathbf{y})p(\mathbf{y}'|\theta,\mathbf{y}) \\ &= p(\theta|\mathbf{y})p(\mathbf{y}'|\theta) \\ &= p(\theta)p(\mathbf{y}|\theta)p(\mathbf{y}'|\theta) \\ &= p(\theta)p(\mathbf{y},\mathbf{y}'|\theta) \end{aligned}$$

which yields the same result by updating on the basis of all the data at hand $(\mathbf{y},\mathbf{y}')$ directly. By induction it can be easily shown that this algorithm allows information to be updated continually as more data arrive sequentially over time.

Some practical problems in statistics involve a statistical model which contains more than one unknown parameter. However it is often the case that only a subset of the model parameters are of particular interest, and other parameters, called nuisance parameters, are required in order to construct a realistic model.

In this case, the ultimate aim of a Bayesian analysis is to obtain the marginal posterior distribution of the parameters of interest. For example, suppose that the model parameters $\boldsymbol{\theta}$ can be partitioned into two parts, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ is the subvector of interest and $\boldsymbol{\theta}_2$ is the complementary subvector of $\boldsymbol{\theta}_1$. With a prior density $p(\boldsymbol{\theta})$ and a likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$, The Bayes theorem leads to the joint posterior density

$$\begin{aligned} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) &= p(\boldsymbol{\theta}|\mathbf{y}) \\ &\propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= p(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \end{aligned}$$

The marginal posterior distribution of $\boldsymbol{\theta}_1$ is simply obtained by integrating out $\boldsymbol{\theta}_2$ from the joint posterior distribution. Thus,

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = \int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) d\boldsymbol{\theta}_2.$$

2.3 The predictive distributions

One important aspect for Bayesian analysis is to make prediction, which is often the real goal of formulating a statistical analysis. To this end, suppose that Y_{new} is a new independent future observation generated under similar experimental conditions. Before the data $\mathbf{y}$ are obtained, the prior predictive distribution of Y_{new} is

$$p(y_{\text{new}}) = \int p(y_{\text{new}}|\theta)p(\theta)d\theta = \int p(y_{\text{new}}, \theta)d\theta. \tag{2.3}$$

Intrinsic to the idea is the fact that θ can never be observed, but all the available information about θ is summarized by the density $p(\theta)$. Therefore the predictive distribution should be obtained as an average of conditional predictions over the prior distribution of θ. Having observed $\mathbf{y}$, the prediction is based on the distribution of $Y_{\text{new}}|\mathbf{y}$, that is,

$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|\theta)p(\theta|\mathbf{y})d\theta = \int p(y_{\text{new}}, \theta|\mathbf{y})d\theta. \tag{2.4}$$

This is called the posterior predictive distribution of Y_{new}. It summarizes the information concerning the likely values of Y_{new} by averaging over the values of θ according to $p(\theta|\mathbf{y})$, which contains all that we have learned about θ so far.

There are a few other variants of predictive distributions available in the literatures. One of which is developed by using the cross-validatory ('leave one out') approach. For notation convenience, let $\mathbf{y}_{(i)}$ denotes the complete data $\mathbf{y}$ excluding the ith component y_i. The conditional posterior predictive distribution of Y_{new} given $\mathbf{y}_{(i)}$, sometimes called the cross-validation predictive density, is expressible as

$$\begin{aligned} p(y_{\text{new}}|\mathbf{y}_{(i)}) &= \int p(y_{\text{new}}|\theta, \mathbf{y}_{(i)}) p(\theta|\mathbf{y}_{(i)}) d\theta \\ &= \int p(y_{\text{new}}|\theta) p(\theta|\mathbf{y}_{(i)}) d\theta. \end{aligned} \tag{2.5}$$

Note that $\mathbf{y}_{(i)}$ is dropped from $p(y_{\text{new}}|\theta, \mathbf{y}_{(i)})$ due to conditional independence. In words, this is distribution of future replicate data sets acquired by eliminating parameter uncertainty against the 'degraded' posterior knowledge $p(\theta|\mathbf{y}_{(i)})$. The motivation for this predictive distribution is that it can be used as benchmark for detecting whether y_i supports the current model. Henceforth, the terms conditional predictive ordinate (CPO) will be referred to the actual values of $p(y_i|\mathbf{y}_{(i)})$, as is customary.

2.4 Model choice

2.4.1 The Bayes factor

In practice, it is typically the case that more than one model is being contemplated as possible descriptions of the observed data. A Bayesian solution to model comparison, and also selection, is to consider the relative performance of each model via the Bayes factor (BF), defined by

$$\text{BF}_{01} = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)} = \frac{\int p(\mathbf{y}|\theta_0, M_0) p(\theta_0) d\theta_0}{\int p(\mathbf{y}|\theta_1, M_1) p(\theta_1) d\theta_1}$$

for the comparison of models M_0 and M_1. In other words, the model choice criterion is the ratio of the marginal likelihoods, or normalizing constant of the posterior distributions, under the two competing models. Therefore, by considering Bayes factors (see Kass and Rafttery, 1995, for a review), we give support to the model for which the marginal likelihood of the data is highest.

BF_{01}	$\log_{10}BF_{01}$	evidence in favor of M_0
below 1	below 0	negative
1 to 3	0 to 0.5	poor
3 to 10	0.5 to 1	substantial
10 to 100	1 to 2	strong
above 100	above 2	decisive

Table 2.1: Calibration values of the Bayes factor.

Proposed by Jeffreys (1961), Table 2.1 provides a rough calibration for judging BF_{01}.

The use of the Bayes factor has wide advocacy within the Bayesian community. However, there is a serious inherent problem with this criterion in that it cannot be calibrated in the case of improper prior specification. This is because prior predictive distribution (2.3) must necessarily be improper when the prior is improper. For improper prior choices, as long as the resultant posterior distribution is proper, model comparison techniques based on posterior predictive densities may be appropriate. We discuss some of these methods next. Note that a density is considered as improper if its integral over the real line is infinity, which is in conflict with the assumption prescribed by the theory of probability.

2.4.2 The pseudo-Bayes factor

Using the cross-validatory ideas (2.5), Geisser and Eddy (1979) proposed the pseudo-Bayes factor (PsBF) as a surrogate for the Bayes factor. The model selection criterion for comparing two given models M_0 and M_1 is defined as

$$\mathrm{PsBF}_{01} = \prod_{i=1}^{n} \frac{p(y_i|\mathbf{y}_{(i)}, M_0)}{p(y_i|\mathbf{y}_{(i)}, M_1)}.$$

Similar to the Bayes factor, a larger than unity PsBF_{01} shows that there is positive evidence in favor of model M_0 against model M_1 in light of the current

data. Jeffreys' scale of evidence in Table 2.1 can again be employed as reference for deciding how strong the evidence is. Notice now that the pseudo-Bayes factor under improper prior distributions is still meaningful provided the degraded posterior density $p(\theta|\mathbf{y}_{(i)})$ is proper for each i.

Besides the single summary measure PsBF, a plot of CPOs for the candidate models versus observation number is a useful model selection tool as well. Higher value of CPO suggests more support for a model from the corresponding observation. Accordingly, our best model for data is the one which yields the most number of bigger CPOs. Comparing individual CPOs also enables us to guard against any surprising observations concealing the general trend (indicated by the PsBF). This advantage together with the graphical flavor make the cross-validatory approach even more attractive than the Bayes factor. More about PsBF and CPO can be found in Gelfand, Dey and Chang (1992) and Gelfand (1996).

2.4.3 The deviance information criterion

Recently, Spiegelhalter, Best, Carlin and Linde (2002) developed an alternative criterion, known as the deviance information criterion (DIC), for selecting a suitable model from a group of plausible models. They defined the model choice criterion as follows

$$\text{DIC} = D\{E(\theta|\mathbf{y})\} + 2p_D,$$

where

$$D(\theta) = -2log\{p(\mathbf{y}|\theta)\} + 2log[p\{\mathbf{y}|E(\mathbf{Y}|\theta) = \mathbf{y}\}]$$
$$p_D = E\{D(\theta)|\mathbf{y}\} - D\{E(\theta|\mathbf{y})\}.$$

A model yielding the smallest DIC is chosen to be the best model for data. Note here that $D(\theta)$ represents the 'Bayesian saturated deviance' measuring the precision of model fit, whilst p_D is a penalty factor interpreted as the effective number of parameters in a model. Therefore, DIC takes model complexity into account and supports model parsimony.

2.4.4 A minimum posterior predictive loss approach

Another way of choosing an appropriate parametric model is to base the decision on posterior predictive loss. With prediction in mind, suppose $\mathbf{y}_{\text{new}}$ is a future set of observations from a replicated experiment. Following Laud and Ibrahim (1995), see also Gelfand and Ghosh (1998), consider the L^2-criterion

$$\begin{aligned} L^2 &= E\{(\mathbf{Y}_{\text{new}} - \mathbf{y})^T(\mathbf{Y}_{\text{new}} - \mathbf{y})\} \\ &= \sum_{i=1}^{n}[\{E(Y_{\text{new},i}) - y\}^2 + var(Y_{\text{new},i})], \end{aligned} \tag{2.6}$$

where the expectation is taken with respect to the posterior predictive distribution (2.4). It seems obvious that a good model should have $\mathbf{y}_{\text{new}}$ close to what already observed. Hence the best model, among those under consideration, is the one which minimizes the model selection criterion.

Note that modification is needed when dealing with censored data since the actual value of some y_i's are unavailable, refer to Chapter 6 for more discussion about censoring. We accommodate right censored observations to (2.6) via a technique similar to that of Gelfand and Ghosh (1998). If the ith data point is right censored at c_i then estimate y_i by

$$y_i = \begin{cases} c_i & \text{if } E(Y_{\text{new},i}) \leq c_i \\ E(Y_{\text{new},i}) & \text{if } E(Y_{\text{new},i}) > c_i \end{cases}$$

Thus, after replacing all censored observations with the relevant estimates, (2.6) can be calculated in the usual way. The criterion under logarithmic responses (i.e. $\mathbf{y}$ and $\mathbf{y}_{\text{new}}$ are substituted by their natural logarithms) will be illustrated using a survival data set in Chapter 6.

2.5 The prior distributions

As mentioned in the previous section, in addition to the likelihood, Bayesian analysis requires the specification of a prior distribution for all model parameters. Whilst this prior density serves as the best way to summarize one's genuine prior belief about the parameters, its determination will certainly influence the resulting inference. Therefore considerable care is required when making a

choice for the prior distribution. In this section, we discuss two common techniques, the conjugate prior approach and the non-informative approach.

2.5.1 Conjugate prior distributions

Prior information from past studies about similar statistical investigation or opinions of subject-area experts are often available. In practice, one might have many probability densities that are compatible with this information. Among these densities, it would be helpful to select a prior distribution that simplifies the subsequent computational burden. The implementation of the Bayes theorem can be computationally difficult due to the normalizing integral in (2.1). One can eliminate the need to evaluate this integral by introducing the so-called conjugate prior. A class of prior distributions $\mathcal{P}$ is said to be conjugate for a likelihood $p(\mathbf{y}|\theta)$ if the resulting posterior distribution also belongs to $\mathcal{P}$. The property that the posterior follows the same parametric form as the prior is called conjugacy, and $\mathcal{P}$ is said to be a conjugate family for the posterior distribution.

However, conjugacy is a rather vacuous idea in the sense that if $\mathcal{P}$ is the class of all distributions, then it will obviously lead to a posterior belonging to the same distributional family as the prior, no matter what class of sampling distributions is used. This is, of course, useless for the choice of a prior density. Therefore, the main interest of conjugacy arises when $\mathcal{P}$ is a set of densities having the same functional form as the likelihood. This specific type of conjugate prior distributions can only be obtained easily for data models within the exponential family, which take the following form

$$p(y|\theta) = g(y)h(\theta)\exp\{\phi(\theta)u(y)\}, \tag{2.7}$$

where $g(\cdot), h(\cdot), \phi(\cdot)$ and $u(\cdot)$ are suitable functions. The class might seem restrictive, but in fact it includes many common continuous and discrete distributions such as the normal, binomial, exponential and Poisson distributions. Important distributions that do not belong to the exponential family include the uniform and Students t distributions.

Assume that the set of independent observations $\mathbf{Y}$ follows the exponential

family (2.7). Then, with a prior density $p(\theta)$, the posterior distribution of θ is given by

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\theta)p(\mathbf{y}|\theta) \\ &= p(\theta)\prod_{i=0}^{n}[g(y_i)h(\theta)\exp\{\phi(\theta)u(y_i)\}] \\ &\propto p(\theta)h(\theta)^n\exp\{\phi(\theta)\sum_{i=1}^{n}u(y_i)\}. \end{aligned}$$

The identity of the conjugate prior distribution can be determined by regarding the likelihood as a function of θ. Thus defining a prior distribution in the form of

$$p(\theta) \propto h(\theta)^{n_0}\exp\{\phi(\theta)u_0\}$$

leads to

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto h(\theta)^{n_0+n}\exp\left[\phi(\theta)\{u_0+\sum_{i=1}^{n}u(y_i)\}\right] \\ &= h(\theta)^{\eta}\exp\{\phi(\theta)\mu\} \end{aligned}$$

where $\eta = n_0 + n$ and $\mu = u_0 + \sum_{i=1}^{n} u(y_i)$. Therefore the chosen prior density has managed to retain the posterior density in the same algebraic form. In fact, switching from prior to posterior distribution is reduced to the task of updating the corresponding parameters. Table 2.2 presents the conjugate distributions for some commonly used distributions belonging to the exponential family.

Although conjugacy can permit posterior distributions to emerge without numerical integration, analysis with the conjugate prior distributions must be used with care. The analytic tractability of this specification comes with a price due to the restrictions they impose on the form of the prior distributions. For instance, the conjugate prior distributions may not deliver an adequate representation of one's prior beliefs, or they may not even exist for complicated sampling models. Nonetheless, if they do exist and can provide a sufficiently close description of one's prior state of uncertainty, the advantages brought by conjugacy are still irresistible.

Likelihood	Prior
Normal(known variance)	Normal
Poisson	Gamma
Exponential	Gamma
Binomial	Beta
Negative Binomial	Beta
Normal(known mean)	Inverse Gamma

Table 2.2: Conjugate priors for some well known exponential families.

2.5.2 Non-informative prior distributions

When no reliable prior information about the model parameters is available, or when an inference based solely on the data is desirable, the imperative is then to minimize the influence of the prior distribution on the resulting posterior distribution. Such prior distributions are sometimes called non-informative, vague, flat, diffuse or reference priors, representing the states of 'prior ignorance' and 'to let the data speak for themselves'. Different formulations lead to different types of non-informative prior distributions. One of the most widely accepted formulation was proposed by Jeffreys (1961), based on considering the consistency of prior ignorance across one-to-one parameter transformations. Jeffreys' choice of non-informative prior distribution is

$$p(\theta) \propto |\mathcal{I}(\theta)|^{1/2} \tag{2.8}$$

where $\mathcal{I}(\theta)$ is the Fisher information for θ, given by

$$\mathcal{I}(\theta) = -E\left(\frac{d^2 \log p(y|\theta)}{d\theta^2}\right) = E\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2.$$

Jeffreys' invariance under re-parametrization is justified in the following sense: any procedure for deriving the prior distribution should grant the same result if applied to the transformed parameter. To see this, suppose $p(\theta)$ is the prior density of θ and let $\phi = h(\theta)$ be a one-to-one transformation of θ. Then

the prior density of ϕ is

$$\begin{aligned} p(\phi) &= p(\theta)\left|\frac{d\theta}{d\phi}\right| \\ &\propto \mathcal{I}(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|. \end{aligned}$$

Now, evaluating $\mathcal{I}(\phi)$ at $\theta = h^{-1}(\phi)$ yields

$$\begin{aligned} \mathcal{I}(\phi) &= -E\left(\frac{d^2 \log p(y|\phi)}{d\phi^2}\right) \\ &= -E\left(\frac{d^2 \log p\{y|\theta = h^{-1}(\phi)\}}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) \\ &= \mathcal{I}(\theta)\left(\frac{d\theta}{d\phi}\right)^2. \end{aligned}$$

Thus,

$$p(\phi) \propto \mathcal{I}(\phi)^{1/2}$$

which is equivalent to the prior distribution for ϕ determined directly by using (2.8), as required.

While Jeffreys' method provides an automated technique for obtaining non-informative prior distributions, it can lead to paradoxes in multi-parameter models. There are two distinct ways to apply Jeffreys' principle in multi-dimensional cases which often bring about different results. The obvious one is to evaluate the joint non-informative prior distribution directly by using the Jeffreys' rule above. A simpler approach, based on the assumption of independence, is to derive the joint prior distribution as the product of the Jeffreys' non-informative prior distributions for each components of the vector parameter. Normally, the latter procedure is preferable as it is more mathematically convenient and it enjoys a specific kind of coherence in that 'ignorance' is in some sense parallel to 'independence'.

In general, the Jeffreys' non-informative approach leads to prior distributions in the form $p(\theta) \propto 1$ for location parameters, and $p(\theta) \propto \theta^{-1}$ for scale parameters. Notice that these densities are improper. Although there are some further difficulties involved, Bayesian inference using improper priors is still possible provided the resulting posterior distribution is proper. However, it is worth

emphasizing that verification of propriety in many complex models is far from trivial. Thus non-informative prior distributions are to be used with caution. Box and Tiao (1973) suggest that if the likelihood function truly dominates the prior distribution then the precise form of the non-informative prior distribution does not matter. The main issue is, therefore, to find a prior distribution which states that little is known a priori relative to the data about the model parameters.

2.6 Summarizing posterior information

Let us assume at this point that both the sampling distribution $p(\mathbf{y}|\theta)$ and the prior distribution $p(\theta)$ of the model parameter θ are available. The Bayes theorem (2.1) can now be used to combine the experimental evidence and the prior knowledge to produce the posterior density $p(\theta|\mathbf{y})$. Since this updated distribution represents all the current extensive information about θ, ideally one may report the entire distribution as a basis for all posterior inference. For instance, a graphical display of the corresponding density function might prove useful in providing the insight about the behavior of the parameter. Still, for many practical purposes, it is often desirable to summarize the posterior information through a point estimate or interval estimate. Although we only work with the posterior distribution in this section, most of the discussions can be applied equally well to other distributions such as the prior and posterior predictive distributions given in equations (2.3) and (2.4), respectively.

2.6.1 Point estimation

To select a summary feature from the posterior distribution which in some way 'best' reflects the parameter under study confronts us with a decision-making problem. The consequences of making a selection or decision can be studied using the concept of loss functions. A loss function is supposed to evaluate the penalty or loss associated with a decision, and the aim is to take the decision which minimizes the expected loss. Suppose that the posterior density $p(\theta|\mathbf{y})$ of θ has been formally derived. Then, for any particular decision $\ddot{a}$, the posterior

expected loss is

$$E\{L(\theta, \ddot{a})|\mathbf{y}\} = \int L(\theta, \ddot{a})p(\theta|\mathbf{y})d\theta$$

where $L(\theta, \ddot{a})$ is a loss function, e.g. a function of $|\theta - \ddot{a}|$. A Bayes point estimate of θ is the value $\ddot{a}$ that gives rise to the minimum value of $E\{L(\theta, \ddot{a})\}$. There are many different choices of loss functions, and the particular choice for any specified problem will depend on the context. We study some of the most widely used loss functions and the associated Bayes estimators below.

1. Quadratic loss function
   Proposed by Legendre in 1805 and also by Gauss in 1810, the loss function $L(\theta, \ddot{a}) = (\theta - \ddot{a})^2$ is called the quadratic loss. The Bayes estimate of θ with respect to this loss function is the posterior mean, i.e. the expected value of θ under $p(\theta|\mathbf{y})$.

   Proof: Since

   $$\begin{aligned} E\{L(\theta, \ddot{a})|\mathbf{y}\} &= E\{(\theta - \ddot{a})^2|\mathbf{y}\} \\ &= E(\theta^2|\mathbf{y}) - 2\ddot{a}E(\theta|\mathbf{y}) + \ddot{a}^2. \end{aligned}$$

   The expected loss actually attains its minimum at $\ddot{a} = E(\theta|\mathbf{y})$.

   □

   Quadratic loss can be extended to multi-parameter cases:

   $$L(\boldsymbol{\theta}, \ddot{\mathbf{a}}) = (\boldsymbol{\theta} - \ddot{\mathbf{a}})^T \mathbf{H}(\boldsymbol{\theta} - \ddot{\mathbf{a}})$$

   where $\ddot{\mathbf{a}}$ and $\boldsymbol{\theta}$ are vectors and $\mathbf{H}$ is a positive definite matrix. As in one-parameter cases, for the quadratic loss the Bayes estimator is the mean of the joint posterior distribution, assuming that it exists.
   This particular loss function is appropriate in situations where losses are approximately symmetric in $|\boldsymbol{\theta} - \ddot{\mathbf{a}}|$ and large deviations need to be penalized heavily.

2. Absolute error loss function
   For absolute error loss, $L(\theta, \ddot{a}) = |\theta - \ddot{a}|$, the Bayes decision rule is to

estimate θ by the posterior median, which divides the parameter space into two equal probability parts.

Proof: The posterior expected loss is

$$\begin{aligned} E[L(\theta, \ddot{a})|\mathbf{y}] &= \int_{-\infty}^{\infty} |\theta - \ddot{a}| p(\theta|\mathbf{y}) d\theta \\ &= \int_{-\infty}^{\ddot{a}} (\ddot{a} - \theta) p(\theta|\mathbf{y}) d\theta + \int_{\ddot{a}}^{\infty} (\theta - \ddot{a}) p(\theta|\mathbf{y}) d\theta. \end{aligned}$$

Differentiating, using Leibniz' rule, with respect to $\ddot{a}$ and setting this equal to zero gives

$$\begin{aligned} &\int_{-\infty}^{\ddot{a}} p(\theta|\mathbf{y}) d\theta - \int_{\ddot{a}}^{\infty} p(\theta|\mathbf{y}) d\theta = 0 \\ &\Longrightarrow \int_{-\infty}^{\ddot{a}} p(\theta|\mathbf{y}) d\theta = \int_{\ddot{a}}^{\infty} p(\theta|\mathbf{y}) d\theta. \end{aligned}$$

Adding $\int_{-\infty}^{\ddot{a}} p(\theta|\mathbf{y}) d\theta$ to both sides yields

$$2 \int_{-\infty}^{\ddot{a}} p(\theta|\mathbf{y}) d\theta = 1 \iff Pr(\theta < \ddot{a}|\mathbf{y}) = \frac{1}{2}.$$

□

A generalization of the absolute error loss function is the linear loss function:

$$L(\theta, \ddot{a}) = \begin{cases} a_1(\theta - \ddot{a}) & \text{if } \theta \geq \ddot{a} \\ a_2(\ddot{a} - \theta) & \text{if } \theta < \ddot{a}. \end{cases}$$

In this case, the Bayes estimator is the quantile of the posterior distribution such that $P(\theta \leq \ddot{a}) = \frac{a_1}{a_1 + a_2}$.

A linear loss function is useful when the losses are assumed to be approximately linear in $(\theta - \ddot{a})$. Note that it increases more slowly than the quadratic loss function and hence does not over-penalize large deviations. In addition, the constants a_1 and a_2 can be chosen as a measure of the relative importance of underestimation and overestimation.

3. Step function loss

   The Bayes estimator associated with the step function loss

$$L(\theta, \ddot{a}) = \begin{cases} 0 & \text{if } |\theta - \ddot{a}| \leq \delta \\ 1 & \text{otherwise.} \end{cases}$$

is the posterior mode, i.e. the most likely value under the posterior distribution.

Proof: Let $I(\cdot)$ be an indicator function (taking value one if its argument is true, and zero otherwise). Then the step function loss can be rewritten as $L(\theta, \hat{\theta}) = 1 - I(|\theta - \hat{\theta}| \leq \delta)$. As $\delta \to 0$,

$$\begin{aligned} E\{L(\theta, \ddot{a})|\mathbf{y}\} &= 1 - E\{I(|\theta - \ddot{a}| \leq \delta)|\mathbf{y}\} \\ &= 1 - \int_{|\theta - \ddot{a}| \leq \delta} p(\theta|\mathbf{y}) d\theta. \end{aligned}$$

$E\{L(\theta, \ddot{a})|\mathbf{y}\}$ is minimized when the integral $\int_{|\theta - \ddot{a}| \leq \delta} p(\theta|\mathbf{y}) d\theta$ is maximized. Thus the posterior mode is the Bayes estimate for θ here.

□

A multivariate extension of the step function loss would be

$$L(\boldsymbol{\theta}, \ddot{\mathbf{a}}) = 1 - I(|\boldsymbol{\theta} - \ddot{\mathbf{a}}| \leq \boldsymbol{\delta}).$$

This is called the zero-one loss. The Bayes decision rule in this case is to choose $\ddot{\mathbf{a}}$ to be the mode of the posterior distribution.
Although $\ddot{\mathbf{a}}$ has the interpretation of being the most likely value of $\boldsymbol{\theta}$, in practice, zero-one loss will rarely be a good approximation to the true loss because of its non-quantitative nature.

While easy to work with, it should be made clear that these losses need not necessarily be appropriate for a given problem. The mean, median and mode of the posterior distribution are commonly used because they are often reasonable point estimates of θ and, for a conjugate family of distributions, they are relatively easy to compute. In the case where the posterior distribution is unimodal and symmetric, the posterior mean, median and mode are all identical. In general, however, they might not coincide and the difference can be quite substantial. Therefore, unless there is a very clear need for a point estimate and a strong rational for a specific loss function, the provision of a single number to summarize the posterior density may be extremely misleading.

2.6.2 Interval estimation

Point estimates give no measure of accuracy. Accordingly, it is always important to report posterior uncertainty by using indices such as the posterior variance, posterior quantiles and posterior intervals. Interval summaries are particularly useful in Bayesian inference in that it facilitates a common sense interpretation of having a certain probability of containing the parameter of interest. Formally, the Bayesian analogue of a confidence interval in classical statistics is referred to as a credible set, defined as follows.

Definition 2.1 For a posterior distribution $p(\theta|\mathbf{y})$, a set C is said to be a $100(1-\alpha)\%$ credible set for θ if

$$Pr(\theta \in C|\mathbf{y}) = 1 - \alpha.$$

□

In other words, θ has a probability of $(1-\alpha)$ to belong to a fixed interval C. One difficulty with credible sets is that, for any given α, they are not uniquely defined. To tackle the problem, an additional constraint needs to be imposed. One way of doing this is to consider an interval that has the smallest width or, equivalently, an interval that includes only the most plausible values of θ. Such an interval is called a highest posterior density (HPD) credible set.

Definition 2.2 For a posterior density $p(\theta|\mathbf{y})$, a set C is said to be a $100(1-\alpha)\%$ HPD credible set for θ if it can be written under the form

$$C = \{\theta : p(\theta|\mathbf{y}) \geq k_\alpha\}$$

where k_α is chosen to ensure that

$$Pr(\theta \in C|\mathbf{y}) = 1 - \alpha.$$

□

HPD credible set are not invariant under a non-linear parameter transformation.

One other choice of constraint is simply to take the posterior $\alpha/2$ and $1-\alpha/2$ quantiles as the limits of a $100(1-\alpha)\%$ credible set. This central or equal tail

credible set will be equal to the HPD credible set if the posterior is unimodal and symmetric, but will be a bit wider otherwise. Moreover, it is invariant to one-to-one transformations of the parameter and is usually easier to compute. Although the idea of credible set can be extended in exactly the same way to parameters of higher dimension, it may not be easy to comprehend the picture in more than three dimensions. Occasionally, inspection of credible regions of some appropriate conditional and marginal distributions will greatly assist understanding of the posterior state of knowledge in higher dimensions.

2.7 Bayesian computation

Recall that, in general, the posterior density $p(\theta|\mathbf{y})$ of θ has no closed form expression as the normalizing constant

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$$

appearing in (2.1) is often not tractable. Fortunately, the entire posterior distribution can still be accurately approximated by means of sampling-based methods even if the posterior distribution is only known up to a constant of normalization. In particular, the development of Markov chain Monte Carlo (MCMC) computing methods has eliminated many constraints on the prior distributions and made it practically possible to fit models with increasing complexity. The idea of MCMC is very straightforward: "One is interested in simulating from a target distribution with density π but cannot do this directly. Instead, one constructs a Markov chain with equilibrium distribution π and runs it long enough until convergence has been obtained. Following a sufficiently long burn-in period, simulated values of the chain will be dependent samples approximately from π". In this section, we briefly discuss two MCMC procedures to create a sample from π, which in our case is the posterior distribution. A short account of estimation using MCMC output follows.

Note that the presentation of MCMC methods in this section is far from comprehensive. Since the focus of this dissertation is data analysis rather than computation, we only provide sufficient information for these strategies to be implemented. There are many excellent books on a detailed treatment of the

subject such as the books by Gilks, Richardson, and Spiegelhalter (1996) and Chen, Shao, and Ibrahim (2000).

2.7.1 Gibbs sampler

Consider the problem of drawing a sample from a multivariate distribution $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)^T$, each of the d components could be sub-vector of $\boldsymbol{\theta}$. Suppose also that the full conditional distributions

$$\pi(\theta_i|\theta_{-i}) = \pi(\theta_i|\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_d), \qquad i = 1, 2, \cdots, d$$

are completely known and easy to sample from. The Gibbs sampler provides simulations from $\pi(\boldsymbol{\theta})$ based on successive generations from these full conditional distributions. The iterative procedure can be described as follows.

1. Initialize the chain with an arbitrary set of values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \cdots, \theta_d^{(0)})^T$.

2. Simulate $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \cdots, \theta_d^{(j)})^T$ from

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \cdots, \theta_d^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \cdots, \theta_d^{(j-1)}) \\
&\vdots \\
\theta_d^{(j)} &\sim \pi(\theta_d|\theta_1^{(j)}, \cdots, \theta_{d-1}^{(j)})
\end{aligned}$$

   for $j = 1, 2, \cdots, t$.

That is, each component θ_i is updated conditional on the latest values of all other components $\theta_{(i)}$. When convergence is reached, $\boldsymbol{\theta}^{(j)}$ can be regraded as one simulated value from the target distribution $\pi(\boldsymbol{\theta})$. Thus, the requirement of obtaining samples from the joint distribution of $\boldsymbol{\theta}$ has come down to the ability to sample from the d corresponding full conditional distributions.

Note that, the availability of a sample for a multidimensional parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)^T$ means that the ith component of the simulations $\theta_i^{(1)}, \cdots, \theta_i^{(t)}$ is a sample approximately from the marginal posterior distribution of θ_i, $i = 1, \cdots, d$.

2.7.2 Metropolis-Hastings algorithm

Given a target distribution $\pi(\theta)$, where θ can be a scalar or a vector, that can be computed up to a multiplicative constant, the Metropolis-Hastings algorithm generates a Markov chain by the following steps:

1. Start the chain with an arbitrary initial value $\theta^{(0)}$.
2. For $j = 1, 2, \cdots, t$
   (a) Generate θ' from a proposal density $q(\theta'|\theta^{(j-1)})$.
   (b) Compute the ratio
   $$r = \frac{\pi(\theta')q(\theta^{(j-1)}|\theta')}{\pi(\theta^{(j-1)})q(\theta'|\theta^{(j-1)})}$$
   (c) Set
   $$\theta^{(j)} = \begin{cases} \theta' & \text{with probability } \min(1, r), \\ \theta^{(j-1)} & \text{otherwise.} \end{cases}$$

On convergence, the values $\theta^{(j)}$ can be considered as approximate draws from $\pi(\theta)$. Notice that the proposal distribution only defines a candidate value θ' that is substantiated according to the value of r. For obvious reasons, r is generally referred to as the test ratio. Note that the density q can take any form and thus provides a flexible tool for the construction of the algorithm. However, for computational efficiency, it is crucial to have q so that

- for any $\theta^{(j-1)}$, it is easy to sample from $q(\theta'|\theta^{(j-1)})$,
- the test ratio r can be easily evaluated,
- the proposed candidates θ' are not rejected too frequently, and
- each candidate provides a reasonable displacement from the current state.

We now discuss two most common choices for q known as the independent and the random-walk proposals.

Independence sampler

In this case, the proposal distribution does not depend on the previous position $\theta^{(j-1)}$, that is

$$q(\theta'|\theta^{(j-1)}) = q(\theta').$$

Thus, the test ratio reduces to

$$r = \frac{\pi(\theta')/q(\theta')}{\pi(\theta^{(j-1)})/q(\theta^{(j-1)})} = \frac{w(\theta')}{w(\theta^{(j-1)})},$$

where $w(\theta) = \pi(\theta)/q(\theta)$. The choice of q is very important for the practical implementation of the method. The general rule for the independent sampler to work well is to choose a proposal distribution that approximates well the target distribution π but is slightly heavier tailed. However, finding a suitable proposal distribution may be a difficult task and hence limit the applicability of the method.

Random walk algorithm

The chain has proposed candidates θ' according to

$$q(\theta'|\theta^{(j-1)}) = q(\theta' - \theta^{(j-1)}),$$

or, to put it another way,

$$\theta' = \theta^{(j-1)} + w_j$$

where w_j is a symmetric random variable centered at the origin. For this proposal, the test ratio is

$$r = \frac{\pi(\theta')}{\pi(\theta^{(j-1)})}$$

which does not depend on q. The proposal distribution can be formulated independently of the target distribution but care is needed in specifying the scale of q. Small variances will lead to high acceptance rates at the expense of small displacements from the current state. On the other hand, large variances will generate large spreads but with small acceptance rates. Both extremes should be avoided as the chain will mix slowly which will result in low efficiency. It is recommended by Gelman *et al.* (1995) that the dispersion of q should be chosen in order to provide an acceptance rate in the range [0.15, 0.5].

2.7.3 Output analysis

Suppose that a post-convergence correlated sample $\theta^{(1)}, \cdots, \theta^{(t)}$ generated by using a MCMC scheme with stationary distribution $p(\theta|\mathbf{y})$ is available. Assume also the more general case where these are successive values from a single long chain. Using these simulated values, all relevant calculations with the posterior distribution $p(\theta|\mathbf{y})$ can be approximated. In particular,

- A smoothed version of the histogram of the sampled values can be plotted to provide an estimate of the entire posterior density.
- The posterior mean is estimated by the average of the simulated values.
- A $100(1-\alpha)\%$ central or equal tail posterior interval is approximately given by the $[(\alpha/2) \times N]$th and $[(1-\alpha/2) \times N]$th ordered sample values, where N is the total number of simulations and $[a]$ denotes the integer part of a.

Obviously these approximations will become more accurate as, N, the number of simulations increases.

An estimator of posterior point or interval summaries of any parametric transformation $\phi = f(\theta)$ is obtained similarly by using the transformed sample $\phi^{(1)}, \cdots, \phi^{(t)}$, where $\phi^{(j)} = f(\theta^{(j)})$. Specifically, the posterior expectation of ϕ is approximated as

$$\hat{\phi} = E\{f(\theta)|\mathbf{y}\} = \frac{1}{t}\sum_{j=1}^{t}\phi^{(j)}.$$

The batch means method can be used to assess the accuracy of this estimation:

1. Batch or divide the single Markov chain into m successive batches of length b. Generally, we take $m \in (10, 30)$.
2. Compute the batch averages $\bar{B}_1, \cdots, \bar{B}_m$.
3. Check whether the autocorrelation between batches is negligible, say less than 0.05. If it is not the case, select a larger b and repeat the procedure.

4. Sampling variance is approximately

$$\hat{Var}(\hat{\phi}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (B_i - \hat{\phi})^2.$$

Once simulations from the posterior distribution are available, it is typically easy to draw from the posterior predictive distribution of future data Y_{new}. Recall that the posterior predictive distribution for Y_{new} is given by

$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|\theta) p(\theta|\mathbf{y}) d\theta.$$

In other words, $p(y_{\text{new}}|\mathbf{y})$ is a marginal density computed from $p(\theta|\mathbf{y})$. A sample ${y_{\text{new}}}^{(1)}, \cdots, {y_{\text{new}}}^{(t)}$ from the predictive density is obtained by drawing each ${y_{\text{new}}}^{(j)}$ from $p(y_{\text{new}}|\theta^{(j)})$, which is the sampling distribution with parameter $\theta^{(j)}$. The posterior predictive distribution plays an important role in checking the fit of a model to the observed data. If the model is reasonably accurate, the replicated data ${y_{\text{new}}}^{(1)}, \cdots, {y_{\text{new}}}^{(t)}$ should look similar to the data $\mathbf{y}$ that have actually been observed. Any systematic differences between the posterior predictive density and the observed data signal potential failings of the posited model. Inferences can be misleading when a probability model is far from reality.

Chapter 3

The skew-elliptical distributions

3.1 Introduction

Statistical distributions provide the foundation for many statistical procedures and data analysis. The reliability of empirical results lies on the capability of the assumed distribution to model the specific characteristics of the underlying data. Although a huge number of distributions have been proposed and investigated, statistical techniques for continuous data analysis are based largely, both implicitly and explicitly, on the celebrated normal distribution. A major reason for this state of affair is certainly the mathematical implications resulting from the normality assumptions. Computational simplicity is very desirable, but conceptual and flexibility are not unimportant. A well recognized limitation of the normal distribution is the paucity of its competency to accommodate skewness and kurtosis. It is evident that, in real applications, non-normal tail behavior and asymmetry are common traits in the body of data. Accordingly, there have always been resistance to the normal theory methods.

Recently there has been renewed interest in the statistical literature towards robust statistical methods in order to represent features of the data as adequately as possible and reduce unrealistic assumptions. This remark is reflected in the substantial growth in the number of distributional families developed, studied and used for data modeling as alternatives to the normal theory statistics. Some families of distributions which allow for skewness and contain the normal distribution as a proper member or as a limiting case have played

an important role in these developments. Among them are the skew-normal distribution (Azzalini, 1985, 1986), the multivariate skew-normal distribution (Azzalini and Dalla Valle, 1996), the two-piece normal distribution (John, 1982), the epsilon-skew-normal distribution (Mudholkar and Hutson, 2000), the skew-t distribution (Jones and Faddy, 2003, and Jones 2003), the generalized skew-t distribution (Theodossiou, 1998), the two-piece t distribution (Fernandez and Steel, 1998), the skew-elliptical distribution (Sahu, Dey and Branco, 2003), and the generalized skew-elliptical distribution (Genton and Loperfido, 2001).

This chapter extends the previous version of skew-elliptical (SE) distribution, introduced by Sahu *et al.* (2003). The extended class is distinct from the one obtained by Branco and Dey (2001) but contains the Sahu *et al.* (2003) family as a special case. Branco and Dey (2001) develop their multivariate SE distributions by conditioning on one suitable random variable being positive while Sahu *et al.* (2003) impose the non-negativity condition on the same number of random variables. Heuristically, we generalize their ideas by releasing the dimensionality restriction on the conditioned variables. After a short review on various ways of generating the basic univariate skew normal distribution in Section 3.2, Section 3.3 describes the derivation and density function of this new SE distribution. The family is then used in Section 3.4 to define a class of univariate skew normal distributions, which will be the main focus for the remainder of the paper. Central moments of the skew-normal distribution are obtained, along with a discussion of some related properties. Section 3.5 studies an alternative skew-normal distribution that proved to be popular in the literature. We compare three distinct versions of univariate skew normal distributions in Section 3.6. Some possible generalizations of the skew-elliptical distribution are discussed in Section 3.7. The chapter concludes with a few summary remarks in Section 3.8.

3.2 The basic univariate skew-normal model

The term skew normal distribution was first introduced by Azzalini in 1985 as a natural extension of the normal density to accommodate asymmetry. The

main motivation for considering the distribution was the desirability of a class of densities which is mathematically tractable, strictly contains the normal density, and permits a wide range of degrees of skewness and kurtosis. A random variable Z is said to have a skew normal distribution with parameter $\lambda \in \Re$ if it has probability density function

$$f(z|\lambda) = 2\phi(z)\Phi(\lambda z), \qquad z \in \Re \tag{3.1}$$

where here and henceforth we denote the standard normal density and distribution function by $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. The above density is positively skewed when $\lambda > 0$, skewed to the left when $\lambda < 0$, and symmetric when $\lambda = 0$ (in which case it coincides with the standard normal distribution). Therefore it is reasonable to regard λ as the skewness parameter.

A salient feature of the skew normal distribution is that it can be derived in many different settings, i.e. it admits various characterizations.

Scenario 1 (Proposition 1 of Azzalini *et al.*, 1996.) Let U and V be independent $N(0,1)$ variables. Define Z to be equal to U conditionally on $\lambda U > V$. Then Z has density (3.1).

Scenario 2 (Proposition 2 of Azzalini *et al.*, 1996.) Assume that (Z, V) has a bivariate normal distribution with $N(0,1)$ marginals and correlation coefficient $\lambda/\sqrt{1+\lambda^2}$. Then the conditional density of Z given $V > 0$ is equivalent to (3.1).

Scenario 3 (Proposition 2 and 3 of Azzalini, 1986.) If V is a $N(0,1)$ variate, then

$$Z_1 = \begin{cases} V & \text{with probablity } \Phi(\lambda v) \\ -V & \text{with probablity } 1 - \Phi(\lambda v) \end{cases}, \text{ and}$$

$$Z_2 = \begin{cases} |V| & \text{with probablity } \Phi(\lambda |v|) \\ -|V| & \text{with probablity } 1 - \Phi(\lambda |v|) \end{cases}$$

have density (3.1).

Scenario 4 (See Azzalini, 1986, page 201.) Let $\{Z_t\}$ be a stationary process satisfying

$$Z_t = \frac{\lambda}{\sqrt{1+\lambda^2}}|Z_{t-1}| + \varepsilon_t \qquad \text{for } t = 0, \pm 1, \pm 2, \cdots$$

where ε_t is $N(0, 1/\sqrt{1+\lambda^2})$. Then the stationary distribution of Z_t is given by (3.1).

Scenario 5 (Theorem 1 of Henze, 1986.) Suppose U and V are two independent identically distributed $N(0,1)$ random variables. Then

$$Z = \frac{\lambda}{\sqrt{1+\lambda^2}}|U| + \frac{1}{\sqrt{1+\lambda^2}}V$$

has density (3.1).

Scenario 6 (Theorem 2.3 of Loperfido, 2002.) Consider a bivariate random vector (U, V) whose marginal distributions are $N(0,1)$'s and whose correlation coefficient is $(1-\lambda)/(1+\lambda)$. Then the density of the random variable $Z = \max(U, V)$ is (3.1).

All these genesis scenarios provide a physical justification for the skew-normal distribution that may help in understanding its intrinsic structure as well as revealing new applications. For example, Arnold *et al.* (1993, 2002) suggested thinking of Scenario 2 as the marginalization of a hidden truncated bivariate normal density; and Loperfido (2002) has regarded Scenario 6 as selective reporting. The distinct genesis representations may be also useful for simplifying some computations such as moments calculation and random numbers generation. Another attractive implication of the results is that they can each be fruitfully employed for extending the basic skew normal distribution to more general settings. Although this opens the way to the study of particular cases, this chapter will only consider using a method resembling Scenario 2 for multivariate and non-normal extensions.

3.3 Derivation of the skew-elliptical distributions

The present section utilizes a general method for introducing skewness into any symmetric distributions and applies it on the elliptical distributions. To this end, consider two independent random vectors $\mathbf{U}$ and $\mathbf{V}$, both with unimodal and symmetric densities. Now a class of skew distributions can be generated via the following formulation

$$\mathbf{Z} = \mathbf{DU} + \mathbf{V}, \qquad \mathbf{U} > \mathbf{0} \tag{3.2}$$

where $\mathbf{D}$ is a fixed matrix. For the univariate setting in which U and V are chosen to be iid standard normal random variables, a simple convolution computation shows that $Z/\sqrt{D^2+1}$ indeed has a basic skew normal distribution (3.1) with $\lambda = D$. Notice that, in this particular case, equation (3.2) together with the transformation argument is essentially equivalent to Scenario 2. Nonetheless, paradigm (3.2) provides a more general, yet simpler, way of generalizing the basic skew normal density.

The replacement of normal variate in the development of model (3.1) by other statistical distributions has become quite popular. For example, Arnold and Beaver (2000) have substituted the normal component by a suitable heavy tail alternative to obtain the skew Cauchy density. A broader class of multidimensional models, hinted by Azzalini and Capitanio (1999), can be elicited if the normal distribution is replaced by an elliptical distribution. Some related results along these lines can be found in Branco and Dey (2001) and Sahu *et al.* (2003). The probability distribution proposed in this section extends the previous version induced by Sahu *et al.* (2003). Before presenting the new skew elliptical distribution, it is useful to recall the definition of the elliptical distribution.

3.3.1 Elliptical distributions

The *elliptical distribution*, originally defined by Kelker (1970), represents a natural generalization of the concept of symmetry to the multivariate setting. A

comprehensive review of the distribution can be found in Fang, Kotz and Ng (1990). A random vector $\mathbf{X}$ with values in $\Re^k$ has an elliptical distribution with location vector $\boldsymbol{\mu} \in \Re^k$ and covariance matrix Σ if its density function is of the form

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma; g^{(k)}) = |\Sigma|^{-1/2} g^{(k)}\{(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \tag{3.3}$$

for some density generator function defined by

$$g^{(k)}(u) = \frac{\Gamma(k/2)}{\pi^{k/2}} \frac{g(u;k)}{\int_0^\infty r^{k/2-1} g(r;k)dr}, \qquad u \geq 0$$

where $g(u;k)$ is a non-increasing function ensuring that the integral $\int_0^\infty r^{k/2-1}$ $g(r;k)dr$ exists. For simplicity, Σ is assumed to be positive definite. In what follows the notation $El(\boldsymbol{\mu}, \Sigma; g^{(k)})$ will be used to describe the above probability distribution.

The choice of generator function $g^{(k)}(\cdot)$ will determine the distribution of X. Its flexibility enables the elliptical class to acknowledge many well-known symmetrical distributions as proper members, e.g. the multivariate normal, uniform, Student's t, exponential power, and Pearson type II distributions. These densities have a wide range of tail shapes, but the general specification of X being elliptically distributed does not imply either light or heavy tailed distribution. Hence, to some extend, it is admissible to consider (3.3) as a universal model for summarizing kurtosis of a symmetric data. The particular case of normal distribution $N_k(\boldsymbol{\mu}, \Sigma)$ is obtained by defining $g(u;k) = e^{-u/2}$. Note that the function $g(u;k)$ may depend on other parameters. As an example, taking $g(\cdot) = [1 + \frac{u}{v}]^{(-v+k)/2}$, $v > 0$, then the correspondence with the Student's t distribution is apparent.

Elliptical distribution, however, imposes the restriction on symmetry, which does not facilitate the analysis of the effects of skewness. It is accepted that, in real applications, kurtosis and skewness are often observed characteristics of empirical data. Accordingly, statistics employed by assuming ellipticity are not always valid and can be of little value for summarizing the structure in a body of data. The ability to incorporate these pervasive features simultaneously is therefore an important practical consideration. Hence it seems reasonable and appropriate to acquire a skewed version of elliptical distribution so as to enable

a trustworthy analysis of non-normal data.

3.3.2 Skew-elliptical distributions

The general procedure of skewing a symmetric unimodal distribution presented at the beginning of this section provides a simple yet powerful way for generating new distributions. The following theorem applies the previous results to develop a general class of skewed multivariate distributions. The proof of the theorem rests mainly on the properties of the elliptical distributions (Chapter 2 of Fang *et al.*, 1990).

Theorem 3.1 Let $\mathbf{U}$ and $\mathbf{V}$ be two independent random vectors distributed as

$$\mathbf{U} \sim El(\mathbf{0}, \mathbf{I}; g^{(p)}) \qquad \text{and} \qquad \mathbf{V} \sim El(\boldsymbol{\mu}, \Sigma; g^{(m)}).$$

Here $\mathbf{0}$ is the zero vector and $\mathbf{I}$ is the identity matrix.
Defining $\mathbf{Z}_{m\times 1} = \mathbf{D}_{m\times p}\mathbf{U}_{p\times 1} + \mathbf{V}_{m\times 1}$, the conditional density of $[\mathbf{Z}|\mathbf{U} > \mathbf{0}]$ will be of the form

$$h(\mathbf{z}|\boldsymbol{\mu}, \Sigma, \mathbf{D}; g^{(m)}) = 2^p f(\mathbf{z}|\boldsymbol{\mu}, \Sigma + \mathbf{D}\mathbf{D}^T; g^{(m)}) Pr(\mathbf{W} > \mathbf{0}|\mathbf{z}), \qquad (3.4)$$

where $f(\cdot)$ is the elliptical density function as that in (3.3), and

$$\mathbf{W}|\mathbf{z} \sim El\{\mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{z}_*, \mathbf{I} - \mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}; g^{(p)}_{q(\mathbf{z}_*)}\},$$

with

$$g^{(p)}_a(u) = \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{g(a+u; m+p)}{\int_0^\infty r^{p/2-1} g(a+r; m+p) dr},$$

$$q(\mathbf{z}_*) = \mathbf{z}_*^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{z}_*, \text{ and}$$

$$\mathbf{z}_* = \mathbf{z} - \boldsymbol{\mu}.$$

Proof: To derive (3.4), we need the following well-known results.
Suppose that $\mathbf{X} \sim El(\boldsymbol{\mu}, \Sigma; g^{(n)})$. Now partition $\mathbf{X}$ into $\mathbf{X}^T = (\mathbf{X}^T_{(1)}, \mathbf{X}^T_{(2)})$ of dimensions m and $n-m$ respectively, with the corresponding partitions of $\boldsymbol{\mu}$ and Σ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Lemma 3.1 (Theorem 2.16 of Fang et al., 1990.) If $\mathbf{B}$ is a non-singular $n \times m$ matrix and $\mathbf{v}$ is an $m \times 1$ vector, then

$$\mathbf{v} + \mathbf{B}^T\mathbf{X} \sim El(\mathbf{v} + \mathbf{B}^T\boldsymbol{\mu}, \mathbf{B}^T\Sigma\mathbf{B}; g^{(m)}).$$

□

Lemma 3.2 (Corollary of Fang et al., 1990: page 43.) The marginal distributions of $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ are given by:

$$\mathbf{X}_{(1)} \sim El(\boldsymbol{\mu}_{(1)}, \Sigma_{11}; g^{(m)}),$$
$$\mathbf{X}_{(2)} \sim El(\boldsymbol{\mu}_{(2)}, \Sigma_{22}; g^{(n-m)}).$$

□

Lemma 3.3 (Theorem 2.18 of Fang et al., 1990.) The conditional distribution $\mathbf{X}_{(1)}|\mathbf{X}_{(2)}$ is given by

$$\mathbf{X}_{(1)}|\mathbf{X}_{(2)} = \mathbf{x}_{(2)} \sim El(\boldsymbol{\mu}_{1.2}, \Sigma_{11.2}; g^{(m)}_{q(\mathbf{x}_{(2)})})$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}),$$
$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$
$$q(\mathbf{x}_{(2)}) = (\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)})^T\Sigma_{22}^{-1}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}),$$
$$g_a^{(m)}(u) = \frac{\Gamma(m/2)}{\pi^{m/2}} \frac{g(a+u; m+n)}{\int_0^\infty r^{m/2-1} g(a+r; m+n)dr}.$$

□

An alternative and convenient expression for (3.2) is the following

$$\begin{pmatrix} \mathbf{Z}_{m\times 1} \\ \mathbf{W}_{p\times 1} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{D}_{m\times p} \\ \mathbf{0}_{p\times m} & \mathbf{I}_{p\times p} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{m\times 1} \\ \mathbf{U}_{p\times 1} \end{pmatrix},$$

from which the p.d.f. (3.4) can be obtained by computing the conditional density $\mathbf{Z}|\mathbf{W} > \mathbf{0}$.

It can be easily verified from Lemma 3.1 that

$$\begin{pmatrix} \mathbf{Z}_{m\times 1} \\ \mathbf{W}_{p\times 1} \end{pmatrix} \sim El\left\{ \begin{pmatrix} \boldsymbol{\mu}_{m\times 1} \\ \mathbf{0}_{p\times 1} \end{pmatrix}, \begin{pmatrix} \Sigma_{m\times m} + \mathbf{D}\mathbf{D}^T_{m\times m} & \mathbf{D}_{m\times p} \\ \mathbf{D}^T_{p\times m} & \mathbf{I}_{p\times p} \end{pmatrix}; g^{(m+p)} \right\}.$$

It follows easily using Lemma 3.2 that

$$\mathbf{Z} \sim El(\boldsymbol{\mu}, \Sigma + \mathbf{D}\mathbf{D}^T; g^{(m)}),$$
$$\mathbf{W} \sim El(\mathbf{0}, \mathbf{I}; g^{(p)}).$$

Symmetry of the elliptical distribution and the Bayes theorem implies that

$$h(\mathbf{z}|\mathbf{W} > \mathbf{0}) = 2^p f\{\mathbf{z}|\boldsymbol{\mu}, \Sigma + (\mathbf{D}\mathbf{D}^T); g^{(m)}\} Pr(\mathbf{W} > \mathbf{0}|\mathbf{z}).$$

The proof is completed by specifying the conditional density of $\mathbf{W}|\mathbf{Z}$ using Lemma 3.3:

$$\mathbf{W}|\mathbf{Z} = \mathbf{z} \sim El\{\mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{z}_*, \mathbf{I}_{p\times p} - \mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}; g^{(p)}_{q(\mathbf{z}_*)}\},$$

where

$$q(\mathbf{z}_*) = \mathbf{z}_*^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{z}_*,$$
$$\mathbf{z}_* = \mathbf{z} - \boldsymbol{\mu}, \text{ and}$$
$$g_a^{(p)}(u) = \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{g(a+u; m+p)}{\int_0^\infty r^{p/2-1} g(a+r; m+p) dr}.$$

□

Non-singularity of $(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}$ is a prerequisite for ensuring the existence of the resulting density (3.4). The matrix $\mathbf{D}$, in a broad sense, controls the degree of asymmetry of the density via the probability function $Pr(\mathbf{W} > \mathbf{0}|\mathbf{z})$. Henceforth $\mathbf{D}$ will be interpreted as the skewness parameter and $Pr(\mathbf{W} > \mathbf{0}|\mathbf{z})$ as the 'skewing function' (following Cartinhour, 1990). It is clear that the particular case $\mathbf{D} = \mathbf{0}$ corresponds to the one of the elliptical distribution $El(\boldsymbol{\mu}, \Sigma; g^{(m)})$. Consequently, the random vector $\mathbf{Y} = [\mathbf{Z} > \mathbf{0}]$ can reasonably be regarded as having an m-dimensional *skew-elliptical distribution.* For brevity, the symbols $SE(\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p}; g^{(m)})$ are employed to denote the sampling density in (3.4). Note that, in general, the quantities $\boldsymbol{\mu}$ and Σ are not the mean and the scale matrix of $\mathbf{Y}$ as the density may not be symmetric with respect to $\boldsymbol{\mu}$.

The use of an elliptical model in the development of (3.4) is motivated by its desirable property of including thin and thick tailed distributions as special cases. As a result, in addition to the obvious increased flexibility in skewness, the family $SE(\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p}; g^{(m)})$ should allow for a variety of tail thickness. In the case where $\mathbf{D} = \delta\mathbf{I}$, the proposed class closely parallels to the one given in

Branco and Dey (2001). Moreover, it agrees with the skew elliptical densities mentioned in Sahu *et al.* (2003) when $\mathbf{D}$ is diagonal of order m. Therefore the present class includes the earlier version obtained by Sahu *et al.* (2003) as a special case. Another appealing feature of the skew elliptical in (3.4) is its coherence under marginalization operation, i.e. it has marginal distributions that still belong to the same family. This is essentially an implicit result in the genesis of the distribution. Although the skewing function $Pr(\mathbf{W} > \mathbf{0}|\mathbf{z})$ may prove to be hard to evaluate, it need not be computed for practical MCMC model fitting. Such details will be clear in the subsequent chapters. Summing up, this new skew distribution should be valuable in modeling multivariate random phenomena which display both skewness and kurtosis.

3.4 Skew-normal distributions

As pointed out in the last section, construction (3.2) is a vigorous technical tool for transforming a symmetric distribution into a skewed one. Clearly, joint consideration of asymmetry and tail behavior can now be achieved by applying this method to a suitable fat or thin tailed distribution. From the inferential viewpoint it means that the resulting skewed distribution is made up of two components, $\mathbf{DU}$ and $\mathbf{V}$ in the preceding notations. Skewness is driven only by a single vector $\mathbf{U}$ and its sensitivity is dependent on $\mathbf{D}$. Although it is not obvious in the context, operation (3.2) does have an effect on other distributional characteristics. The principal purpose of the current section is to examine how procedure (3.2) influences the shape of the skewed density. Since normal distribution has been the standard point of reference for many characteristic measurements, attention will be held on its skewed counterpart from this time onwards. After presenting the density function of the m-dimensional version, this section will focus on the general univariate case. Specifically, mathematical moments and some properties of the latter will be presented in streamlined form.

3.4.1 Multivariate skew-normal distributions

As an immediate use of Theorem 3.1, consider the particular case $g(u;m) = e^{-u/2}$. Now, since the generator function simplifies to $g^{(m)}(u) = (2\pi)^{-m/2}e^{-u/2}$ and $g^{(p)}_{q(\mathbf{Z}_*)}(u)$ is free of $q(\mathbf{z}_*)$, it is straightforward to verify that the joint density of $\mathbf{Y} = [\mathbf{Z}|\mathbf{U} > \mathbf{0}]$ is of the form

$$h(\mathbf{y}|\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p}) = 2^p|\Sigma + \mathbf{D}\mathbf{D}^T|^{-1/2}\phi_m\{(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\}Pr(\mathbf{W} > \mathbf{0}|\mathbf{y}), \tag{3.5}$$

where ϕ_m is the multivariate normal density of $N_m(\mathbf{0}, \mathbf{I})$, and

$$\mathbf{W}|\mathbf{Y} = \mathbf{y} \sim N_p\{\mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}(\mathbf{y} - \boldsymbol{\mu}), \mathbf{I}_{p\times p} - \mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}\}.$$

It follows that $\mathbf{Y}$ has a multivariate skew normal distribution, indicated henceforth by the notation $\mathbf{Y} \sim SN_m(\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p})$. As expected the original normal density is retrieved when $\mathbf{D} = \mathbf{0}$. Conversely, deviation of the parameter $\mathbf{D}$ from $\mathbf{0}$ measures the departure of the distribution from normality. Therefore the above family nests the normal distribution as a proper member and permits a continuous departure from normality to non-normality.

3.4.2 Univariate skew-normal distributions

Density function

Specifying $m = 1$ in (3.5), the matrix D becomes a column vector $\boldsymbol{\delta}^T = (\delta_1, \cdots, \delta_p) \in \Re^p$ and Σ reduces to a scalar σ^2. In this case, Y is a univariate skew normal variate with density function given by

$$h(y|\mu, \sigma^2, \boldsymbol{\delta}) = \frac{2^p}{\sqrt{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2}}\phi\left(\frac{y - \mu}{\sqrt{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2}}\right) Pr(\mathbf{W} > \mathbf{0}|y), \tag{3.6}$$

where

$$\mathbf{W}|Y = y \sim N_p\left(\frac{y - \mu}{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2}\boldsymbol{\delta}, \mathbf{I}_{p\times p} - \frac{1}{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2}\boldsymbol{\delta}\boldsymbol{\delta}^T\right).$$

In what follows, (3.6) will be referred to the general form of the univariate skew normal distribution.

Moments

As mentioned previously computation of the skewing function $Pr(\mathbf{W} > \mathbf{0}|y)$ can be obstructive. As a consequence, direct evaluation of the moments of the general univariate skew normal distribution will not be straightforward. A convenient way of proceeding is the following one. According to the representation (3.2), $Y \sim SN(\mu, \sigma^2, \boldsymbol{\delta})$ is the upshot of a linear combination of independent normal and standard half normal random variables. That is

$$\begin{aligned} Y &= \boldsymbol{\delta}^T \mathbf{U}_{p\times 1} + V \qquad\qquad \mathbf{U} > \mathbf{0}, \\ &= \delta_1 U_1 + \cdots + \delta_p U_p + V \end{aligned} \tag{3.7}$$

where

$U_1, \cdots, U_p$ are iid standard half normal,
$V \sim N(\mu, \sigma^2)$, and
U's and V are independent.

Using this fact as well as the properties of moment generating function, expressions for the mean and the central moments of orders two through four can be explicitly evaluated.

Result 1 The random variable Y has

$$E(Y) = \mu + (\delta_1 + \cdots + \delta_p)\sqrt{\frac{2}{\pi}},$$

$$Var(Y) = \sigma^2 + (\delta_1^2 + \cdots + \delta_p^2)\left(1 - \frac{2}{\pi}\right),$$

$$m_3(Y) = E[\{Y - E(Y)\}^3] = (\delta_1^3 + \cdots + \delta_p^3)\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right),$$

$$\begin{aligned} m_4(Y) &= E[\{Y - E(Y)\}^4] \\ &= 3\sigma^4 + (\delta_1^4 + \cdots + \delta_p^4)\left\{3 - \frac{4}{\pi}\left(\frac{3}{\pi} + 1\right)\right\} \\ &\quad + 6\left\{(\delta_1^2 + \cdots + \delta_p^2)\left(1 - \frac{2}{\pi}\right)\sigma^2 + \delta_1^2 \cdots \delta_p^2\left(1 - \frac{2}{\pi}\right)^2\right\}. \end{aligned}$$

Proof: The derivation of the above equations is straightforward but lengthy, the basic steps are presented as follows. For convenience, we provide the details for the case

$p = 2$, the proof for the general case is similar. Now, as a result of (3.7), the moment generating function of Y can be written as

$$M_Y(t) = M_{Z_1}(t)M_{Z_2}(t)M_V(t).$$

Here $Z_i = \delta_i U_i$, $i = 1, 2$. In general, $M_Y(t)$ has no closed form expression since $M_{Z_1}(t)$ and $M_{Z_2}(t)$ do not lend themselves to explicit computation. Nevertheless, the above formulation can still be used as an indirect tool to obtain the moments.
Together with the fact $\frac{d^r}{dt^r} M_X(t)\big|_{t=0} = E(X^r)$, it is easy to check that the central moments of Y satisfy the relationships

$$\begin{aligned}
E(Y) &= E(Z_1) + E(Z_2) + E(V),\\
Var(Y) &= Var(Z_1) + Var(Z_2) + Var(V),\\
m_3(Y) &= m_3(Z_1) + m_3(Z_2) + m_3(V),\\
m_4(Y) &= m_4(Z_1) + m_4(Z_2) + m_4(V) +\\
&\qquad 6\{Var(Z_1)Var(Z_2) + Var(Z_1)Var(V) + Var(Z_2)Var(V)\},
\end{aligned}$$

where $m_i(X) = E[\{X - E(X)\}^i]$. These results essentially reduce the problem to the evaluation of moments of normal and standard half normal distributions.
The r-th noncentral moments of Z_i, $i = 1, 2$, is

$$E(Z_i^r) = \begin{cases} \delta^r \frac{1}{\pi} 2^{r/2} \left(\frac{r-1}{2}\right)! & \text{for odd } r,\\ \delta^r 2^{-(r-2)/2} 1 \cdot 3 \cdot 5 \cdots (r-1) & \text{for even } r. \end{cases}$$

The r-th central moments of V is

$$E\{(V-\mu)^r\} = \begin{cases} 0 & \text{for odd } r,\\ \frac{r!}{(r/2)!} \frac{\sigma^r}{2^{r/2}} & \text{for even } r. \end{cases}$$

The proof follows immediately by direct substitution.

□

In order to illustrate the influence of the parameter $\boldsymbol{\delta}$, it is necessary to adopt some suitable measures of skewness and kurtosis. To this end, a natural choice is the two classical distributional measurements stated in the next definition.

Definition 3.1 The skewness and kurtosis measures of a random variable X are respectively defined as the third and fourth standardized central moments of X, i.e.

$$Sk(X) = \frac{E[\{X - E(X)\}^3]}{\{Var(X)\}^{3/2}}, \qquad \text{and} \qquad Ku(X) = \frac{E[\{X - E(X)\}^4]}{\{Var(X)\}^2} - 3.$$

□

Thus skewness and kurtosis of Y can be readily obtained from the moments reported in Result 1. Intuitively, symmetrical distributions have skewness measure equal zero, positive values correspond to distributions skewed to the right and negative values to those skewed to the left. Kurtosis, on the other hand, measures the degree of flatness of a density. Intrinsically positive kurtosis indicates peaked center and negative one signifies flat center relative to the normal curve.

An elementary calculation demonstrates that the skewness approaches its supremum (infimum) as $\delta_i \to \infty(-\infty)$, $i = 1, \cdots, p$, with

$$\sup\{Sk(Y)\} = -\inf\{Sk(Y)\} = \sqrt{2}(4-\pi)(\pi-2)^{-3/2} \simeq 0.9953.$$

Similarly, the bounds of the kurtosis can be shown to be

$$0 \leq Ku(Y) \leq (3\pi^2 - 4\pi - 12)(\pi-2)^{-2} - 3 \quad (\simeq 0.8692).$$

Therefore it may come to a conclusion that, with other parameters fixed, $\boldsymbol{\delta}$ in (3.6) can only produce more central peakedness than those in the original distribution.

In addition to creating some savings in moment calculations, relation (3.7) leads to an efficient algorithm for computer generation of skew normal random samples. The method can be described as follows. First, sample a p-dimensional vector $\mathbf{U}$ from $N_p(\mathbf{0}, \mathbf{I})$ and a scalar V from $N(\mu, \sigma^2)$. Then a random number Y from density (3.6) is obtained by setting

$$Y = \boldsymbol{\delta}^T |\mathbf{U}_{p\times 1}| + V.$$

This construction avoids rejection of sampling. The role played by $\boldsymbol{\delta}$ will be further highlighted in the coming sections.

Some simple properties

Some basic properties of the general univariate skew normal distribution are presented as below.

Property 1 The density (3.6) reduces properly to the $N(\mu, \sigma^2)$ density when $\boldsymbol{\delta} = \mathbf{0}$.

Property 2 Reversing the sign of $\boldsymbol{\delta}$ and μ in (3.6) yields the density of $-Y$, i.e. the distribution $SN(-\mu, \sigma^2, -\boldsymbol{\delta})$ is the reflection of the distribution of $SN(\mu, \sigma^2, \boldsymbol{\delta})$ about $y = 0$.

Property 3 The way in which $\boldsymbol{\delta}$ intervenes in the central moments implies that

$$Sk(Y|\sigma^2, -\boldsymbol{\delta}) = -Sk(Y|\sigma^2, \boldsymbol{\delta}) \text{ and } Ku(Y|\sigma^2, -\boldsymbol{\delta}) = Ku(Y|\sigma^2, \boldsymbol{\delta}).$$

Property 4 The parameter $\boldsymbol{\delta}$ regulates skewness, which is positive if $\Lambda > 0$ and negative if $\Lambda < 0$ where $\Lambda = \sum_{i=1}^{p} \delta_i^3$. Clearly, symmetric distribution can be obtained by taking $\Lambda = 0$.

Property 5 The skewness $Sk(Y)$ is an increasing function of δ_i while the kurtosis $Ku(Y)$ is an increasing function of $|\delta_i|$, $i = 1, \cdots, p$.

Property 6 Large $\boldsymbol{\delta}$ will have momentous impact on the spread on (3.6) as $Var(Y)$ grows without bound with the absolute value of δ_i, $i = 1, \cdots, p$.

Property 7 Since $\frac{d \log h(y)}{dy}$ is a decreasing function of y it follows that the density (3.6) is unimodal.

Property 8 The mode of $SN(\mu, \sigma^2, \boldsymbol{\delta})$ is at the right of μ when $\sum_{i=1}^{p} \delta_i^3 > 0$ and vice versa. Except for the symmetric cases, it is in general not possible to find the mode analytically.

Furthermore, the distribution function of Y does not admit a closed form expression.

3.4.3 Two specific cases of univariate skew-normal distributions

Generally speaking, a one-parameter distribution can model only one empirical characteristic while greater flexibility is necessarily accompanied by increasing

complexity in probability distribution. Therefore, the choice of p in (3.6) should depend on the level of difficulty in modeling the distributional characteristics in a body of data. From a pragmatic perspective, normal distribution ($p = 0$) is often sufficient for reflecting the structure underlying a population distribution. Other selections of p can be useful for analyzing data with the presence of possible skewness or kurtosis. For ease of exposition, only two particular cases of (3.6) are examined extensively for the rest of the research.

1. The case $p = 1$ is of special interest, since it coincides with the univariate skew normal distribution obtained by Sahu *et al.* (2003). After some straightforward computations, it follows that the density of Y is of the form

$$h(y|\mu, \sigma^2, \delta) = \frac{2}{\sqrt{\sigma^2 + \delta^2}} \phi \left\{ \frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} \right\} \Phi \left\{ \frac{\delta}{\sigma} \frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} \right\}. \tag{3.8}$$

   We write $Y \sim \mathrm{SN}_{\mathrm{sdb}}(\mu, \sigma^2, \delta)$ for future reference. By way of illustration, Figure 3.1 depicts the shape of the probability density function for certain values of δ. As the diagram indicates, the effect of increasing δ is to magnify both the dispersion and asymmetry of the distribution.

2. Considering $p = 2$, it is not difficult to verify that

$$h\left\{ y|\mu, \sigma^2, \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \right\} = \frac{4}{\sqrt{\sigma^2 + \delta_1^2 + \delta_2^2}} \phi \left(\frac{y - \mu}{\sqrt{\sigma^2 + \delta_1^2 + \delta_2^2}} \right) F(\mathbf{0}), \tag{3.9}$$

   where F stands for the cumulative density function of the bivariate normal

$$N_2 \left\{ -\frac{y - \mu}{\sigma^2 + \delta_1^2 + \delta_2^2} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}, \frac{1}{\sigma^2 + \delta_1^2 + \delta_2^2} \begin{pmatrix} \sigma^2 + \delta_2^2 & -\delta_1\delta_2 \\ -\delta_1\delta_2 & \sigma^2 + \delta_1^2 \end{pmatrix} \right\}.$$

   Throughout the sequel of the paper, we shall denote this distribution by $\mathrm{SN}_{\mathrm{new}}(\mu, \sigma^2, \binom{\delta_1}{\delta_2})$. Examples of the densities for different combinations of values for δ_1 and δ_2 are presented in Figure 3.2. Observe that the graphs plotted below the diagonal are duplications of those above the diagonal. This is a consequence of the exchangeability property of δ_1 and δ_2.

Besides the discrepancy in the level of algebraic complications, densities (3.8) and (3.9) differ in the sense that skewness of the latter is driven by differences

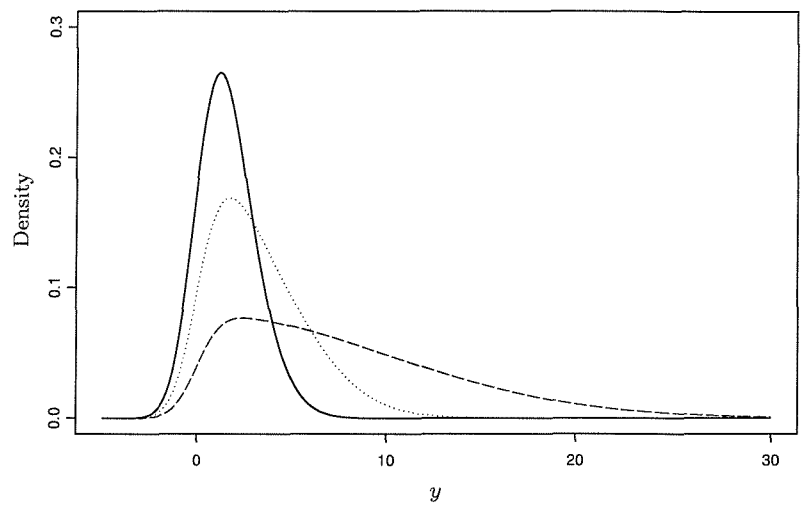

Figure 3.1: Plot of the density functions of $\mathrm{SN}_{\mathrm{sdb}}(0, 1, \delta)$; solid line is for $\delta = 2$, dotted line is for $\delta = 4$, and dashed line is for $\delta = 10$.

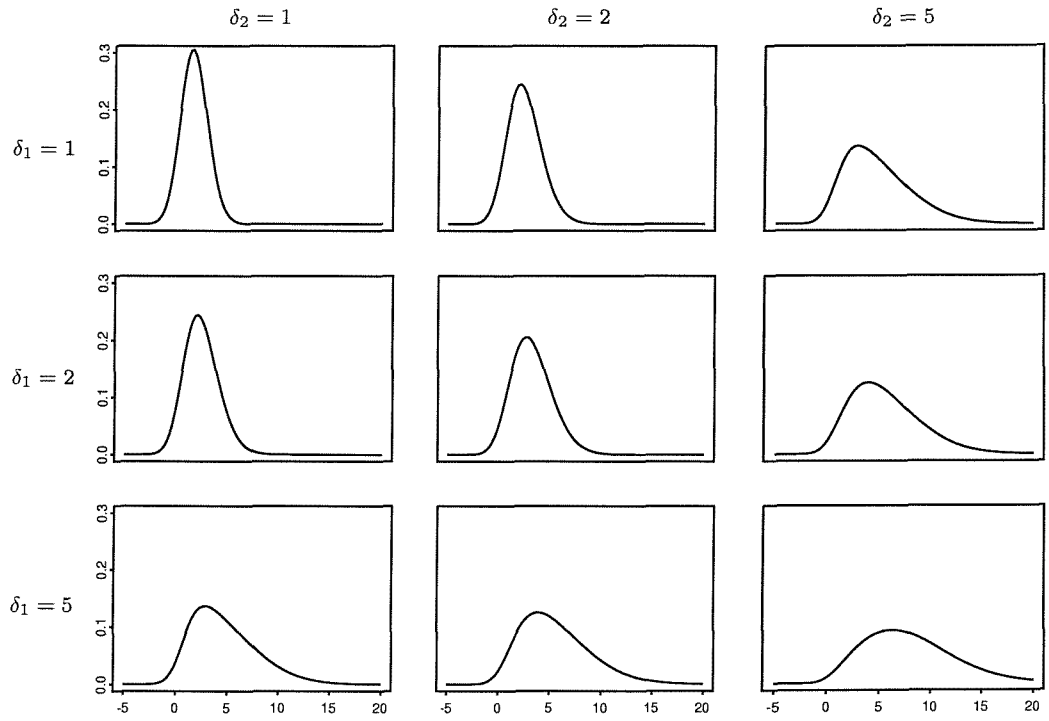

Figure 3.2: Plot of the density functions of $\mathrm{SN}_{\mathrm{new}}(0, 1, \binom{\delta_1}{\delta_2})$ for $\delta_1, \delta_2 = 1, 2, 5$.

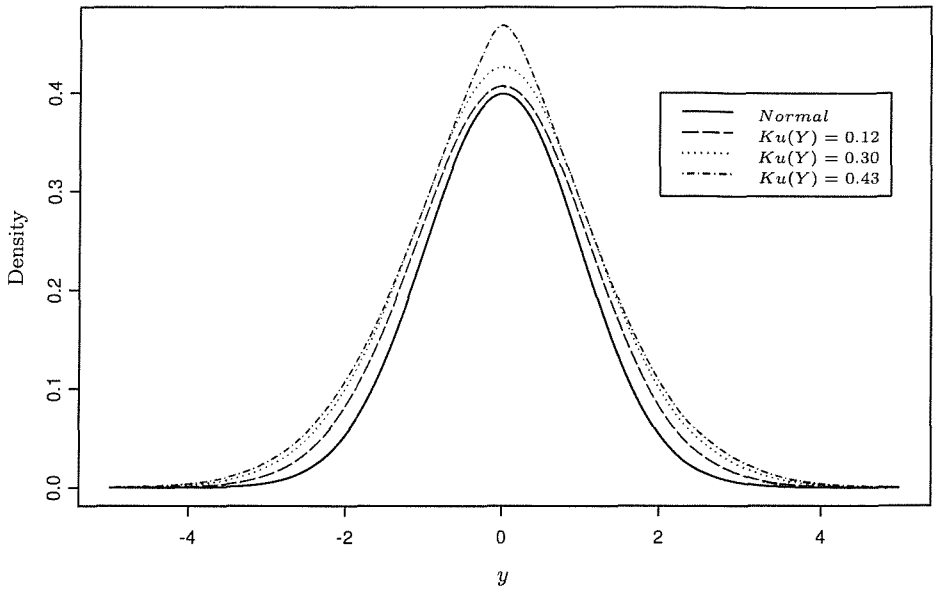

Figure 3.3: Plot of the density functions of $\mathrm{SN}_{\mathrm{new}}(0, \sigma^2, \binom{\delta_1}{\delta_2})$ where $\delta_1 = -\delta_2$ and $Var(Y) = 1$.

in $(\delta_1, -\delta_2)$, while the former by a single parameter δ. Trivially, density (3.9) reduces to the one in (3.8) when $\delta_1 = 0$ or when $\delta_2 = 0$. As it might be expected, the major improvement involving the presence of an additional parameter is the flexibility in kurtosis variation. The three-parameter density (3.8) imposes some restraints on the kurtosis as soon as the skewness is fixed. In contrast, a broad range of the kurtosis of (3.9) can be covered by appropriate choices of δ_1 and δ_2 for any degree of skewness. To gain more insight in the impact of (3.2) on the kurtosis, consider Figure 3.3 which provides plots of (3.9) when δ_1 is set to equal $-\delta_2$ with variance of Y appointed at unity. All graphs in the diagram illustrate greater peakedness around the center as compared to the normal density.

3.5 Two-piece skew-normal distributions

Skewed distributions can be originated by differing the scale of a symmetric distributions on each sides of its mode, as observed by Runnenberg (1978) .

Obviously, a family of skew normal distributions distinct from those discussed earlier can be obtained by employing this method on the normal distribution. The resulting distribution, namely the two piece skew normal distribution, has been studied by many authors, including Gibbons and Mylroie (1973), John (1982), Kimber (1985) and Kimber and Jeynes (1987). See also Fernandez and Steel (1998) for a generalization. The density function of the two piece skew normal distribution is defined as follows

$$h(y|\mu,\sigma^2,\delta) = \frac{2}{\sigma(\delta+1/\delta)}\left\{\phi\left(\frac{y-\mu}{\sigma\delta}\right) I(y-\mu \geq 0) + \phi\left(\frac{\delta(y-\mu)}{\sigma}\right) I(y-\mu < 0)\right\}. \tag{3.10}$$

where I is an indicator function with $I(Q) = 1$ if Q is true and equals 0 otherwise. For notation ease, we say that the random variable Y is of the class $\mathrm{SN}_{\mathrm{tpn}}(\mu,\sigma^2,\delta)$ henceforth. The parameter $\delta \in (0,\infty)$ controls the allocation of probability mass to each side of the mode. More formally, it can be shown that

$$\frac{Pr(Y \geq \mu)}{Pr(Y < \mu)} = \delta^2.$$

Clearly, the normal distribution is a special case ($\delta = 1$) and the half normal distribution is a limiting case. Sketches in Figure 3.4 display the effect of various values of δ on the shape of the distribution.

The mode of the distribution (3.10) is retained at μ for any value of δ. Using the results from Fernandez and Steel (1998), the expected value, variance and

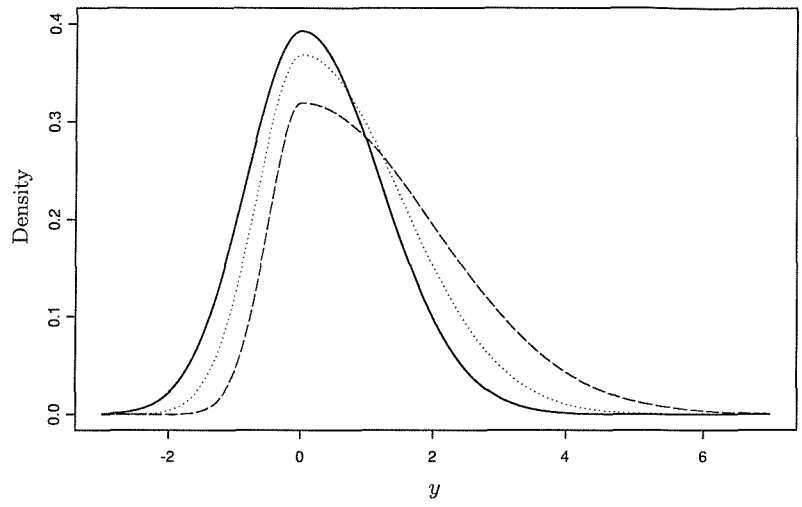

Figure 3.4: The density functions of $\mathrm{SN}_{\mathrm{tpn}}(0, 1, \delta)$. The solid line is for $\delta = 1.2$, the dotted line is for $\delta = 1.5$ and the dashed line is for $\delta = 2$.

central moments of orders three and four of (3.10) are respectively obtained as

$$
\begin{aligned}
E(Y) &= \mu + \sqrt{\frac{2}{\pi}}\sigma\left(\delta - \frac{1}{\delta}\right), \\
Var(Y) &= \sigma^2\left\{\left(1 - \frac{2}{\pi}\right)\left(\delta^2 + \frac{1}{\delta^2}\right) - 1 + \frac{4}{\pi}\right\}, \\
m_3(Y) &= \sqrt{\frac{2}{\pi}}\sigma^3\left\{\left(\frac{4}{\pi} - 1\right)\left(\delta^3 - \frac{1}{\delta^3}\right) - 4\left(\frac{3}{\pi} - 1\right)\left(\delta - \frac{1}{\delta}\right)\right\}, \\
m_4(Y) &= \sigma^4\left\{\left(3 - \frac{4}{\pi} - \frac{12}{\pi^2}\right)\left(\delta^4 + \frac{1}{\delta^4}\right) - \right. \\
&\qquad \left.\left(3 + \frac{4}{\pi} - \frac{48}{\pi^2}\right)\left(\delta^2 + \frac{1}{\delta^2}\right) + 3 + \frac{16}{\pi} - \frac{72}{\pi^2}\right\}.
\end{aligned}
$$

Expressions for the classical skewness and kurtosis measures described in Definition 3.1 can be immediately written down by using the moments above. Simple algebraic manipulations reveal that skewness and kurtosis of the two piece skew normal distribution have the same ranges as those in the univariate skew normal distribution (3.6). This might be expected since both families are skew extension of the normal density and admit the same limiting distribution. A useful stochastic representation of (3.10), which provides an easy way for ran-

dom samples simulation, is the following one. Let X be a standard normal random variable and define Y by

$$Y = \begin{cases} \mu + \sigma X/\delta & \text{with probability } 1/(\delta^2+1) \\ \mu + \sigma X\delta & \text{with probablity } \delta^2/(\delta^2+1) \end{cases}.$$

Then Y has density function (3.10).

3.6 Graphical comparisons

The specific aim of the current section is to compare different variants of skew normal distributions discussed earlier by means of some graphical plots. In short, the distributions to be considered are

1. $\mathrm{SN}_{\mathrm{sdb}}(\mu, \sigma^2, \delta)$: the Sahu *et al.* (2003) distribution displayed in (3.8).
2. $\mathrm{SN}_{\mathrm{new}}(\mu, \sigma^2, \binom{\delta_1}{\delta_2})$: the new distribution specified by (3.9).
3. $\mathrm{SN}_{\mathrm{tpn}}(\mu, \sigma^2, \delta)$: the two piece skew normal distribution with density (3.10).

The disparities of these skewed models in distributional structure may be illustrated effectively by drawing their densities in the same diagram, in which they admit the same amount of skewness. Yet, in order to have a fair comparison, it is necessary to require common mean and variance across the distributions. Figure 3.5 pictures the densities for a selection of values of mean, variance and skewness. Note that, unlike $\mathrm{SN}_{\mathrm{sdb}}$ and $\mathrm{SN}_{\mathrm{tpn}}$ distributions, there is a series of $\mathrm{SN}_{\mathrm{new}}$ densities complying with the imposition. In spite of that, only two of these densities are plotted in the figure so as to enhance visualization. The supplementary flexibilities of $\mathrm{SN}_{\mathrm{new}}$ over $\mathrm{SN}_{\mathrm{sdb}}$ in terms of height and tails controls should be apparent from the graphic display. As one can anticipate, the behavior of $\mathrm{SN}_{\mathrm{tpn}}$ exhibits a manifest difference from the other densities. This is because $\mathrm{SN}_{\mathrm{tpn}}$ has a much lighter tail and its distinct style of descending from the mode. Similar figures could be constructed for other combinations of characteristic measures.

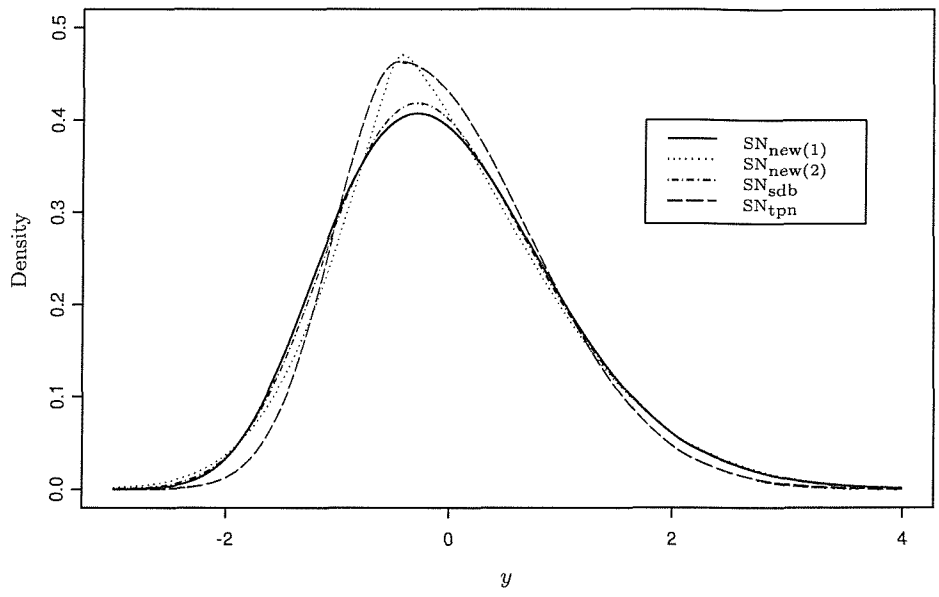

Figure 3.5: Plots of the density functions of various skew-normal distributions. All distributions are scaled to have zero mean, unit variance and $Sk(Y) = 0.5$.

It is of interest to acquire a visual summary of the relationship between the skewing parameter and the degree of asymmetry of the distributions. On this basis, Figures 3.6 and 3.7 delineate the level of skewness measure $Sk(Y)$ as the parameter $\boldsymbol{\delta}$ changes for representative values of σ^2. Although Figure 3.7 only copes with positive values of δ_1, analogous plots for the negative domain can be obtained by a simple 180° rotation. It appears from Figure 3.6 that $\mathrm{SN}_{\mathrm{tpn}}$ is very sensitive to variations in δ. In fact, a wide range of its skewness can be covered for δ varying in $(0.2, 2)$. The rate with which $\mathrm{SN}_{\mathrm{sdb}}$ diverges from symmetry as $|\delta|$ increases is substantially influenced by the parameter σ^2. Smaller values of σ^2 will be associated with greater steeply sloped skewness curves and vice versa. Evidently the parameter has a similar impact on $\mathrm{SN}_{\mathrm{new}}$, as shown in Figure 3.7. It is important to note that a rise in $|\delta_1^3 + \delta_2^3|$ does not necessarily mean a greater asymmetry, especially when $\mathrm{SN}_{\mathrm{new}}$ is already highly skewed.

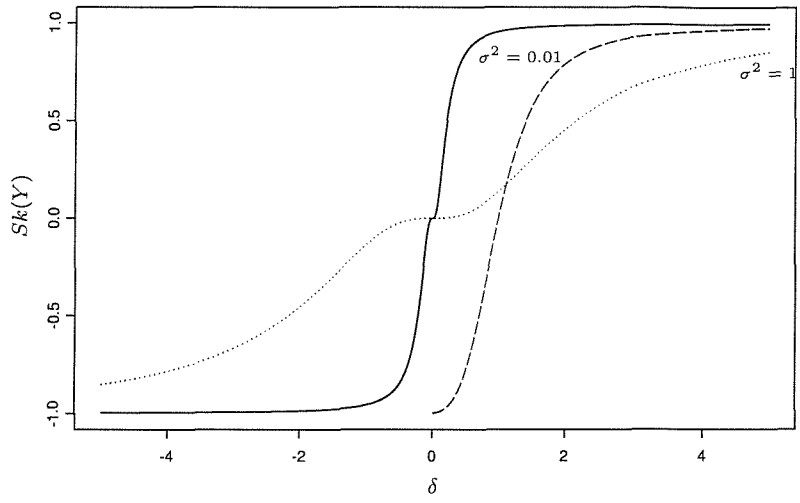

Figure 3.6: Plots of the skewness measure $Sk(Y)$ against δ. Dashed line is for $\mathrm{SN}_{\mathrm{tpn}}$; solid line is for $\mathrm{SN}_{\mathrm{sdb}}$ with $\sigma^2 = 0.01$ and the dotted line is for $\sigma^2 = 1$. Note that skewness of $\mathrm{SN}_{\mathrm{tpn}}$ does not depend on σ^2.

Figure 3.8 gives an additional insight into the achievable kurtosis $Ku(Y)$ as a function of skewness $Sk(Y)$. For $\mathrm{SN}_{\mathrm{tpn}}$ and $\mathrm{SN}_{\mathrm{sdb}}$ distributions, greater asymmetry will inevitably result in larger values for the kurtosis. Similarly, smaller magnitude of skewness will correspond to less central peakedness. Nonetheless the two distributions depart from normality in a quite different manner. The gap between the dotted and solid lines shows the advantage of $\mathrm{SN}_{\mathrm{new}}$ over $\mathrm{SN}_{\mathrm{sdb}}$ in the form of kurtosis variation. It is seen that the advantage is most perceptible for near normal cases and gradually melted away as asymmetry increases.

3.7 Further generalizations

There are many possible ways of generalizing the skew elliptical distribution (3.4). Without going into details, some feasibilities are listed as follows.

- From representation (3.2), it may be claimed that skewness is instigated by some unobserved additive random effects $\mathbf{U}$ which were truncated at a

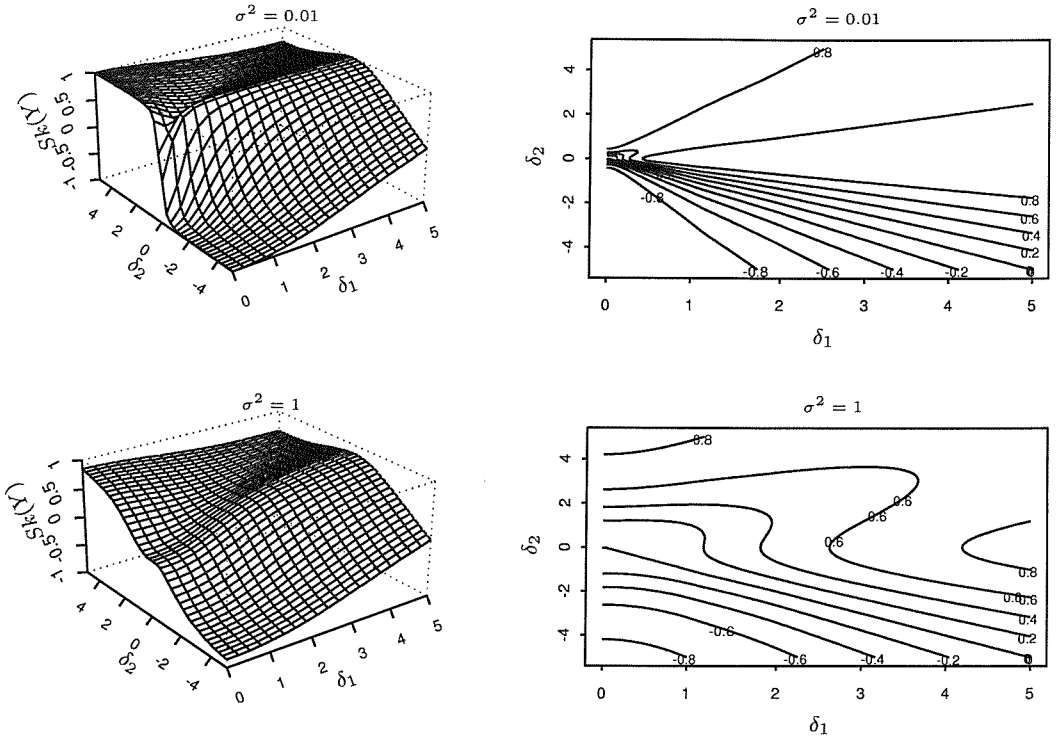

Figure 3.7: Surface and contour plots of the skewness measure $Sk(Y)$ of $\mathrm{SN}_{\mathrm{new}}$ against δ_1 and δ_2 for two different values of σ^2.

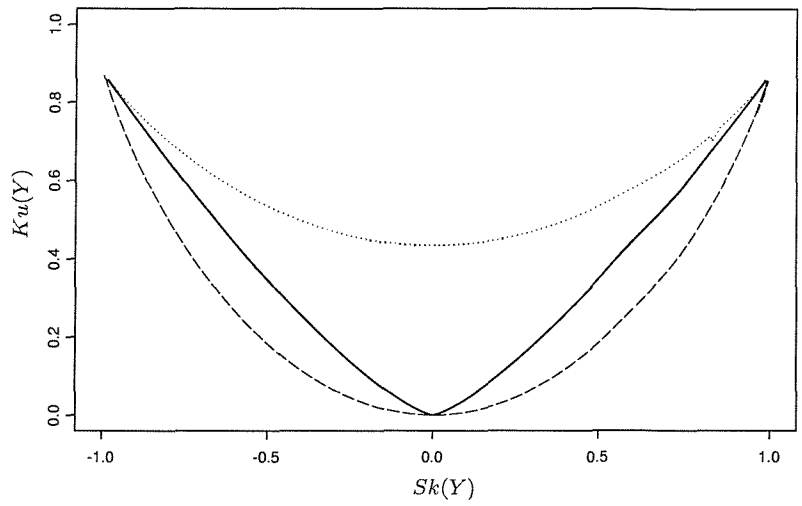

Figure 3.8: Plots of kurtosis $Ku(Y)$ versus skewness $Sk(Y)$; dashed line is for $\mathrm{SN}_{\mathrm{tpn}}$, solid line is for $\mathrm{SN}_{\mathrm{sdb}}$, and dotted line is for the maximum achievable kurtosis of $\mathrm{SN}_{\mathrm{new}}$.

specific threshold. This suggests that further flexibility should be annexed to the model (3.4) by adopting a more general threshold or permitting broader style of truncation on $\mathbf{U}$.

- The random variables $\mathbf{U}$ and $\mathbf{V}$ used in the development of elliptical distribution were assumed to have come from the same standard family. Allowing a combination of assorted distributions will result in new classes of skewed distributions. See Azzalini (1985) for some related ideas.

- A natural way of extending (3.4) is to utilize a comprehensive transformation mechanism admitting the representation: $\mathbf{DU} + \mathbf{BV}$. Obviously, a joint density for $\mathbf{U}$ and $\mathbf{V}$ can be used instead for the sake of releasing the independence assumption, which in turn will lead to extra level of generalization.

If the research was to be extended, it would be interesting to perform further analysis on each circumstance and consequently merit our consideration.

3.8 Conclusion

A new class of univariate skewed normal distributions is obtained by using simple transformation and conditioning. The family represents a mathematically tractable extension of the normal density, with the addition of a vector of parameters $\boldsymbol{\delta}$ to regulate distributional shape. Our focus in this chapter has been concentrated on the scalar and the 2 dimensional $\boldsymbol{\delta}$ cases. We find the latter case quite appealing for some of its attractive features:

1. It contains the normal distribution by strict inclusion, thus allowing a smooth transition from normality to non-normality.

2. It admits the Sahu *et al.* skew normal density (equivalent to the scalar $\boldsymbol{\delta}$ case) as a proper member, but possesses an extra parameter to account for kurtosis.

3. It is a flexible unimodal density that is able to reflect practical values of skewness and some levels of non-normal peakedness.

4. Simulating random samples from this distribution is straightforward.

Therefore, the proposed four-parameter distribution is potentially useful for data modeling, statistical analysis and robustness studies of normal theory methods.

Chapter 4

Applications to linear regression models

4.1 Introduction

The asymmetry and non-normal peakedness of many practical data sets can contaminate empirical results of normal theory statistics. Routine use of methods resulting from more robust distributional assumptions has long been hampered by the corresponding technical difficulties in likelihood evaluation. Advances in readily-available computer power and Markov Chain Monte Carlo techniques in recent years have diminished the computational burden and hence broadened the scope of probability models that can be fitted to real data. The work, starting from this chapter, is intended to have a very practical focus, exploring the applications of the skew normal distribution from the viewpoint of reliability analysis. More specifically, our objective is to examine the suitability of the skew normal distribution in fitting linear statistical models outside normality. We start our expository analysis with linear regression models under independent skewed error structure.

Our plan for the chapter is the following. Section 4.2 provides a concise description of the statistical model where residual follows the $\mathrm{SN}_{\mathrm{new}}$ distribution defined in (3.9). We cast the model as a problem in Bayesian inference and employ the MCMC method known as the Gibbs sampler (see Section 2.7), to overcome the difficulties in computation. All the needed conditional posterior

distributions are derived in Section 4.3. In Section 4.4, we justify and demonstrate the usefulness of the skewed model through some numerical examples. The last section offers some concluding remarks. To enhance readability BUGS routines developed for implementing the numerical procedures are deferred to the Appendix.

4.2 Linear regression models

Regression models have become an integral component of many data analyses. A quantity of interest is observed and its value is known to be affected by other quantities, called explanatory variables or covariates. This quantity is referred to as the response or outcome variable and is regarded as random due to the sampling process and the natural variation of the population. Generally, the primary goal of setting up a regression study is to discover the association between the response variable and one or more explanatory variables using sample data. In nearly all applications of regression analysis, the precise relationship between these variables will never be known. It is preferable and usually appropriate to adopt a reasonably simple formulation so as to characterize this dependency. In this chapter, we restrict our attention to the simplest case where the influence of the explanatory variables is linear and additive on the mean of the response variable, which is assumed to be continuous. However, instead of considering the ordinary assumption of normality, we use the skew normal $\mathrm{SN}_{\mathrm{new}}$ defined in (3.9) to model the conditional distribution of the response given the covariates. From now on, such a model will be known as the skew normal linear regression model.

Suppose the observed data y_i, $i = 1, \cdots, n$, are an independent sample generated from the regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{4.1}$$

where $\mathbf{x}_i \in \Re^k$ are the values of k explanatory variables for the i-th observation and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)^T$ is a vector of regression parameters associated with these variables. The term ε_i is called the error or residual, representing the

variation of Y_i given the regressors $\mathbf{x}_i$ around its expected value. The stochastic environment considered here is

$$\varepsilon_i \sim \mathrm{SN}_{\mathrm{new}}\left\{0, \sigma^2, \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}\right\}$$

in which the residuals are independently, identically distributed random variables and are also independent of the regressors. An important implication of this assumption is that the conditional mean of $Y_i|\mathbf{x}_i$ will be equal to $\mathbf{x}_i^T\boldsymbol{\beta}$ plus the average value of the error distribution. To parallel the conventional analysis, the error distribution can be forced to take mean zero by suitably adjusting β_1 which is the intercept parameter of the regression model.

Since the observations are assumed to be independent given $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, the likelihood function of the model parameters is obtained as the product of the individual density of the observable y_i yielding

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} h\left\{y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}\right\},$$

where $h(\cdot)$ is stated in (3.9). Henceforth, in this section we condition on $\mathbf{x}_i$ without explicit mention. A prior distribution for the model parameters is needed to complete the specification of a Bayesian model. In our investigation, components of both $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are assigned independent normal prior distributions with large variances, and an inverse gamma distribution $IG(\nu, \nu)$ with a small positive choice of ν is used as prior for σ^2. More formally, the adopted joint prior distribution is given by

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) &= p(\boldsymbol{\beta}) \times p(\sigma^2) \times p(\boldsymbol{\delta}) \\ &= N_k(\tilde{\boldsymbol{\beta}}, \Omega) \times IG(\nu, \nu) \times N_2(\mathbf{0}, \Psi) \end{aligned}$$

where Ω and Ψ are diagonal matrices with diagonal elements 10^{10}, $\tilde{\boldsymbol{\beta}} = (\bar{y}, 0, \cdots, 0)$ ($\bar{y}$ is the sample mean), and $\nu = 0.001$. Thus $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta})$ is chosen in such a way that it has little impact on the posterior analysis.

According to the Bayes theorem (2.1), the joint posterior distribution of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\delta}$ is simply proportional to the likelihood function times the joint prior distribution

$$p(\boldsymbol{\beta}, \sigma^2, \delta_1, \delta_2|\mathbf{y}) \propto \prod_{i=1}^{n} h\left\{y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}\right\} \times p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}). \qquad (4.2)$$

Due to the complexity of the likelihood function, it is not possible to evaluate the marginal posterior distributions of the model parameters by analytical means. Hence, we resort to a numerical scheme known as Gibbs sampling (discussed in Section 2.7) to circumvent the calculational impediments. Our advocacy of the method rests essentially on its ease of implementation, yielding output from which functions of interest can be readily computed and inferences made. The necessary conditional distributions for use of the Gibbs sampler are obtained in the next section.

4.3 Numerical implementation

Because the skewing function of the sampling distribution (3.9) does not possess a closed form expression, it is not possible to obtain the required conditional distributions directly from (4.2). A better way to proceed is the following. Notice that the derivation of the sampling distribution allows us to alternatively express the skew normal linear regression model in (4.1) as

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\delta}^T \mathbf{Z}_i + \epsilon_i, \qquad Z_i > 0$$

with

$$\mathbf{Z}_i = \begin{pmatrix} Z_{1i} \\ Z_{2i} \end{pmatrix} \sim N_2(\mathbf{0}, \mathbf{I}) \qquad \text{and} \qquad \epsilon \sim N(\mu, \sigma^2),$$

where $\mathbf{Z}_i$ and ϵ_i are independent, and $\boldsymbol{\delta} = (\delta_1, \delta_2)^T$. Evidently, by treating the auxiliary variables $\mathbf{Z}_i$ as covariates, model (4.1) is seen to have an underlying normal linear regression model on the observations $\mathbf{y} = (y_1, \cdots, y_n)^T$. Thus casting the model in this form eliminates the need for the skewing function evaluations which in turn greatly facilitates the computation of the conditional distributions.

Now the complete Bayesian model of all the unknowns ($\mathbf{Z}$, $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\delta}$)

can be written hierarchically as

$$
\begin{aligned}
&Y_i|\mathbf{z}_i, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim N(\mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\delta}^T\mathbf{z}_i, \sigma^2),\\
&\mathbf{Z}_i \sim N_2(\mathbf{0}, \mathbf{I})I(\mathbf{z}_i > 0),\\
&\boldsymbol{\beta} \sim N_k(\tilde{\boldsymbol{\beta}}, \Omega),\\
&\tau = \frac{1}{\sigma^2} \sim \Gamma(\nu, \nu),\\
&\boldsymbol{\delta} \sim N_2(\mathbf{0}, \Psi).
\end{aligned}
$$

The corresponding expression for the complete joint posterior density is then

$$p(\mathbf{z}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta})p(\mathbf{z})p(\boldsymbol{\beta})p(\sigma^2)p(\boldsymbol{\delta}).$$

All relevant conditional distributions can be easily obtained by regarding other observables and unobservables as constant. Specifically, the conditional distributions of the regression parameters take the form

$$\beta_j|\{\beta_s : s \neq j\}, \sigma^2, \boldsymbol{\delta}, \mathbf{z}, \mathbf{y} \sim N\left[\frac{\Omega_{jj}\sum_{i=1}^n\{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} + x_{ij}\beta_j - \boldsymbol{\delta}^T\mathbf{z}_i)x_{ij}\} + \sigma^2\tilde{\beta}_j}{\Omega_{jj}\sum_{i=1}^n x_{ij}^2 + \sigma^2}, \frac{\sigma^2\Omega_{jj}}{\Omega_{jj}\sum_{i=1}^n x_{ij}^2 + \sigma^2}\right], \qquad j = 1, \cdots, k.$$

For the scale and skewness parameters we obtain

$$\tau = \frac{1}{\sigma^2}|\boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{z}, \mathbf{y} \sim \Gamma\left\{\frac{n}{2} + \nu, \frac{1}{2}\sum_{i=1}^n(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\delta}^T\mathbf{z}_i)^2 + \nu\right\}$$

and

$$\delta_j|\boldsymbol{\beta}, \sigma^2, \{\delta_s : s \neq j\}, \mathbf{z}, \mathbf{y} \sim N\left[\frac{\Psi_{jj}\sum_{i=1}^n\{y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\delta}^T\mathbf{z}_i + \delta_j(\mathbf{z}_j)_i\}(\mathbf{z}_j)_i}{\Psi_{jj}\sum_{i=1}^n(\mathbf{z}_j)_i^2 + \sigma^2}, \frac{\sigma^2\Psi_{jj}}{\Psi_{jj}\sum_{i=1}^n(\mathbf{z}_j)_i^2 + \sigma^2}\right], \qquad j = 1, 2.$$

Finally, $\mathbf{Z}_i$'s have full conditional distributions defined by

$$\mathbf{Z}_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, y_i \sim N_2\left\{\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\delta_1^2 + \delta_2^2 + \sigma^2}\begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}, \frac{1}{\delta_1^2 + \delta_2^2 + \sigma^2}\begin{pmatrix}\delta_2^2 + \sigma^2 & -\delta_1\delta_2\\ -\delta_1\delta_2 & \delta_1^2 + \sigma^2\end{pmatrix}\right\}I(\mathbf{Z}_i > 0).$$

These distributions are all of standard functional forms in which sample generation is relatively straightforward. So, Gibbs sampling can be easily implemented.

Before proceeding with the examples, there is still a technical issue needing to be addressed. The roles played by the two skewness parameters in SN_{new} are exchangeable, thus allowing their Gibbs realizations to travel from one's target distribution to the other. This sort of behavior will cause severe identifiability problems in determining the true marginal distributions of δ_1 and δ_2. Bimodality will, not surprisingly, be a typical characteristic in the estimated posterior densities. As a consequence, point and interval estimations based upon such marginal distributions can be rather misleading about the actual distributional structures. To resolve this we adopt the following simple strategy in all examples subsequently analyzed. When the estimated marginal posterior distributions are bimodal with relatively negligible probability mass in between the modes we recommend using the following steps:

1. Rearrange the MCMC simulated values by setting

   - $\delta_1^{(j)} = \max(\delta_1^{(j)}, \delta_2^{(j)})$ and
   - $\delta_2^{(j)} = \min(\delta_1^{(j)}, \delta_2^{(j)})$,

   where $j = 1, \cdots, t$ represents jth cycle of the Gibbs sampler.

2. Carry out all marginal calculations using the resulting samples.

This re-estimation proposal should be exercised prudently as it can have an adverse effect in other situations. For example, it is quite plausible that we might have skewness parameters with identical true value. Two unimodal marginal distributions (estimated using the original Gibbs outputs) will then be encountered. These should themselves provide a good approximation to the underlying distributions, thus there is no benefit in using the proposed scheme. In the case where there are two distinct but close modes in the original marginal distributions, the above approach will inevitably lead to densities with obvious truncations. So, neither the original nor the re-estimated marginal distributions are representative of the actual distributions. Use of a joint marginal to

make probability statements will be more sensible in this situation. However, visual inspection of the original marginal distributions is recommended before choosing an appropriate estimation method.

4.4 Examples

To illustrate the proposed methodologies, the Bayesian model described in Section 4.2 is used for the analysis of some real data sets. It is of interest to judge the appropriateness of new sampling distribution $\mathrm{SN}_{\mathrm{new}}$ in modeling non-Gaussian data and compare it with other existing skew-normal distributions. A total of four sampling models have been fitted to each of the following examples. Briefly they are $\mathrm{SN}_{\mathrm{new}}$ in (3.9), $\mathrm{SN}_{\mathrm{sdb}}$ in (3.8), $\mathrm{SN}_{\mathrm{tpn}}$ in (3.10) and the usual normal model. Legitimate comparison is elicited by allocating $\boldsymbol{\beta}$ and σ^2 in the latter models the same prior distributions as those specified in Section 4.2. The remaining parameter δ is given the $N(0, 10^{10})$ prior under the $\mathrm{SN}_{\mathrm{sdb}}$ model whilst a-priori $\delta \sim N(0, 10^{10})I(\delta > 0)$ is specified for the $\mathrm{SN}_{\mathrm{tpn}}$ model. The Gibbs sampler outlined earlier has been executed by using the `WinBUGS` software. Inferences are based on $200,000$ sequential version of Gibbs realizations, following a burn-in period of $10,000$ iterations to mitigate the impact of starting points.

4.4.1 Example 1: Non-academic scores

Data

Skewness of a sample distribution is often a consequence of screening operations in which experimental units are included in the study only if they have achieved certain pre-specified requirements. The particular data set that we consider here concerns admission to a Welsh medical institution in 1996. Non-academic scores for home applicants meeting the school academic criteria were recorded after reading the corresponding Universities & Colleges Admissions Service (UCAS) application forms. (UCAS is the central organization that processes applications for full-time undergraduate courses at Great Britain universities and colleges.) Candidates who failed the non-academic standards would be screened in the next level of selection process. The objective of the investigation is to determine

the group of students that is least likely to be successful in this stage of their application. Accordingly, our response variable Y is the non-academic scores of $n = 777$ individuals, and the covariates of interest include: number of GCSE A grades x_2, race (white or non-white) x_3, age in years x_4, and predicted/achieved A Level examination points x_5. Interaction between number of GCSE A grades and age $x_6 = x_2x_4$ is also embraced in the analysis, since it is a significant predictor in classical normal regression. Note that GCSE, stands for General Certificate of Secondary Education, is a national school examination in Britain. A summary illustration of the data is provided by Figure 4.1. The diagram demonstrates significant skewness in the non-academic distributions for both groups of students. As a result, a normal model may not adequately fit the data in this occasion.

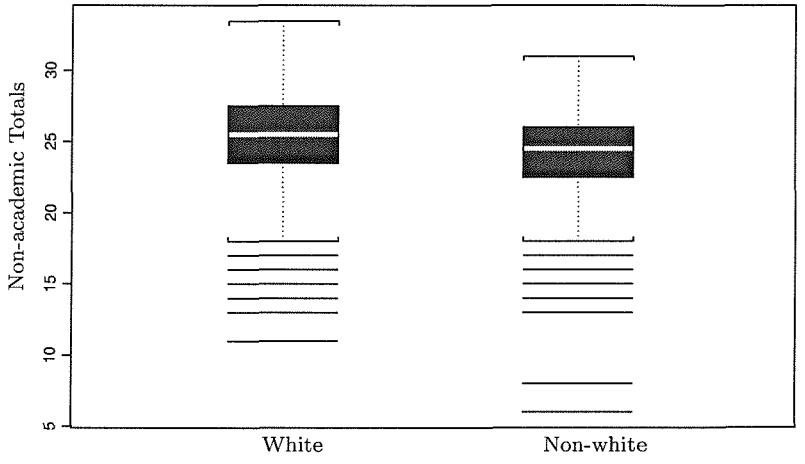

Figure 4.1: Boxplots of non-academic scores for different races.

Convergence diagnostics

The Gibbs sampling scheme is used to fit the skew normal linear regression model (4.1) to the non-academic scores data. We standardize all continuous covariates to help stabilize the posterior computations. It is recalled that a

complete implementation of the algorithm still requires the assessment of convergence. In this regard, we consider the following simple graphical technique that, though naive and less rigorously defined than might be desired, has proved successful in a considerable number of applications.

1. Run several (two to five) pilot independent chains with over-dispersed starting points and different random-number seeds.

2. Visually monitor these simulated sequences by overlaying their traces on the same graph for each parameter (or a representative subset of parameters).

3. Convergence is assumed when the simulated values mix together and settle around common values. Increase the number of drawings if this is not the case.

In the current example, the above method provides indication of satisfactory convergence after around 9000 cycles, which is less than the number of iterations intended to be discarded. This slow convergence behavior is caused by the lackadaisical movement of the simulation algorithm. In other words, there are strong positive autocorrelations between the successive iterations. Similar impression can be obtained from the after convergence samples shown in Figure 4.2. Accordingly, our single Gibbs cycle needs to be run a large number of times so as to appropriately cover the entire parameter space. Although retaining sample chain values at every kth cycle can reduce the first order dependence of the Markovian sequence, we have not found it useful as no efficiency is gained in posterior estimations.

Results

Table 4.1 reports the parameter estimates for all four models under consideration. Inspection of the table indicates little alteration in the estimates of β_3 and β_5, but inferences on β_2, β_4 and β_6 are noticeably affected by allowing for skewness. A closer examination reveals that the posterior mean of the latter parameters are very close to zero under the skew normal models. In other words,

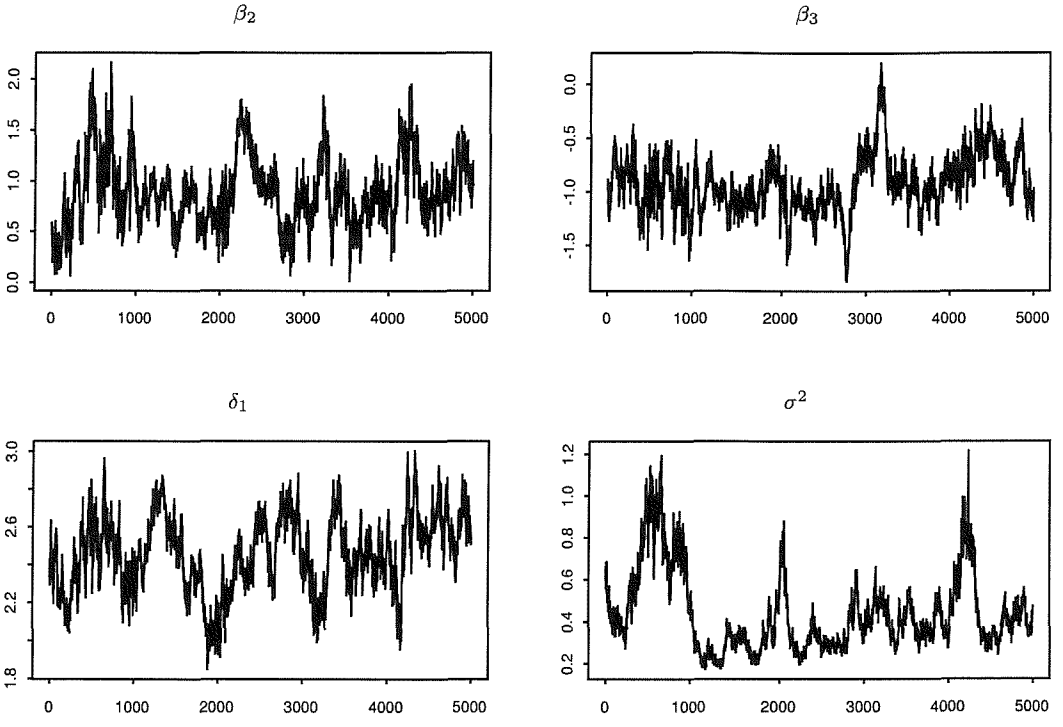

Figure 4.2: Times series plots for some model parameters in the non-academic scores example.

	β_1	β_2	β_3	β_4	β_5	β_6	σ^2	δ or δ_1	δ_2
Normal	25.1	1.51	-0.91	0.40	0.05	-0.05	9.36	–	–
	(0.123)	(0.377)	(0.278)	(0.106)	(0.017)	(0.020)	(0.478)		
SN_{sdb}	28.2	1.07	-0.87	0.29	0.05	-0.03	3.71	-3.90	–
	(0.217)	(0.387)	(0.260)	(0.111)	(0.016)	(0.020)	(0.550)	(0.253)	
SN_{tpn}	26.4	1.04	-0.94	0.29	0.05	-0.03	8.21	0.75	–
	(0.260)	(0.388)	(0.261)	(0.111)	(0.016)	(0.021)	(0.487)	(0.039)	
SN_{new}	26.92	0.93	-0.89	0.27	0.05	-0.03	1.16	-4.20	1.88
	(0.793)	(0.403)	(0.253)	(0.110)	(0.017)	(0.021)	(1.184)	(0.327)	(1.140)

Table 4.1: Parameter estimates and the associated standard deviations (given in parentheses) for the non-academic scores example.

the covariate effects are attenuated by assuming skewed models. Further insight into the behavior of these parameters is obtained through graphical representations of their marginal densities in Figure 4.3. As shown in the diagrams, the marginal posterior distributions based on SN_{sdb} and SN_{tpn} modeling are remarkably cohesive. Interestingly, there seems to be evidence of association between sampling model flexibility and marginal locality (simpler sampling model possesses marginal distributions that are farther from zero). One concern in the plots is the effect of the interaction variable on the analysis. The skewed sampling assumptions induce β_6 to have substantial posterior mass around zero, thus lessening the interaction variable momentousness in predicting the non-academic scores. Therefore, according to the skewed normal models, there is an improvement in regression additivity in the sense that the main covariates are emphasized relative to the interaction.

On the basis of the 95% credible intervals, it appears that modeling using different versions of error distribution can have considerable influences on the posterior of β_1. This is not surprising because β_1 is not the true regression intercept in the skew normal cases. From the theoretical results in Chapter 3,

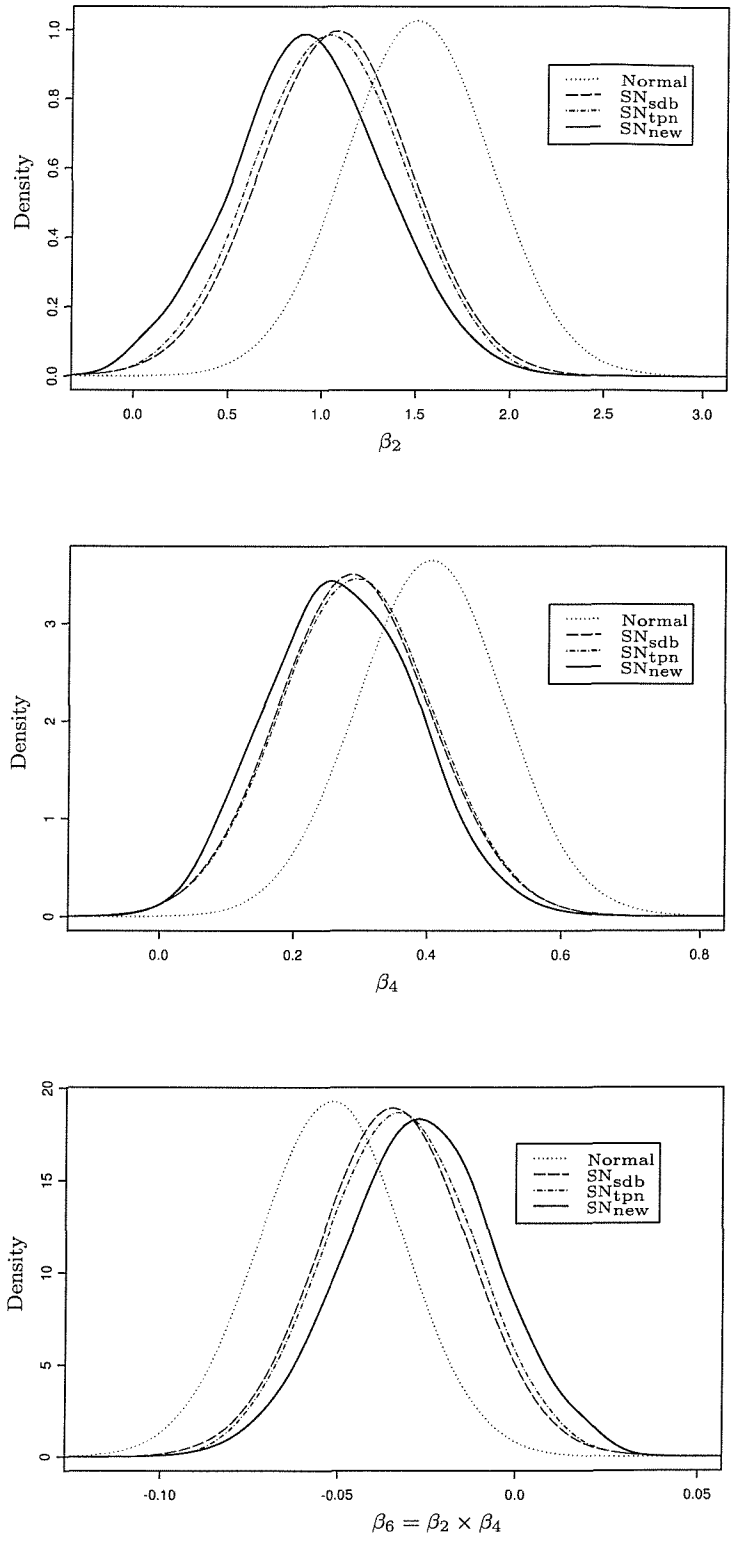

Figure 4.3: Marginal posterior densities of β_2, β_4 and β_6 for the non-academic scores example.

one can express the actual intercept parameter α as

$$\begin{aligned} \mathrm{SN}_{\mathrm{sdb}}: &\qquad \alpha = \beta_1 + \delta\sqrt{\frac{2}{\pi}}, \\ \mathrm{SN}_{\mathrm{tpn}}: &\qquad \alpha = \beta_1 + \sigma(\delta - \frac{1}{\delta})\sqrt{\frac{2}{\pi}}, \text{ and} \\ \mathrm{SN}_{\mathrm{new}}: &\qquad \alpha = \beta_1 + (\delta_1 + \delta_2)\sqrt{\frac{2}{\pi}}. \end{aligned}$$

Thus a meaningful location comparison should be obtained via α instead of β_1. Markov chain simulations of the intercept are readily computed from the existing Gibbs output. The resulting α estimates for $\mathrm{SN}_{\mathrm{sdb}}$, $\mathrm{SN}_{\mathrm{tpn}}$ and $\mathrm{SN}_{\mathrm{new}}$, together with their estimated standard deviations in parentheses, are given by 25.1 (0.122), 25.0 (0.126), and 25.1 (0.122) respectively. As expected, these values are in good agreement and are consistent with the findings in the normal model.

Consider now the inferences on the shape parameters: σ^2 and $\boldsymbol{\delta}$. Table 4.1 shows notably different estimates of σ^2 under the normal model as compared to the skewed models. This is justifiable since the parameter has dissimilar interpretations for all these models. Variability of the data is represented solely by σ^2 in the normal case, but non-zero skewness parameter(s) also share part of the variability in the skew normal cases. This distinction has made the estimated parameter values not comparable. Posterior estimates of δ for both $\mathrm{SN}_{\mathrm{tpn}}$ and $\mathrm{SN}_{\mathrm{sdb}}$ reiterate the finding evident from Figure 4.1, i.e. moderate right skewness is present in the data. Statistical significance of the parameter under the two models reinforces the fact that normal family would be unsuitable for modeling the original non-academic scores. The reported δ_1 and δ_2 estimates (successfully obtained using the approach described in Section 4.3) are also significant and lead to similar conclusion.

Model comparisons

To assess model adequacy, Figure 4.4 displays the data histogram with superimposed posterior predictive densities under each of the four models. All skewed models seem to provide an adequate fit to the non-academic scores, with the predictive from $\mathrm{SN}_{\mathrm{new}}$ most closely resembles the histogram. Observe that the

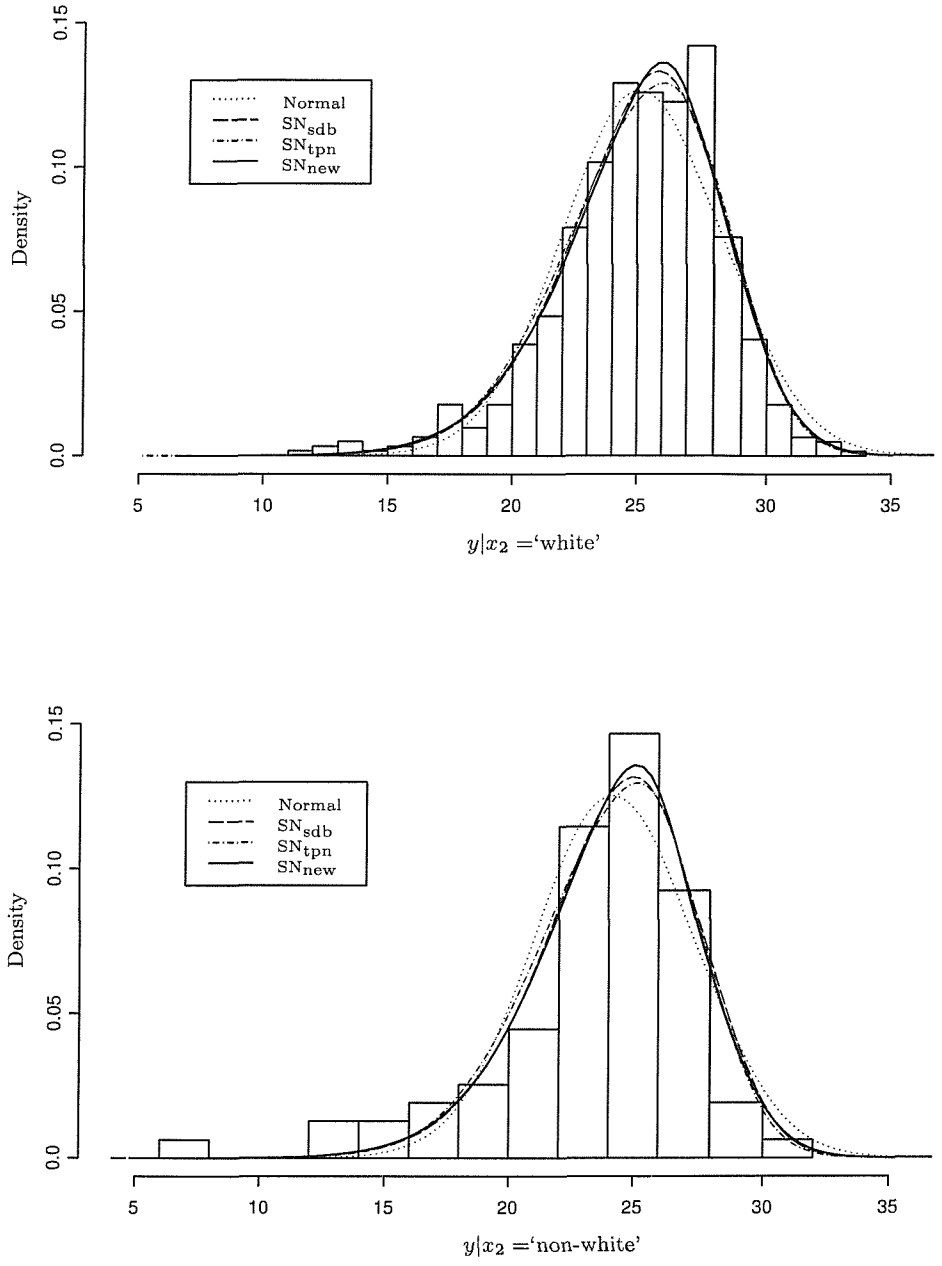

Figure 4.4: Histograms of the non-academic scores for both the races with superimposed posterior predictive densities under the SN_{new}, SN_{sdb}, SN_{tpn} and normal models.

	SN_{new}	SN_{sdb}	SN_{tpn}	Normal
SN_{new}	1	4.76	5.92E3	4.87E8
SN_{sdb}	–	1	1.24E3	1.02E8
SN_{tpn}	–	–	1	8.23E4
Normal	–	–	–	1

Table 4.2: Bayes factors based on the Laplace-bridge method for non-academic scores data. Entry (i, j) indicates the evidence in favor of model i versus model j. (Note: $\Re\text{E}\aleph \equiv \Re \times 10^{\aleph}$.)

predictive distribution under the normal model needs to be shifted to the left in order to account for the skewness in the data. This has an adverse effect on the model ability in capturing the peak of the histogram. A formal model comparison can be conducted through the use of Bayes factors. We compute the criterion by exercising the methods advocated by Meng and Wong (1996) and DiCiccio, Kass, Raftery and Wasserman (1997). Table 4.2 lists the resulting Bayes factors. Jeffreys' scale of evidence in Table 2.1 is employed to facilitate the interpretation of the criterion values relative magnitudes. The upshots indicates a dramatic improvement in the skew normal fits over the normal fit. In addition, SN_{new} is substantially better than SN_{sdb}, which is in turn definitely preferable to SN_{tpn}. Hence, the Bayes factor approach selects SN_{new} as the best model for the empirical data.

Conclusions

The analysis based on our best model SN_{new} suggests that non-academic scores are strongly related to number of GCSE A grades, race, age and number of predicted/achieved A Level points. Individual scores improve with GCSE results at approximately 0.93 credit for each A grade. Good non-academic outcomes are more prevalent among white students, who have 0.89 higher scores than non-white candidates generally. Age of the applicants also have a positive impact on the non-academic totals. The relative increment is about 0.27 unit per age year. As for the GSCE results, number of predicted/achieved A level points

is positively related to the non-academic outcome. However, the magnitude of influence is much smaller, approaching 0.05 score for each A level point gained. The analysis indicates no evidence of association between the response and the interaction effect, contradicting the upshot under the normal regression model. Putting these results together, we conclude that young non-white students with unfavorable GCSE and A level outcomes are those most probable to achieve inferior non-academic scores.

4.4.2 Example 2: Martin Marietta data

Data

Our second example is based on the data set reported in Table A1 of Butler *et al.* (1990). It contains records of monthly excess rate of return on equity for the Martin Marietta company, y, as well as the excess rate of return for the New York market as a whole, x, in the years 1982–1986. The primary objective of the study is to characterize the dependence of y on x. An illustration of the data is provided in Figure 4.5, showing quite clearly that there is a relationship between the two variables. However, the plot also appears to indicate presence of distributional peculiarity, in which all prominent outlying points tend to recline above the main body of the data. As such, proceeding the analysis by using a normal linear regression model may prove to be inappropriate. This impression has motivated a comparison of the adequacy of normal model to the skew-normal models in the subsequent investigation.

The simple linear model $Y_i = \alpha + \beta x_i + \varepsilon_i$, with ε_i follows the $\mathrm{SN}_{\mathrm{new}}$ distribution, is used to fit the Martin Marietta data. In order to conduct a comprehensive Bayesian analysis, we have employed the Gibbs sampling algorithm for obtaining a simulated sample of size $200,000$ from the corresponding joint posterior distribution. (The first $10,000$ iterations have been discarded as a burn-in phase.) But much smaller MCMC drawings already lead to reliable results. Convergence of the Gibbs sampler has been assessed via methods mentioned in the previous example. There was no sign of any instability in the marginal density estimates, providing an indication for convergence.

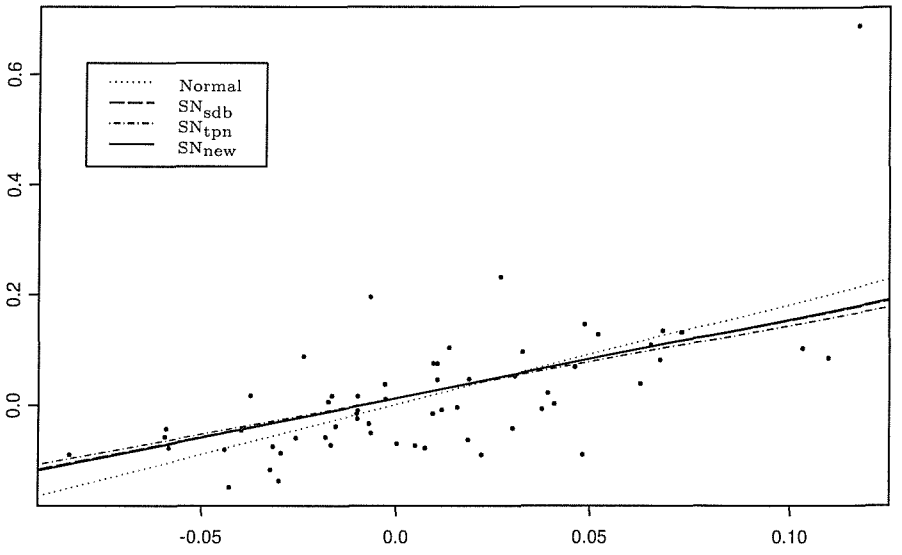

Figure 4.5: Scatter plot and fitted lines for the Martin Marietta data.

Results

Table 4.3 contains parameter estimates for the skew normal $\mathrm{SN}_{\mathrm{new}}$ linear regression model. For comparisons, posterior results using the standard normal model and the other skew normal models are also included in the same table. A check on the estimated kernel density of δ_1 has demonstrated evidence for bimodality with substantial probability mass rested between the modes. Thus, neither the original Gibbs estimates nor the proposed re-estimation scheme estimates are useful in summarizing the underlying distributional structure of δ_1 and δ_2 in this case, as remarked in Section 4.3. Rather than reporting their individual moments, we feel that a scatter plot of the simulated samples would provide a better picture about the distribution of these parameters. This is displayed in Figure 4.6.

Apart from revealing the strong correlation between the parameters, Figure 4.6 clearly illustrates an appreciable frequency of posterior drawings lying around the origin. This latter observation suggests that at least one of the

Model	Parameter	Posterior Mean	Posterior Standard Deviation	95% credible interval
Normal	α	0.002	0.013	(-0.023,0.027)
	β	1.803	0.295	(1.224,2.382)
	σ^2	0.0092	0.0018	(0.0064,0.0133)
SN_{sdb}	α	-0.093	0.014	(-0.120,-0.066)
	β	1.396	0.264	(0.868,1.910)
	σ^2	0.0017	0.0010	(0.0006,0.0043)
	δ	0.133	0.019	(0.097,0.171)
SN_{tpn}	α	-0.070	0.024	(-0.121,-0.027)
	β	1.303	0.307	(0.667,1.855)
	σ^2	0.0043	0.0016	(0.0008,0.0077)
	δ	2.320	1.181	(1.378,5.937)
SN_{new}	α	-0.095	0.035	(-0.158,-0.028)
	β	1.416	0.269	(0.887,1.947)
	σ^2	0.0015	0.0009	(0.0004,0.0039)
	δ_1	NA	–	–
	δ_2	NA	–	–

Table 4.3: Parameter estimates under each competing model for the Martin Marietta example.

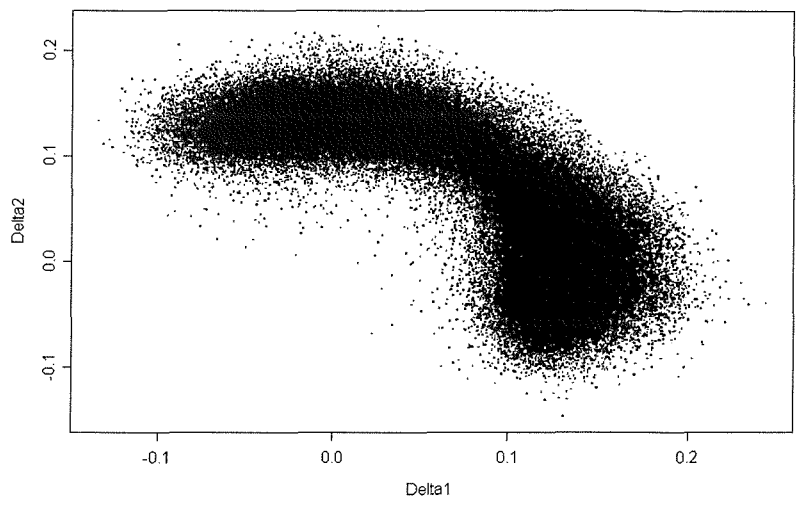

Figure 4.6: Simulated values of the skewness parameters (δ_1, δ_2) in the Matin Marietta data example.

skewness parameters is not needed to fit the empirical data. In other words, it is enough to consider a simpler model (normal or $\mathrm{SN}_{\mathrm{sdb}}$) as the underlying data generating mechanism. However, from Table 4.3, the skewness parameter δ under the $\mathrm{SN}_{\mathrm{sdb}}$ model is estimated to be positive with 95% credible interval excluding the point zero. This delivers a clear indication of significant positive skewness in the residual distribution. Accordingly, it would be inadequate to fit the data using a simple normal model. Similar conclusions can also be obtained through examining the posterior estimate of δ in the $\mathrm{SN}_{\mathrm{tpn}}$ case.

It is of interest to check if the form of the assumed error distribution affects the inferences on the regression coefficients. To this end, the estimated posterior densities of β and the true intercept α' under each of the four alternative models are displayed in Figure 4.7 for comparison. As seen in the diagram, $\mathrm{SN}_{\mathrm{sdb}}$ and $\mathrm{SN}_{\mathrm{new}}$ lead to practically indistinguishable posterior distributions for intercept and slope, giving a further demonstration that the extra skewness parameter is not worthwhile in an overall quality-of-fit sense. While $\mathrm{SN}_{\mathrm{tpn}}$ yields a similar finding for α' (with mean = 0.013 and standard deviation = 0.012 across the three model), the kernel density estimate of the slope parameter is slightly

attenuated relative to those obtained from the other skewed normal models. Nevertheless, the most noticeable feature in the figure is the isolation of the posterior densities under the normal model. That is, our inference about the regression coefficients change greatly after allowing for skewness.

To further quantify the difference between the normal and skewed specifications, regression lines for each of the four competing models are superimposed on Figure 4.5. As expected, all skewed versions provide fairly similar fitted lines to the Martin Marietta data, especially as compared with that of the normal family. A closer inspection of the configuration suggests that the fitted line from the normal model is substantially influenced by the possibly aberrant point on the top right corner of the graph. Intuitively, asymmetry in the residual can be largely attributed to this outlying observation. Hence, because the normal distribution makes no allowance for possible skewness, it needs to adjust its location in order to capture the observation in its upper tail. This also explains why the disturbance of the point seems alleviated under the skewed error assumptions. It is, however, not possible to select the best fitting candidate model on the basis of such plots alone.

Model comparisons

A natural next step after model fitting is to examine issues relating to model adequacy and model choice. In order to investigate the former, Figure 4.8 plots the posterior predictive densities for each candidate models, together with a histogram of the data. The observed data are shown as symbols on the horizontal axis. Notice that all skewed predictives appear to give a satisfactory overall representation to the underlying data. On the contrary, the normal model clearly induces a rightward shift and more dispersion in its predictive distribution, pointing to an inadequate description of the structures present in the data. Bayes factors are reported in Table 4.4 to facilitate model selection. It is interesting to observe that, despite the conclusions evident from Table 4.3, Figure 4.5 and Figure 4.8, this criterion actually expresses strong preference for the normal model.

Although Bayes factor is a pure Bayesian tool for general model comparison,

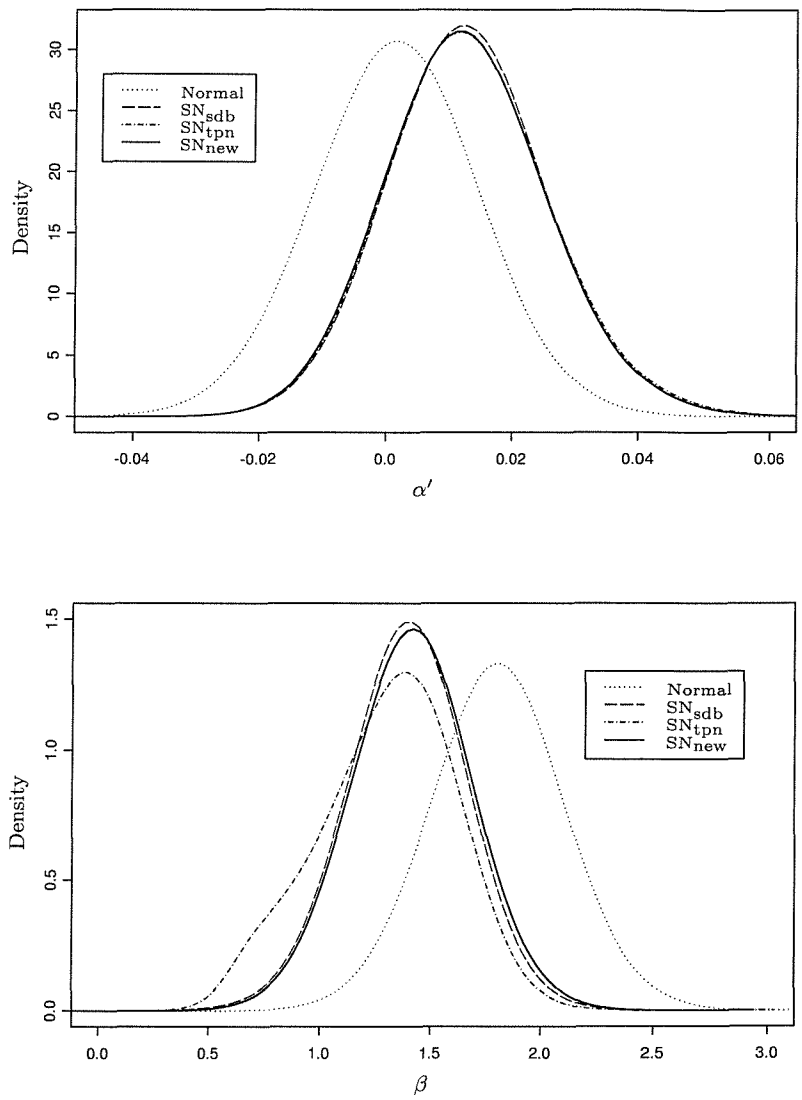

Figure 4.7: Kernel density estimates of the regression coefficients in the Matin Marietta data example.

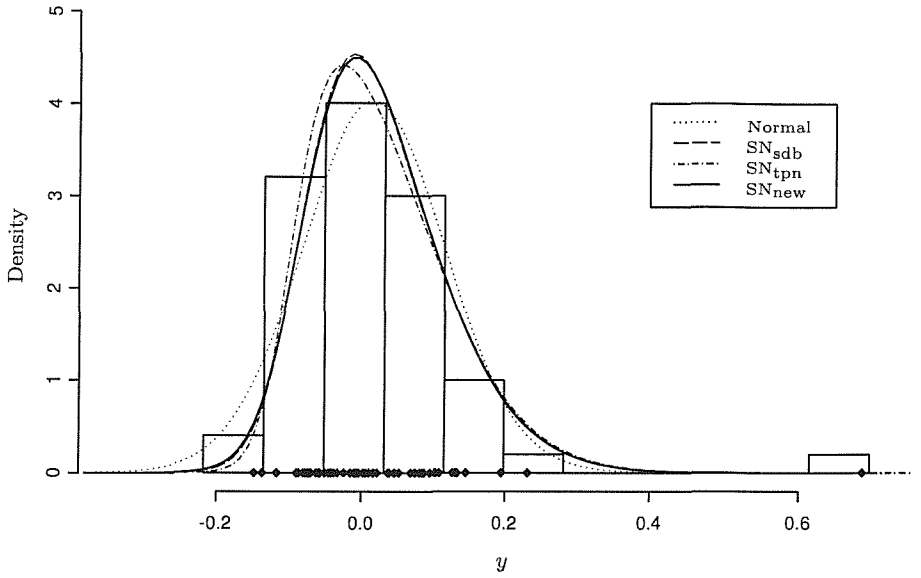

Figure 4.8: Posterior predictive distributions for the models considered in the Martin Marietta data example.

Bayes Factor				◁▷	Pseudo BF			
SN_{new}	SN_{sdb}	SN_{tpn}	Normal		SN_{new}	SN_{sdb}	SN_{tpn}	Normal
1	–	–	–	SN_{new}	1	–	6.90	3.63E3
1.57E6	1	–	–	SN_{sdb}	3.57	1	24.6	1.29E4
2.39E6	1.52	1	–	SN_{tpn}	–	–	1	526
4.02E8	256	168	1	Normal	–	–	–	1

Table 4.4: Model choice for the Martin Marietta data.

it could be misleading in the present example. From Table 4.4, the Bayes factor in favor of SN_{sdb} to SN_{new} is calculated as 1.57×10^6. Considering the similarity between these models in terms of overall fit, it seems that the criterion has allocated too much penalty for model complexity. As such, the choice of a suitable model is not immediately clear without undertaking further statistical investigation. Here we propose to work with the cross validation approach using pseudo-Bayes factor (PsBF) and conditional predictive ordinate (CPO), in aiming to identify a model that best explains the observed data with reasonable trade off against over-fitting. For ease of comparison, the PsBF's are also presented in Table 4.4. Evidence for skewness is quite emphatically pronounced under this criterion. Following Jeffreys' scale of evidence, SN_{sdb} is the best fitted model, though the support for SN_{sdb} is only marginally greater than SN_{tpn}.

Figure 4.9 graphically displays the individual CPOs to help understand better the implications of PsBF's. Recall from Section 2.4 that larger value of CPO indicates more preference for a model from the corresponding observation. Thus it is apparent from the configuration that SN_{sdb} generally outperforms all other candidate models. In particular, there are 42 out of 60 observations which support SN_{sdb} over SN_{tpn}. This conveys strong evidence in favor of the former. Moreover, there seems to be significant improvement in moving from the normal to SN_{sdb}, but not in going from SN_{sdb} to SN_{new}. The CPOs resulting from using SN_{new} and SN_{sdb} are very much alike, confirming again the unworthiness of the additional skewness parameter. In conclusion, according to the cross validation methods, SN_{sdb} is the most appropriate model for the Martin Marietta data.

Conclusions

The New York market excess rate of return (ERR) is important in predicting the Martin Marietta company excess rate of return on equity. Referring to our best model SN_{sdb}, they are associated via the following formulation:

$$\text{Expected company ERR on equity} = 0.013 + 1.40 * \text{Whole market ERR}.$$

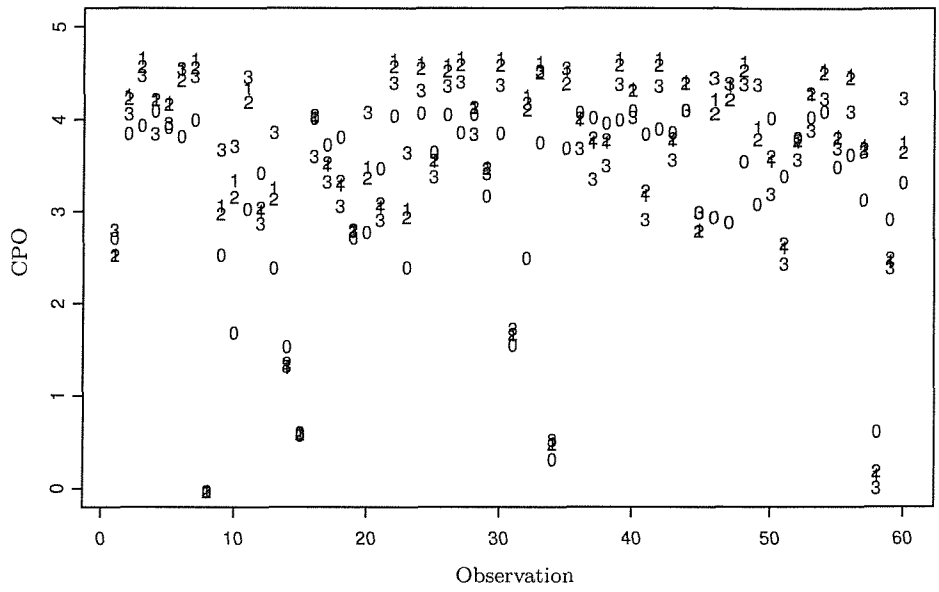

Figure 4.9: Plot of CPO versus observation number for the Martin Marietta data example: 0 =Normal, 1 = SN_{sdb}, 2 = SN_{new}, and 3 = SN_{tpn}.

Nevertheless, the data contains an outlier suspicious of being responsible for a major part of the residual skewness. More data would seem the best arbiter of whether this outlier corresponds to a maverick observation. If that is indeed the case, it might be preferable to have the observation removed from the analysis. This is likely to give rise to quite different statistical conclusions. However, in the absence of such information, we do not consider there is sufficient evidence to cast doubt on the appropriateness of the SN_{sdb} specification.

4.5 Closing remarks

Our efforts in this chapter have essentially been focused on linear regression modeling through the use of SN_{new}. Naturally in a regression problem, numerous distributions are possible candidates but SN_{new} is intriguing in a number of respects.

- The family shows promise with its flexibility to yield significantly bet-

ter fits than other existing skew normal distributions. It adapts itself to accommodate skewness or kurtosis whenever a data possesses these characteristics.

- Model fitting is performed on the original scale of the data. This may lead to a more meaningful interpretation of the quantities involved.

- Empirical analysis based on MCMC methods is quite attractive. In particular, Gibbs sampling can be easily implemented using publicly available software BUGS.

Therefore, $\mathrm{SN}_{\mathrm{new}}$ provides a viable alternative to the symmetric normal distribution often assumed in regression analysis.

Still it might be meaningful to consider re-expressing the skewed distribution in terms of explicit skewness and kurtosis parameters. Such approach would permit all model parameters to have clearer defined modeling purposes, i.e. each of them are now intuitively interpretable. The approach may also grant a radical solution to $\mathrm{SN}_{\mathrm{new}}$ identifiability problem in Gibbs sampling, but this requires thorough investigations. We should not pursue reparameterization further here since it is beyond the scope of the present research. We remark that $\mathrm{SN}_{\mathrm{new}}$ endures no fat tails per se, which makes the distribution unsuitable for prediction of extreme outcomes. Faced with this situation, data analysts are advised to turn to other members of the skew elliptical family (3.4). A useful choice in this context would be the skew-t distribution, see, for instance, Sahu *et al.* (2003).

Chapter 5

Applications to variance components models

5.1 Introduction

This chapter further illustrates the use of the skew-normal distribution in realistic modeling. The statistical problems of interest here are concerned with variance components. In particular, our main purpose is to develop skew modeling of random effects for the balanced one-way classification. All discussions are within the Bayesian framework, with the Gibbs sampling scheme (see Section 2.7 for details) being adopted to obtain marginal inferences about the general mean and the components of variance. Bayesian approach for estimating the variance components has several practical advantages over the sampling theory methods. Firstly, there is no such issue as a negatively estimated variance under the Bayesian paradigm. Secondly, posterior credible intervals are never empty or contain negative values. Thirdly, experimenters can report the whole posterior distribution and, with little computational effort, make some measure for posterior precision. These superiorities justify the consideration of the Bayesian procedure.

The remainder of the chapter is structured as follows. Section 5.2 introduces the variance components model. We model random effect using $\mathrm{SN}_{\mathrm{sdb}}$ distribution (3.8) but exercise the customary normality assumption on the error term. The prior distributions are discussed in Section 5.3. Full conditional den-

sities essential for implementing the Gibbs sampler are detailed in Section 5.4. Section 5.5 takes up illustrations based on simulated data sets generated using varying amount of information. The results are contrasted with those acquired under symmetric normal random effects. Finally, in Section 5.6, major findings of the chapter are summarized. The pertinent BUGS codes are placed in the Appendix.

5.2 Variance components models

Sets of observations are frequently obtained in clusters. For instance, the clustering may be due to subsampling of the primary sampling units, or to repeated measurements on a collection of individuals. Such data require special treatments for handling the correlations which typically present among the responses in the same group. The customary approach for analyzing clustered responses is to introduce as covariates a set of classifying variables into the traditional regression specification. More formally, in its simplest non-trivial setup, the linear model holds the form

$$Y_{ij} = \mu + g_i + \varepsilon_{ij}, \qquad i = 1, \cdots, I; j = 1, \cdots, J, \tag{5.1}$$

where Y_{ij} represents the jth observation in the ith group, μ is a common location parameter, g_i symbolizes the effect associated with the ith group, and ε_{ij} characterizes the deviation of the (i, j)th observation from the structure of the model $\mu + g_i$. Intrinsic to the idea is that, by conditioning on the indicator variable g_i, independence of ε_{ij}'s may now become a realistic assumption.

The term g_i can be treated as either an unknown constant or a random variable. The former arises when the particular clusters occurring in the experiment are of primary interest, i.e. our principal purpose of the analysis is to measure and compare their effects. In the alternative scenario, experimental concentration lies, rather, in drawing inferences about the whole population from which groups in the data are considered as a random sample. Nothing important has conditioned our choosing any one group over another, and there is no particular interest in specific comparisons between the selected clusters. It is with such random nature that this chapter is concerned. The classifying

variable g_i is commonly referred to as random effect in this context. Accordingly, the resultant model (5.1) is called a random effects model, or sometimes a variance-components model. This latter terminology emerges from the following relationship

$$Var(Y_{ij}) = Var(g_i) + Var(\varepsilon_{ij}).$$

Hence the respective dispersion measures of error terms and random effects are 'variance components' of the responses.

Like other parametric models, model (5.1) requires assumptions about the observations before proceeding with any statistical analysis. In the current case assumptions must be placed upon the distributions of both g_i and ε_{ij}. Most investigators specify a normally distributed random effect in their analysis. However, when robustness to asymmetry is a concern, use of a distribution that gives flexibility to cope with possible skewness seems more appropriate. In what follows, attention will be directed at modeling g_i using the skew normal distribution $\mathrm{SN}_{\mathrm{sdb}}$. We keep the conventional assumption of normality for the residual ε_{ij}. It is also necessary to assume that both clusters and observations within clusters are randomly sampled. To summarize:

$$\begin{aligned} g_i &\sim \mathrm{SN}_{\mathrm{sdb}}(0, \delta, \sigma_g^2), \qquad \text{and} \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2), \end{aligned}$$

for which g_i and ε_{ij} are mutually independent.

The likelihood function for the above model is easily written down, viz

$$\begin{aligned} L(\mu, \sigma_g^2, \sigma_\varepsilon^2, \delta|\mathbf{y}) = \int \cdots \int \left\{ \prod_{i=1}^{I} \prod_{j=1}^{J} \phi(y_{ij}|\mu + g_i, \sigma_\varepsilon^2) \right\} \times \\ \left\{ \prod_{i=1}^{I} h(g_i|0, \sigma_g^2, \delta) \right\} dg_1 \cdots dg_I, \end{aligned} \tag{5.2}$$

where $\phi(\cdot)$ denotes the normal density function and $h(\cdot)$ is the pdf of the skew normal distribution given in (3.8). A full Bayesian analysis demands the specification of suitable prior densities for all model parameters. Leaving our choice of prior initially as $p(\mu, \sigma_\varepsilon^2, \sigma_g^2, \delta)$, the joint posterior distribution becomes

$$p(\mu, \sigma_\varepsilon^2, \sigma_g^2, \delta|\mathbf{y}) \propto L(\mu, \sigma_g^2, \sigma_\varepsilon^2, \delta|\mathbf{y}) p(\mu, \sigma_\varepsilon^2, \sigma_g^2, \delta).$$

Interest in this type of study often centers at variance components and overall mean. Here, these are defined by

$$\text{Within cluster variation } :Var(\varepsilon_{ij}) = \sigma_\varepsilon^2$$
$$\text{Between cluster variation } :Var(g_i) = \sigma_g^2 + \delta^2(1 - 2/\pi)$$
$$\text{General mean } :\alpha = \mu + \delta\sqrt{\frac{2}{\pi}}.$$

We propose to use the Gibbs sampler for obtaining the desired marginal posterior distributions.

5.3 Prior distributions

Consider the following hierarchical interpretation of the likelihood (5.2).

$$Y_{ij}|g_i, \mu, \sigma_\varepsilon^2 \sim N(\mu + g_i, \sigma_\varepsilon^2)$$
$$g_i|z_i, \sigma_g^2, \delta \sim N(\delta z_i, \sigma_g^2)$$
$$Z_i \sim N(0,1)I(Z_i > 0).$$

Introducing z_i into the variance component model eliminates the need for evaluating the skewing functions of SN_{sdb}. As a consequence, conditional densities used in the Gibbs sampler can now be derived straightforwardly. More precisely, they are all proportional to the complete joint posterior given below.

$$p(\mathbf{z}, \mathbf{g}, \mu, \sigma_\varepsilon^2, \sigma_g^2, \delta|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{g}, \mu, \sigma_\varepsilon^2)p(\mathbf{g}|\mathbf{z}, \sigma_g^2, \delta)p(\mathbf{z})p(\mu, \sigma_\varepsilon^2, \sigma_g^2, \delta).$$

The choice of individual prior distributions is now discussed. We assume a priori that all model parameters are independently distributed, i.e.

$$p(\mu, \sigma_\varepsilon^2, \sigma_g^2, \delta) = p(\mu)p(\sigma_\varepsilon^2)p(\sigma_g^2)p(\delta).$$

Although a great variety of continuous densities can be specified, we embrace a widely used conditional conjugate prior for the unknowns, taking

$$\mu \sim N(0, \omega)$$
$$\tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2} \sim \Gamma(\nu, \nu)$$
$$\tau_g = \frac{1}{\sigma_g^2} \sim \Gamma(\kappa, \kappa)$$
$$\delta \sim N(0, \psi).$$

In eliciting the hyperparameters, our objective is to keep the prior information as vague as possible so that posterior inference is not driven by the prior. The values of both ν and κ are set at 0.001, thus leading to a prior spread of 1000 for the inverse variances $(\tau_\varepsilon, \tau_g)$. The specification is completed by picking the hyperparameters ω and ψ to be 10^{10}.

5.4 Full conditional posterior distributions

Now it is easy to verify that in each iteration the Gibbs sampler draws samples from the following full conditional distributions:

$$\bullet\ \mu|\sigma_\varepsilon^2, \mathbf{g}, \mathbf{y} \sim N\left(\frac{\omega\sum_i\sum_j(y_{ij}-g_i)}{IJ\omega+\sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2\omega}{IJ\omega+\sigma_\varepsilon^2}\right),$$

$$\bullet\ \tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2}|\mu, \mathbf{g}, \mathbf{y} \sim \Gamma\left(\frac{I+J}{2}+\nu, \frac{1}{2}\sum_i\sum_j(y_{ij}-\mu-g_i)^2+\nu\right),$$

$$\bullet\ \tau_g = \frac{1}{\sigma_g^2}|\delta, \mathbf{z}, \mathbf{g} \sim \Gamma\left(\frac{I}{2}+\kappa, \frac{1}{2}\sum_i(g_i-\delta z_i)^2+\kappa\right),$$

$$\bullet\ \delta|\sigma_g^2, \mathbf{z} \sim N\left(\frac{\psi\sum_i z_i g_i}{\sum_i z_i^2\psi+\sigma_g^2}, \frac{\psi\sigma_g^2}{\sum_i z_i^2\psi+\sigma_g^2}\right),$$

$$\bullet\ Z_i|\sigma_g^2, \delta, g_i \sim N\left(\frac{\delta g_i}{\delta^2+\sigma_g^2}, \frac{\sigma_g^2}{\delta^2+\sigma_g^2}\right) I(Z_i>0),$$

$$\bullet\ g_i|\mu, \sigma_\varepsilon^2, \sigma_g^2, \delta, z_i, \mathbf{y} \sim N\left(\frac{\sigma_g^2\sum_j(y_{ij}-\mu)+\sigma_e^2\delta z_i}{J\sigma_g^2+\sigma_\varepsilon^2}, \frac{\sigma_g^2\sigma_\varepsilon^2}{J\sigma_g^2+\sigma_\varepsilon^2}\right).$$

While conditional conjugacy simplifies the Gibbs sampling implementation, it is not an essential element in the analysis. Other form of priors could as well be employed if so desired.

5.5 Simulation study

Data sets

In order to assess the potential of the skewed random effect model (5.1) in data analysis, we have conducted a simulation study. A total of 16 data sets

reflecting different amount of statistical information were generated. Number of groups was chosen to be $I = (10, 50, 100, 200)$, and observation numbers per cluster J took the values $(5, 10, 20, 50)$. Thus the smallest data set comprised 10 groups of 5 observations each; the largest one were made out of 10,000 total data points. Data were drawn randomly from model (5.1) using parametric values of $\mu = 0$, $\sigma_\varepsilon^2 = 1$, $\sigma_g^2 = 0.1$ and $\delta = 2$ respectively. This implies a highly skewed group effects g_i (index of skewness equals 0.9) with true between clusters variation $Var(g_i) = 1.55$ and overall mean $\alpha = 1.60$. Our principal focus is in making inferences about the variance components $[Var(\varepsilon_{ij}), Var(g_i)]$ as well as the general mean α. As such, no attention will be paid hereafter to the parameters μ, σ_g^2 and δ.

For each simulated data set, the Gibbs sampler was run in `WinBUGS` for 210,000 cycles. Allowing an initial transient phase of 10,000 realizations, posterior inferences are based upon the latter 200,000 Gibbs iterations. We monitor the iterative process by observing the raw trace plots of $[Var(\varepsilon_{ij}), Var(g_i)]$, and α in a univariate fashion. Typical sets of successive drawings are graphically documented in Figure 5.1. As seen in the diagram, sequence of chain values tend to concentrate around the same patterns. This qualitative behavior gives an indication of convergence for the quantities involved.

Results

Table 5.1 presents summary values of the numerical work. Posteriors of α are reasonably symmetric while those of $Var(\varepsilon_{ij})$ and $Var(g_i)$ can be very skewed with extremely long right tails. To see this more clearly, Figure 5.2 depicts boxplots of the variance component marginals. For such asymmetric distributions, a median appears to be a better single summary measure than a mean, since it is less sensitive to strong skewness. Hence, we provide values of the posterior median in the table instead of the posterior mean. It seems obvious from both Table 5.1 and Figure 5.2 that $Var(g_i)$ is harder to estimate as compared with $Var(e_{ij})$. Marginals of the former exhibits significantly greater posterior uncertainty for all simulated data sets. The variability is particularly noticeable for small number of clusters I. This may be expected since the information

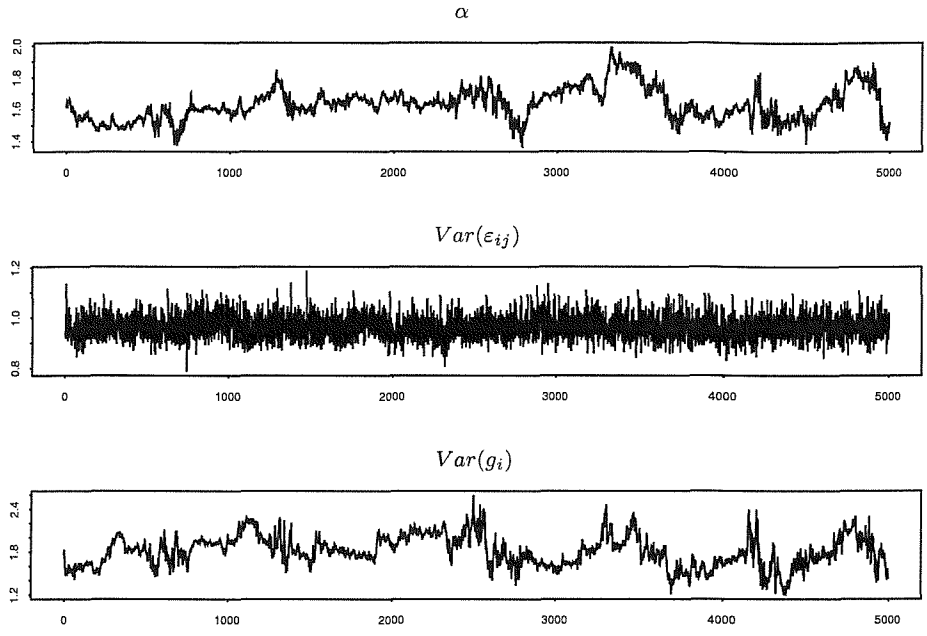

Figure 5.1: Post-convergence times series plots of general mean and variance components for the $(I = 50, J = 10)$ case.

on $Var(g_i)$ is largely determined by I rather than the total observations. Accordingly, the posterior distribution of $Var(g_i)$ benefitted far more from adding groups into the experiment than from increasing group sizes. On the contrary, one gains more accurate inference for $Var(e_{ij})$ (and α) as I or J becomes larger.

As mentioned earlier, normality is a popular assumption for the random effect g_i. It is now of practical interest to determine whether asymmetry causes systematic differences between the two specifications for g_i. For comparative purposes, the prior distributions for the parameters present in the normal random effect model are kept the same as those described in Section 5.3. Table 5.2 gives the parameter estimates for the normal model fitted to the simulated data sets. All estimates are again based on 200,000 post-convergence Gibbs replications. The results in Tables 5.1 and 5.2 indicate fairly similar findings from both models, especially when I is large. However, further examination shows that employing $\mathrm{SN}_{\mathrm{sdb}}$ random effect leads to greater estimated medians for α. Notice also the slight reductions of $Var(\varepsilon_{ij})$ estimates under the skewed ver-

		General Mean (α)		Error Term (σ_ε^2)		Random Effect (σ_g^2)	
		Posterior	95% Credible	Posterior	95% Credible	Posterior	95% Credible
I	J	Median (s.d.)	Interval	Median (s.d.)	Interval	Median (s.d.)	Interval
10	5	1.64 (0.62)	(0.66,3.16)	0.75 (0.18)	(0.50,1.21)	2.35 (2.50)	(0.83,9.30)
	10	1.46 (0.47)	(0.70,2.60)	1.14 (0.18)	(0.86,1.55)	1.22 (1.43)	(0.40,5.47)
	20	1.48 (0.49)	(0.66,2.67)	0.99 (0.103)	(0.81,1.22)	1.55 (1.62)	(0.58,6.17)
	50	1.69 (0.56)	(0.73,2.95)	0.95 (0.061)	(0.84,1.08)	2.14 (2.02)	(0.81,7.77)
50	5	2.00 (0.20)	(1.62,2.41)	0.94 (0.095)	(0.78,1.15)	1.98 (0.50)	(1.26,3.19)
	10	1.57 (0.19)	(1.25,1.99)	1.08 (0.072)	(0.95,1.24)	1.67 (0.43)	(1.07,2.74)
	20	1.76 (0.23)	(1.35,2.25)	0.99 (0.046)	(0.91,1.09)	2.70 (0.61)	(1.83,4.19)
	50	1.72 (0.20)	(1.38,2.16)	1.05 (0.030)	(0.99,1.11)	1.96 (0.44)	(1.30,2.97)
100	5	1.69 (0.15)	(1.40,1.99)	1.02 (0.073)	(0.89,1.18)	2.02 (0.37)	(1.43,2.85)
	10	1.61 (0.13)	(1.38,1.87)	0.96 (0.045)	(0.88,1.05)	1.77 (0.28)	(1.30,2.41)
	20	1.35 (0.13)	(1.11,1.62)	0.94 (0.031)	(0.88,1.00)	1.54 (0.27)	(1.12,2.14)
	50	1.84 (0.13)	(1.60,2.11)	1.00 (0.020)	(0.96,1.04)	1.96 (0.31)	(1.46,2.66)
200	5	1.68 (0.093)	(1.51,1.88)	0.99 (0.049)	(0.90,1.10)	1.62 (0.21)	(1.28,2.12)
	10	1.65 (0.089)	(1.49,1.83)	0.99 (0.033)	(0.93,1.05)	1.73 (0.20)	(1.39,2.15)
	20	1.59 (0.088)	(1.42,1.76)	1.00 (0.023)	(0.95,1.04)	1.63 (0.18)	(1.31,2.03)
	50	1.58 (0.085)	(1.42,1.76)	1.00 (0.014)	(0.97,1.03)	1.51 (0.17)	(1.22,1.88)

Table 5.1: Parameter estimates for the simulated examples under $\mathrm{SN}_{\mathrm{sdb}}$ random effect.

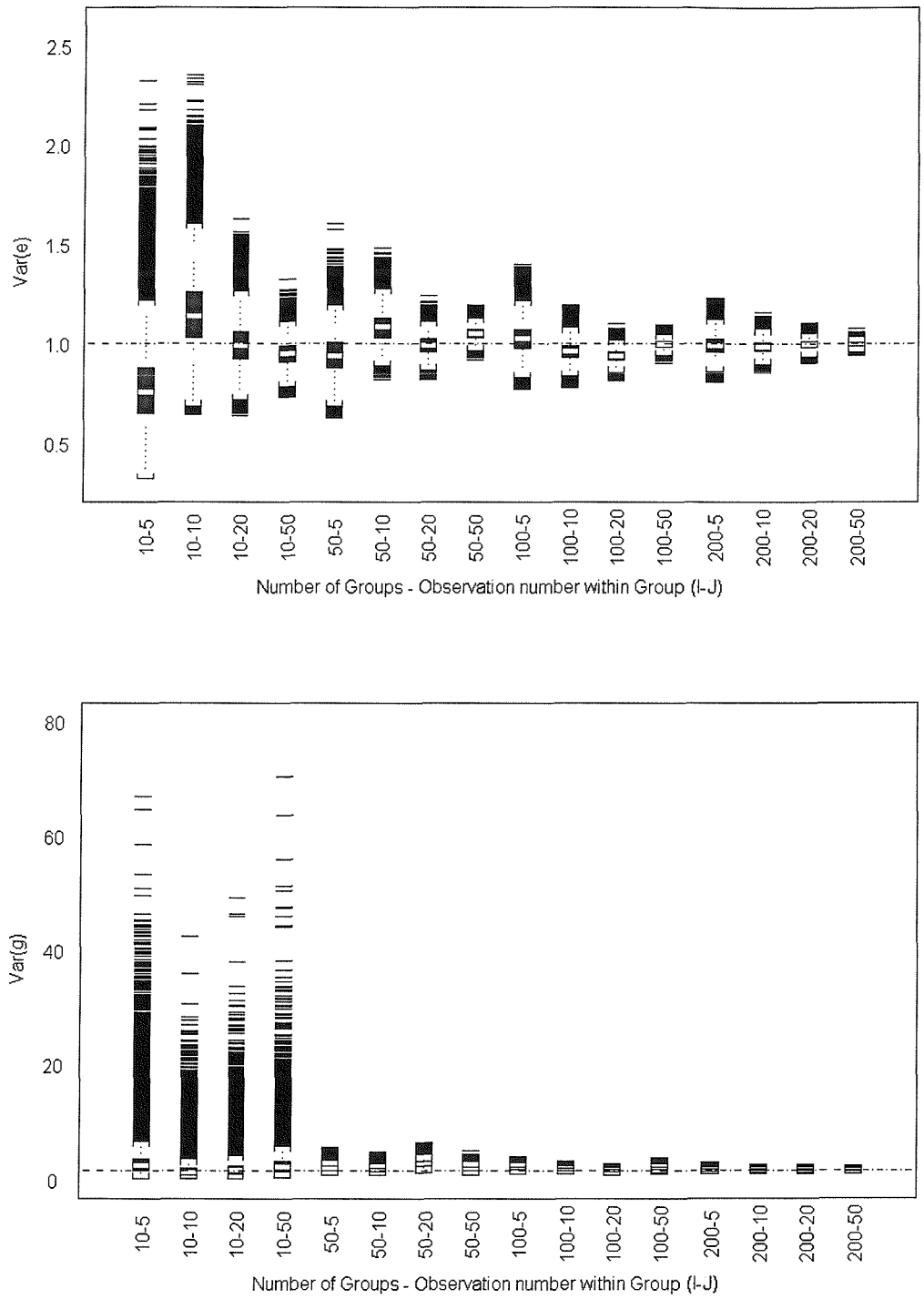

Figure 5.2: Boxplots for marginal posterior of variance components under various combinations of sample size, where random effect follows the SN_{sdb} distribution. The broken horizontal lines marked on each plot represents the respective population value of the variance components.

		General	Mean (α)	Error	Term (σ_e^2)	Random	Effect (σ_g^2)
		Posterior	95% Credible	Posterior	95% Credible	Posterior	95% Credible
I	J	Median (s.d.)	Interval	Median (s.d.)	Interval	Median (s.d.)	Interval
10	5	1.41 (0.54)	(0.33,2.48)	0.76 (0.19)	(0.51,1.23)	2.27 (1.84)	(0.90,7.33)
	10	1.34 (0.38)	(0.59,2.09)	1.15 (0.18)	(0.87,1.56)	1.07 (0.90)	(0.40,3.56)
	20	1.33 (0.43)	(0.47,2.19)	0.99 (0.104)	(0.82,1.22)	1.50 (1.20)	(0.63,4.77)
	50	1.56 (0.49)	(0.58,2.53)	0.95 (0.061)	(0.84,1.08)	2.00 (1.52)	(0.86,6.12)
50	5	1.91 (0.21)	(1.49,2.33)	0.94 (0.096)	(0.78,1.15)	2.03 (0.48)	(1.34,3.21)
	10	1.53 (0.20)	(1.14,1.91)	1.09 (0.073)	(0.95,1.24)	1.75 (0.40)	(1.17,2.74)
	20	1.59 (0.26)	(1.08,2.09)	0.99 (0.046)	(0.91,1.09)	3.13 (0.69)	(2.15,4.83)
	50	1.56 (0.21)	(1.14,1.98)	1.05 (0.030)	(0.99,1.11)	2.18 (0.48)	(1.50,3.36)
100	5	1.67 (0.15)	(1.38,1.97)	1.03 (0.073)	(0.90,1.19)	2.01 (0.33)	(1.49,2.76)
	10	1.58 (0.14)	(1.31,1.86)	0.96 (0.045)	(0.88,1.06)	1.85 (0.29)	(1.39,2.51)
	20	1.34 (0.12)	(1.10,1.59)	0.94 (0.030)	(0.88,1.00)	1.49 (0.23)	(1.13,2.01)
	50	1.84 (0.14)	(1.56,2.11)	1.00 (0.020)	(0.96,1.04)	1.92 (0.29)	(1.46,2.58)
200	5	1.68 (0.097)	(1.49,1.87)	1.00 (0.050)	(0.90,1.10)	1.64 (0.19)	(1.32,2.06)
	10	1.65 (0.095)	(1.46,1.84)	0.99 (0.033)	(0.93,1.06)	1.70 (0.18)	(1.38,2.10)
	20	1.58 (0.092)	(1.40,1.76)	1.00 (0.023)	(0.95,1.04)	1.63 (0.17)	(1.34,2.01)
	50	1.57 (0.089)	(1.39,1.74)	1.00 (0.014)	(0.97,1.03)	1.54 (0.16)	(1.27,1.89)

Table 5.2: Parameter estimates for the simulated examples under normal random effect.

sion. More formal comparisons between these models are disclosed in the next section.

Model comparisons

We utilize posterior predictive distributions to check the validity and adequacy of the fitted models. The results are summarized in Figures 5.3 – 5.6. For number of groups $I = 10$, predictive distributions from the two sets of specifications seem not to differ too much. This is perhaps not surprising because $I = 10$ is of rather small sample size for an adequate fit of a $\mathrm{SN}_{\mathrm{sdb}}$ random effect. Only when number of groups is sufficiently large do discrepancies between the models

become apparent. In those cases ($I \geq 50$), predictive densities assuming the skewed random effect appear to follow satisfactorily the data histograms. Notice that predictive distributions yielded by the customary normal model have failed to take account of the positive data skewness and seem at odds with the histograms. These graphical diagnostics point to the inadequacy of the normal random effect approach in modeling the generated data sets.

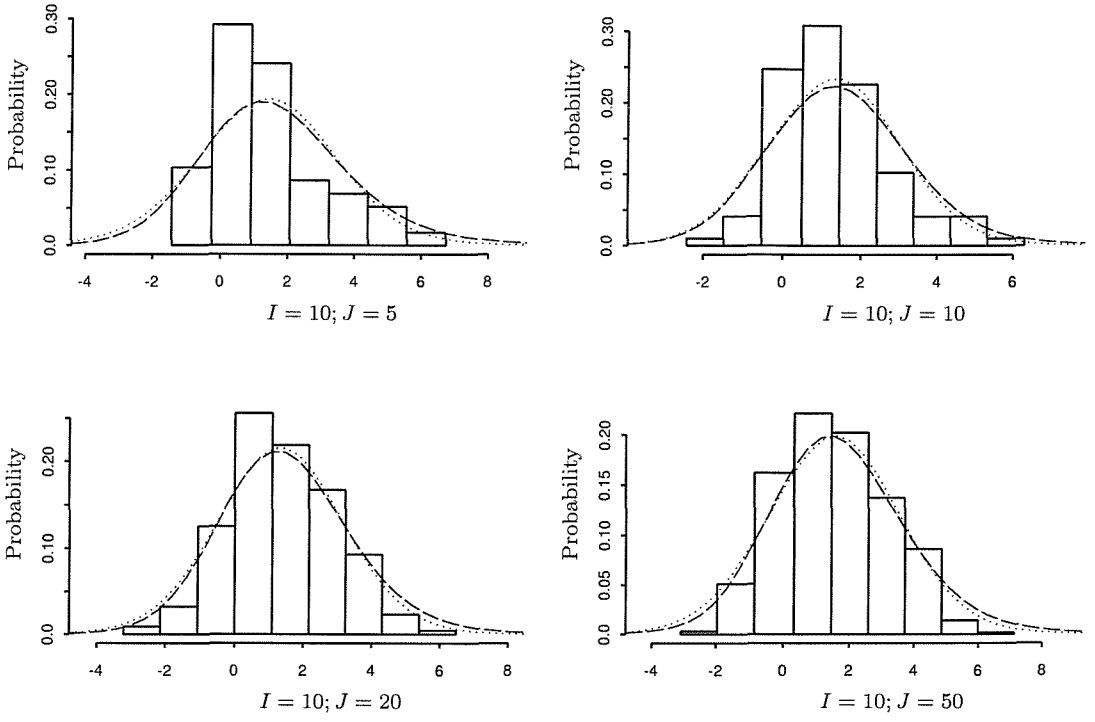

Figure 5.3: Histograms and predictive densities of the simulated examples for $I = 10$ clusters: dotted curves correspond to model with normal random effect; dashed curves are for $\mathrm{SN}_{\mathrm{sdb}}$ random effect model.

The models are now formally compared through the use of pseudo-Bayes factor (PsBF). Figure 5.7 shows a plot of PsBF against cluster size J for various number of clusters I. As would be expected, the values of PsBF under each combination of I and J considered are all larger than unity, thus lending support to the $\mathrm{SN}_{\mathrm{sdb}}$ random effect model. However, it is clear from the graph that evidence in favor of skewness is dependent on the amount of information

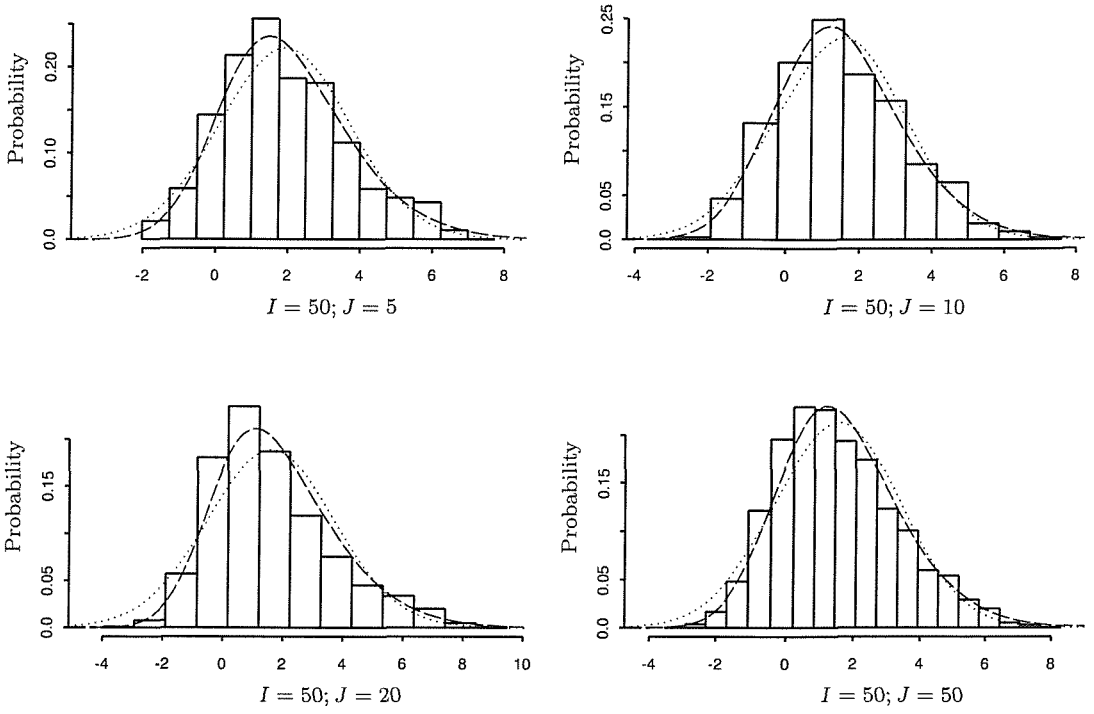

Figure 5.4: Histograms and predictive densities of the simulated examples for $I = 50$ clusters: dotted curves correspond to the model with normal random effect; dashed curves are for $\mathrm{SN}_{\mathrm{sdb}}$ random effect model.

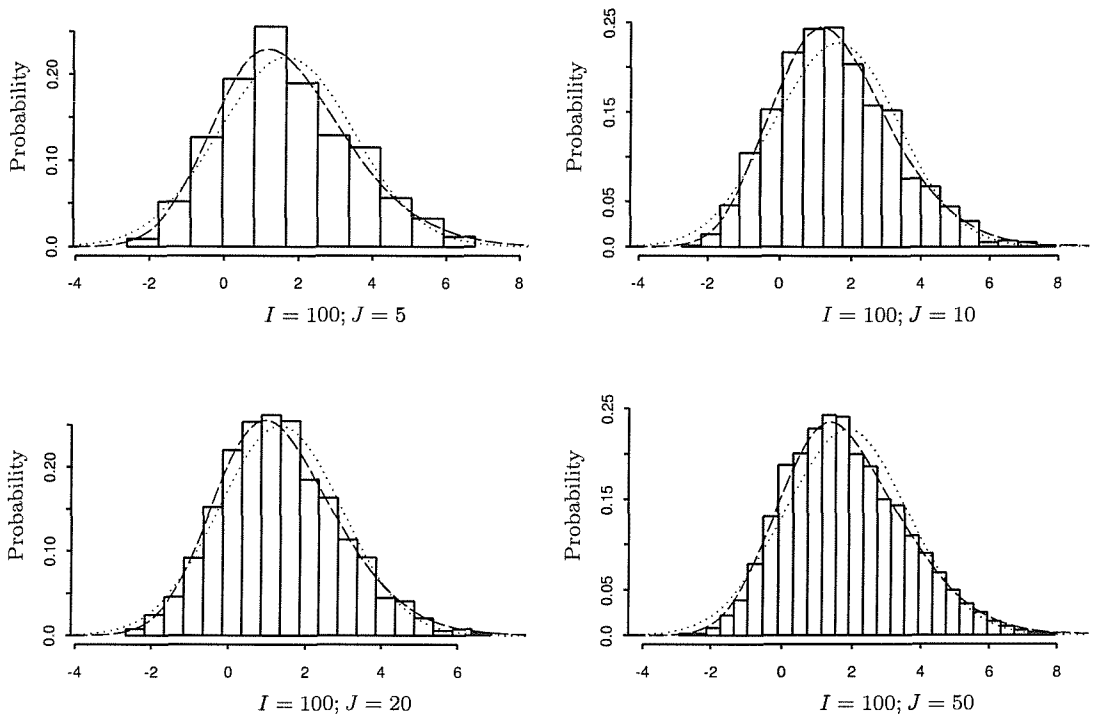

Figure 5.5: Histograms and predictive densities of the simulated examples for $I = 100$ clusters: dotted curves correspond to the model with normal random effect; dashed curves are for $\mathrm{SN}_{\mathrm{sdb}}$ random effect model.

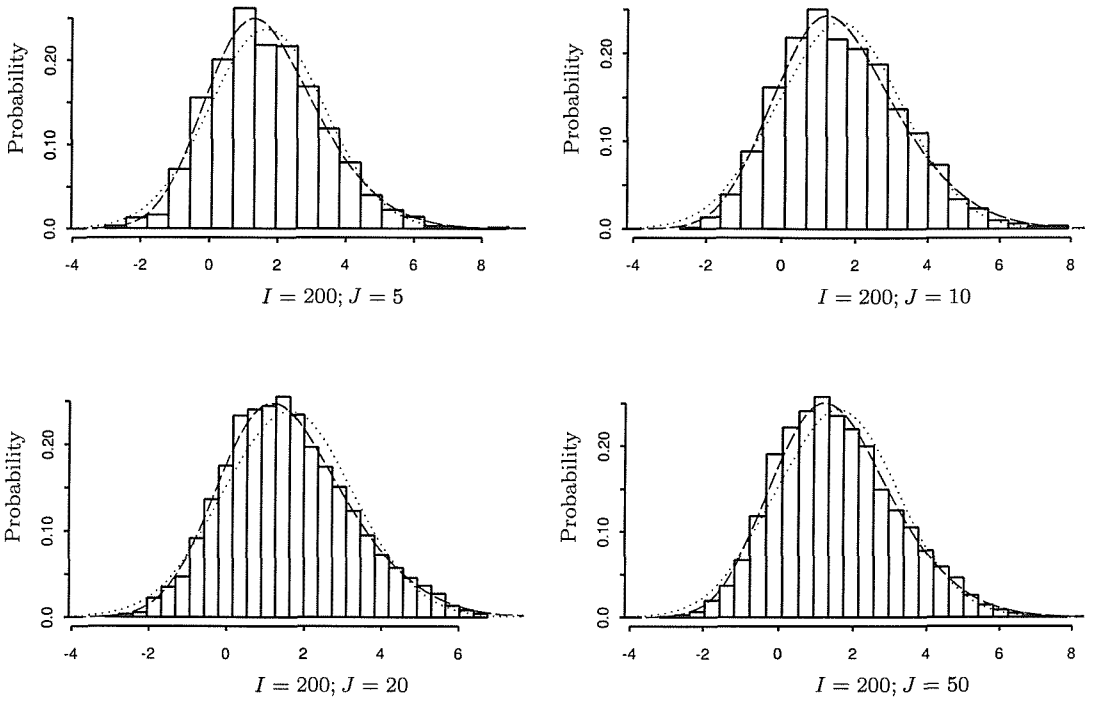

Figure 5.6: Histograms and predictive densities of the simulated examples for $I = 200$ clusters: dotted curves correspond to the model with normal random effect; dashed curves are for $\mathrm{SN}_{\mathrm{sdb}}$ random effect model.

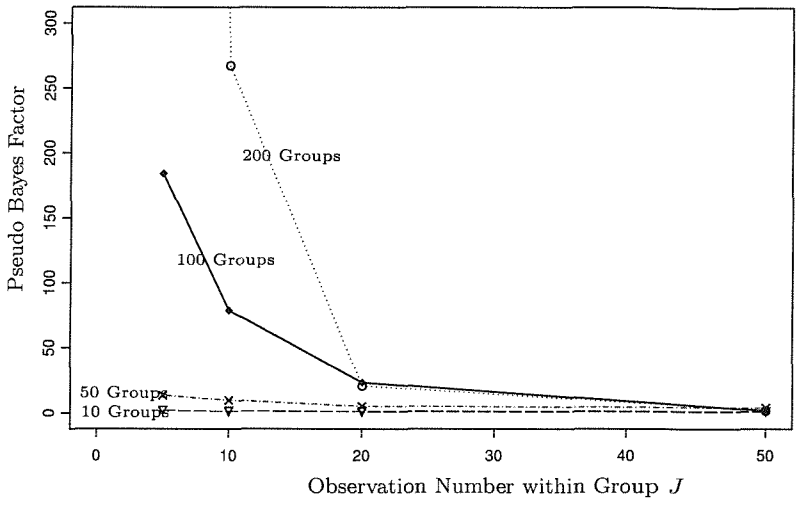

Figure 5.7: The pseudo-Bayes factor for the SN_{sdb} versus the normal random effect models for different sample sizes.

contained in the simulated data. Skewed random effect obviously enjoys the best diagnosis when I is biggest and J is smallest. In general, its superiority over the normal one rises as more groups becomes available but lessens with increasing number of observations within group. Deviance information criterion (DIC) given in Table 5.3 can be employed to confirm the above results. A small DIC value indicates a good model, see Spiegelhalter *et al.* (2002) for a detailed discussion. For ease of comparison, Figure 5.8 presents a graphical illustration of the DIC disparity between the candidate models. The overall findings is essentially in harmony with the PsBF calculations.

5.6 Summary discussion

This chapter has presented Bayesian analysis of variance components model where random cluster effect arises from a SN_{sdb} distribution. Our empirical investigations demonstrate that the overall fit of data can be significantly improved by using the asymmetric group effect instead of the popular normal assumption. The proposed model provides a much more accurate description

I	J	Normal DIC	SN_{sdb} DIC
10	5	139.632	138.357
	10	308.351	307.216
	20	577.108	576.388
	50	1403.720	1403.510
50	5	742.138	738.411
	10	1508.200	1503.330
	20	2880.990	2877.430
	50	7267.510	7264.320
100	5	1525.650	1516.570
	10	2896.020	2887.050
	20	5651.600	5646.350
	50	14277.700	14276.400
200	5	3013.050	2997.160
	10	5840.880	5827.420
	20	11531.600	11525.900
	50	28580.000	28577.900

Table 5.3: The DICs of different models considered for the simulated examples.

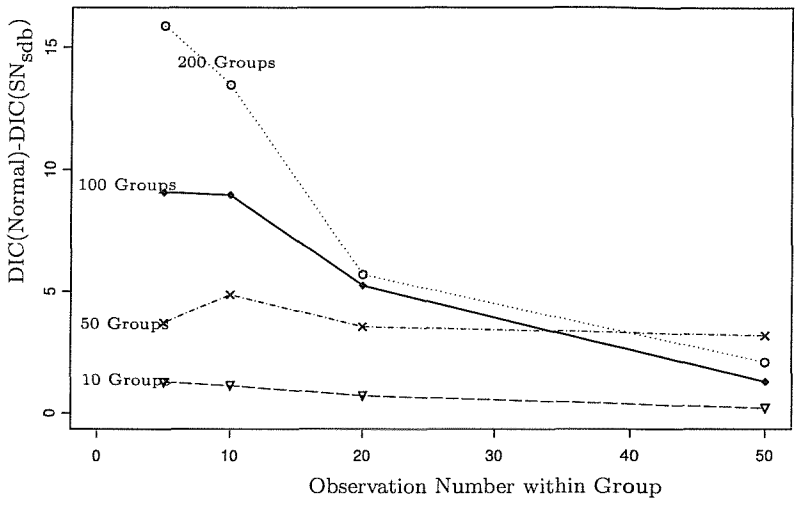

Figure 5.8: Difference in DIC between the SN_{sdb} and normal random effect models for varying number of groups and group sizes.

of the generated data sets, with predictive results very closely matching the underlying skewed histograms. A large number of clusters, however, is required to justify the use of SN_{sdb} random effect. Pseudo-Bayes factor and deviance information criterion have been considered in connection to the problem of model selection. The degree of preference given by these model choice criteria is shown to be strongly related to the sample sizes. Although both criteria deliver no appreciable evidence in support of skewness when the number of observations within group is large, the new methodology may still be more appropriate owing to its superiority in terms of goodness of fit.

The clear message is that the conventional normal variance components model does not always yield legitimate inferences for a given set of data. Analytical and empirical results here could provide a foundation for developing models which give a better approximation of reality. With group effects exhibiting kurtosis and skewness beyond those permitted by the normal distribution, a more flexible assumption, perhaps employing the SN_{new} or skew t families, would seem necessary. Development for the SN_{new} random effect model should involve nothing new in principle, though at some computational expense.

Chapter 6

Applications to survival analysis

6.1 Introduction

Survival analysis is concerned with the modeling of time-to-event data. In general, the data consist of a response variable measuring the duration until some specified event and a set of explanatory variables thought to be associated with this event-time variable. It is the main purposes of a survival analysis to model the underlying event times distribution and to ascertain the relationship between the response and the explanatory variables. This chapter employs methods analogous to the linear regression approach in Chapter 4 for analyzing survival data. More precisely, we assess covariate effects on logarithmic transformation of event times using a linear model representation with independent skew normal residual terms. Since it seems best to illustrate the construction of the statistical model in a particular example, we motivate the development with the laryngeal cancer data reported by Kardaun (1983).

The outline of the present chapter is thus the following. We start in Section 6.2 by introducing the data set. Section 6.3 fully sets out a Bayesian model founded upon skew normal errors assumption in log time. We utilize a numerical procedure, namely the Gibbs sampler, so as to conduct posterior inference with this model. In Section 6.4, simulation results are presented and comparisons are made with the normal residual approach. Finally, Section 6.5 draws some comments concerning the proposed techniques.

6.2 The laryngeal cancer data

The data set of interest here is quoted from Kardaun (1983). It comprises record on survival times of $n = 90$ male patients diagnosed with laryngeal cancer during the period 1970–1978 at a hospital in the Netherlands. Survival times reported measure the intervals (in years) between first treatment and either death or end of the study (January 1, 1983). The pertinent covariate information is contained in two variables: age at the time of diagnosis and stage of the cancer. The variable stage classifies the patients into four different groups according to the severity of the disease. More formally, it is a categorical variable taking four possible values: Stage 1, 2, 3 and 4 (ordered from least to most serious). There were 33 patients in Stage 1, 17 patients in Stage 2, 27 patients in Stage 3 and the remaining 13 patients were in Stage 4. Interest of the study centers on the effects of stage and age on survival time. As preliminary analysis of the data, Figure 6.1 depicts the Kaplan-Meier survival function estimates for patients within each of the four stages. By definition, the survival function is simply the probability of surviving beyond time t. Thus the curves indicate that individuals in higher stages of the cancer tend to have a shorter lifetimes on average. To confirm this, we shall conduct a comprehensive Bayesian analysis for the laryngeal cancer data in the coming sections.

6.3 Model specification

This section discusses a statistical formulation for the above problem. Let T_i be a nonnegative continuous random variable representing the survival time of the ith patient, $i = 1, \cdots, n$. Denoting the covariate values by

x_{i1} = 1if the ith patient is in stage 2, 0 otherwise;

x_{i2} = 1if the ith patient is in stage 3, 0 otherwise;

x_{i3} = 1if the ith patient is in stage 4, 0 otherwise; and

x_{i4} = age of the ith patient,

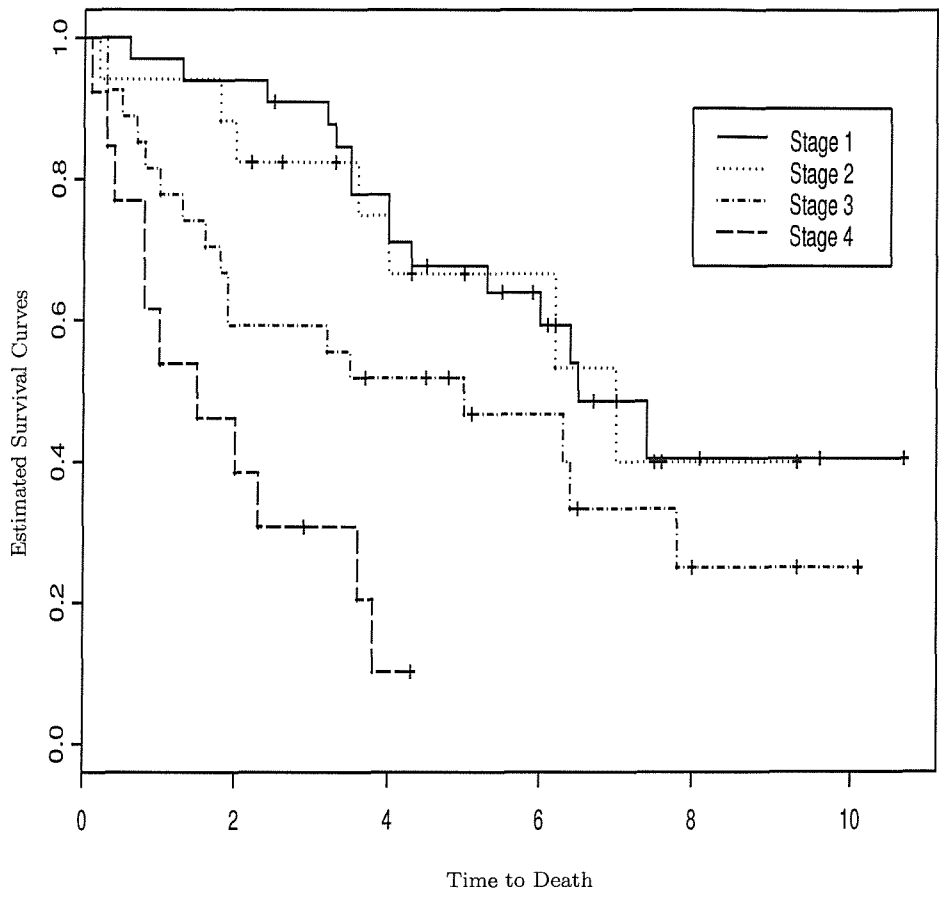

Figure 6.1: Kaplan-Meier survival curves for larynx cancer patients.

we assume that the data set is a realization from

$$\begin{aligned} Y_i = \log(T_i) &= \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i. \end{aligned} \tag{6.1}$$

Here $\boldsymbol{\beta}$ is the vector of regression parameters and ϵ_i typifies the discrepancy between a transformed observation $\log(t_i)$ and $\mathbf{x}_i^T\boldsymbol{\beta}$. The most common distributional assumptions for the random disturbance include the normal, the logistic and the extreme value families. In the present investigation, we consider an alternative density specification using that of a skew normal distribution. Formally,

$$\epsilon_i \sim \mathrm{SN}_{\mathrm{sdb}}(0, \delta, \sigma^2), \qquad i = 1, \cdots, n.$$

Note that the reason for modeling $\log(\mathbf{T})$ instead of $\mathbf{T}$ is to ensure a positive estimation on survival times.

Model (6.1) is, in effect, equivalent to the regression model in Chapter 4. However, the method of analysis is not as straightforward due to incomplete survival information on some individuals. For patients who were still alive at the termination of data collection, we know the lower bound of their survival time but not the actual time to death. Such partial observation of event time, called right censoring, has to be handled properly. Simply ignoring the censored measurements or treating them as if they were uncensored could lead to substantial biased results. For more details on censoring, see Klein and Moeschberger (1997).

Likelihood function of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \delta)$ is derived as follows. Suppose that the skew normal probability density function and distribution function are denoted by $h(\cdot)$ and $H(\cdot)$, respectively. When t_i corresponds to an observed death time, it contributes a term $h\{y_i = \log(t_i)|\boldsymbol{\theta}\}$ to the likelihood in the usual way. If t_i represents a censoring time, then what can only be sure of is that $T_i > t_i$. In other words, the contribution becomes $1 - H\{y_i = \log(t_i)|\boldsymbol{\theta}\}$. Thus the overall likelihood can be written as

$$L(\boldsymbol{\beta}, \sigma^2, \delta|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} \{h(y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \delta\}^{\zeta_i} \{1 - H(y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \delta)\}^{1-\zeta_i},$$

where ζ_i signifies a censoring indicator taking value one if the ith patient died during the study and zero otherwise.

We adopt the Bayesian approach to make statistical inference. Hence, model specification is completed by placing a suitable prior distribution on $\boldsymbol{\theta}$. For illustrative purposes, assume that all components of $\boldsymbol{\theta}$ are independent a-priori taking

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}) \times p(\sigma^2) \times p(\delta).$$

We continue to use normal priors with mean zero and variance 10^{10} for each regression coefficient β_j, $j = 1, \cdots, 5$. The parameter σ^2 is given the proper inverse gamma prior distribution $IG(10^{-3}, 10^{-3})$ as in Chapter 4. Finally, prior opinion about δ is modeled through $N(0, 10^{10})$. Notice that large variances are specified so as to represent lack of prior knowledge about the model parameters. This prior selection may be slightly unrealistic, but we work with them for easy of computations.

Simple use of the Bayes theorem (2.1) reveals the expression

$$p(\boldsymbol{\beta}, \sigma^2, \delta, \mathbf{x}_i|\mathbf{y}) \propto \prod_{i=1}^{n} \{h(y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \delta)\}^{\zeta_i} \{1 - H(y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \delta)\}^{1-\zeta_i} \times p(\boldsymbol{\beta}, \sigma^2, \delta).$$

This joint posterior density is rather complicated to allow exploration using analytical techniques, so numerical methods have to be used instead. Our approach to posterior inference is based upon the Gibbs sampler with data augmentation. The Markovian updating scheme proceeds as below.

1. Fix $j = 0$ and initiate the chain with a random guess, say,
$$\boldsymbol{\theta}^{(j)} = (\beta_0^{(0)}, \cdots, \beta_0^{(0)}, \sigma^{2(0)}\delta^{(0)}).$$

2. If $\zeta_i = 0$, draw $y_{\text{new},i}$ from $h(y_{\text{new}}|\boldsymbol{\theta}^{(j-1)})$ such that $\exp(y_{\text{new},i}) > t_i$, $i = 1, \cdots, n$.

3. Impute $y_{\text{new},i}$ for each of the censored y_i. Denote the new 'complete transformed data' by $\mathbf{y}^*$.

4. Draw $\boldsymbol{\theta}^{(j)}$ from $p(\boldsymbol{\theta}|\mathbf{y}^*)$ via the usual Gibbs sampling procedures.

5. Repeat steps 2–4 for $j = 1, \cdots, J$.

Linear regression structure on $\mathbf{Y} = \log(\mathbf{T})$ means that $p(\boldsymbol{\theta}|\mathbf{y}^*)$ bears precisely the same form as the posterior distribution in Chapter 4. So, details about the relevant full conditional posterior distributions is omitted here.

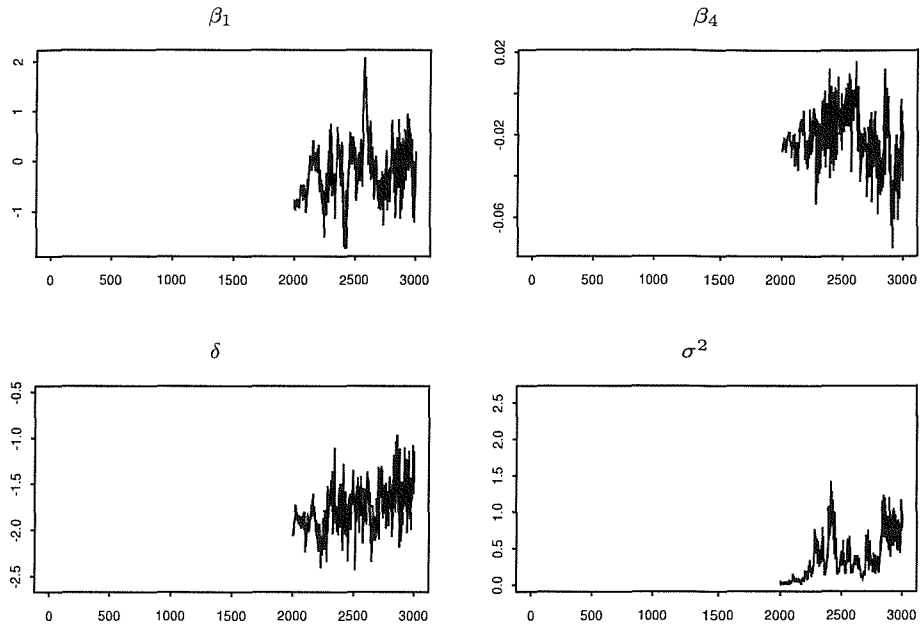

Figure 6.2: Post-convergence sequence plots of certain model parameters for the laryngeal cancer example.

6.4 Inferences

Model (6.1) can be fitted using general purpose Gibbs sampling software `WinBUGS`. We have checked convergence by exercising the graphical methods described in Section 4.4.1. The approach did not detect any problem in convergence. Some successive MCMC outputs are displayed in Figure 6.2. As a summary of the replication outcomes, Table 6.1 tabulates the estimated posterior mean, standard deviation and 95% credible interval of regression coefficients and shape parameters. These statistics are acquired from 200,000 simulated values after a burn in of 10,000 iterations. In order to find out how skewness affects the posterior results, we consider model under normal error assumption for comparison. For ease of referring to, such normal error model will be called (as is customary) the log-normal model, and the name log-skew-normal will be given to our proposed model henceforth. Prior distributions set for $\boldsymbol{\beta}$ and σ^2 in the previous section are again employed in the log-normal model. This serves to keep the prior information as consistent as possible in the comparison sense.

	Log-normal		Log-skew-normal	
	Posterior mean (s.d.)	95% credible interval	Posterior mean (s.d.)	95% credible interval
β_0' (true intercept)	3.55 (1.03)	(1.60,5.63)	3.43 (0.98)	(1.62,5.48)
β_1 (stage 2)	-0.18 (0.48)	(-1.13,0.78)	-0.14 (0.51)	(-1.12,0.95)
β_2 (stage 3)	-0.93 (0.40)	(-1.72,-0.15)	-0.78 (0.42)	(-1.63,0.01)
β_3 (stage 4)	-1.90 (0.48)	(-2.88,-0.97)	-1.73 (0.47)	(-2.74,-0.90)
β_4 (age)	-0.02 (0.01)	(-0.05,0.01)	-0.02 (0.01)	(-0.05,0.01)
σ^2	1.90 (0.45)	(1.21,2.95)	0.70 (0.74)	(0.001,2.34)
δ			-1.07 (1.29)	(-2.35,2.23)

Table 6.1: Parameter estimates from normal error and skew normal error models in the larygeal cancer study.

Sample-based estimates from the log-normal model are summarized in Table 6.1 also. As the table indicates, both alternatives lead to reasonably cohesive inference on regression coefficients. However, a reinspection of the findings shows that all five coefficients are slightly attenuated after accounting for skewness. Figure 6.3 pictorially demonstrates this pattern for several parameters. Noticeably, the estimate of σ^2 is smaller under the skewed model. Such reduction is anticipated since the non-zero skewness parameter also explains some variability of the residuals. According to the 95% credible interval of δ, the skewed modeling seems not worthy in improving the overall fit to the data. Direct examination of the marginal density in Figure 6.4 provides new perspective, however. It turns out that the posterior of δ shrinks towards a negative value but endures an exceptionally heavy upper tail. In view of the location of the clear maximum, introducing skewness in data fitting could be important.

We consider L^2-criterion defined by (2.6) in connection to model determination. The L^2-criterion value for the log-skew-normal model is 258.6. On the other hand, the log-normal case leads to $L^2 = 280.0$, which is 21.4 calibration units larger. The information suggests that model (6.1) is the more likely generating mechanism despite of the statistical insignificance of δ. In an attempt to check the quality of model fit, we compare the non-parametric Kaplan-Meier

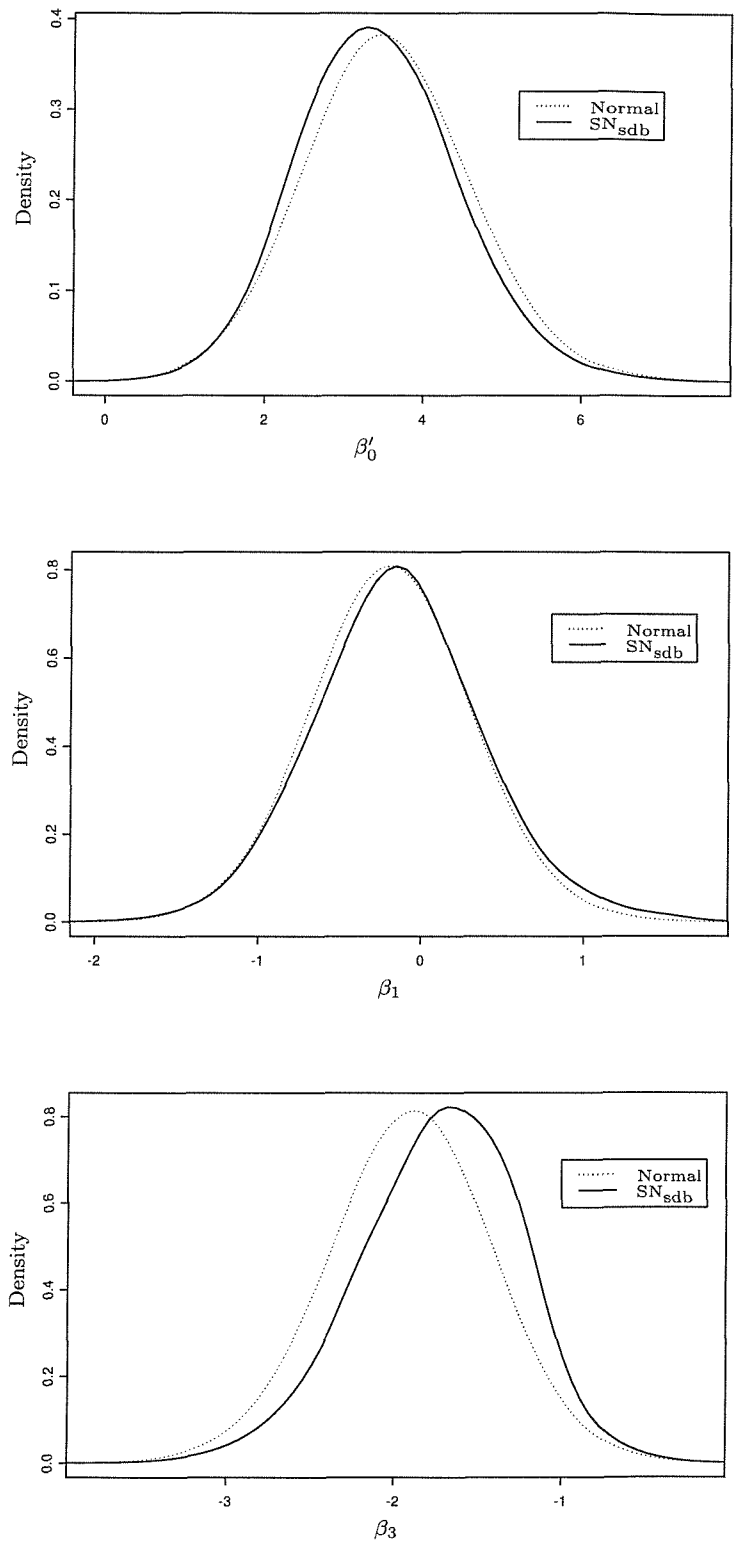

Figure 6.3: Estimated marginal posterior densities for β_0' (true intercept), β_1 and β_3 in the laryngeal cancer example.

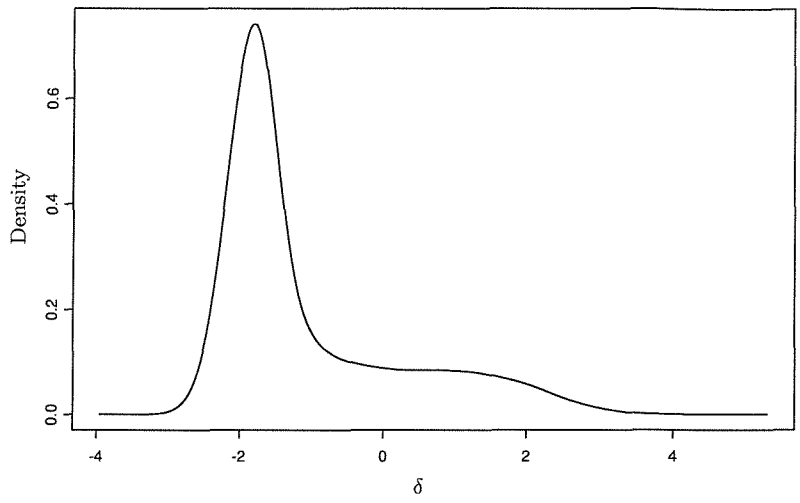

Figure 6.4: Kernel density estimate of skewness parameter for the laryngeal cancer example.

survival function with the relevant predictive estimates. Note that it is invalidated to use data histogram as benchmark for assessing model adequacy because of the censoring involved. Consequently, we have to resort to the non-parametric outcomes as a rough guide. The survival curves are reported in Figures 6.5 and 6.6. Predictive survival functions under the skew normal error specification looks quite in line with the non-parametric ones, hinting an appropriate summary for the data. The log-normal model yields almost identical predictive curves as those obtained from model (6.1) for the first two stages, but it appears to produce relatively inferior estimates for the other cases.

Summing up, our fitted regression line under the preferred log-skew-normal model is expressible as

$$E\{\log(T_i)\} = 3.43 - 0.14X_1 - 0.78X_2 - 1.73X_3 - 0.02X_4.$$

The negative values for the coefficient of X_1, X_2 and X_3 indicate that patients in stages $2-4$ have shorter lifetimes than individuals in the reference group (stage 1). In addition, the chance of dying is greater for persons with higher stage of laryngeal cancer. These conclusions are essentially in agreement with the content of Figure 6.1. However, by considering the pertinent 95% credible intervals, it is determined that times to death only differ substantially between

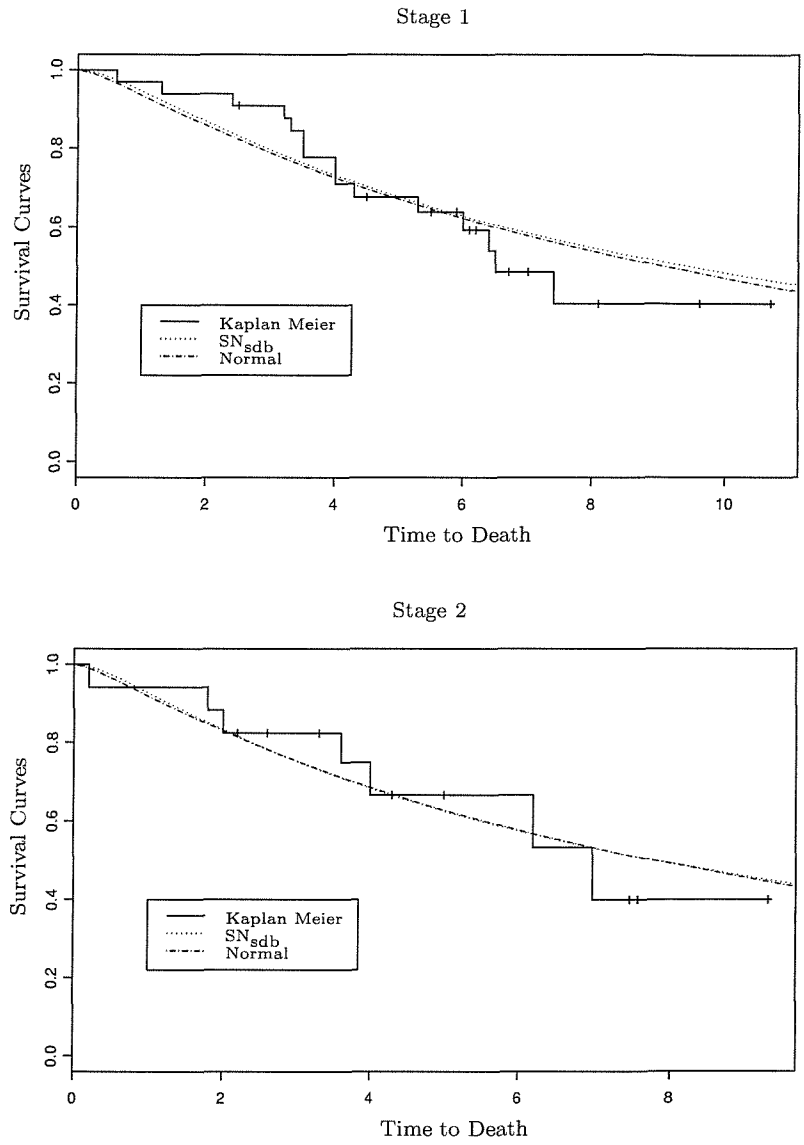

Figure 6.5: Predictive survival curves for larynx cancer patients in Stages 1 and 2.

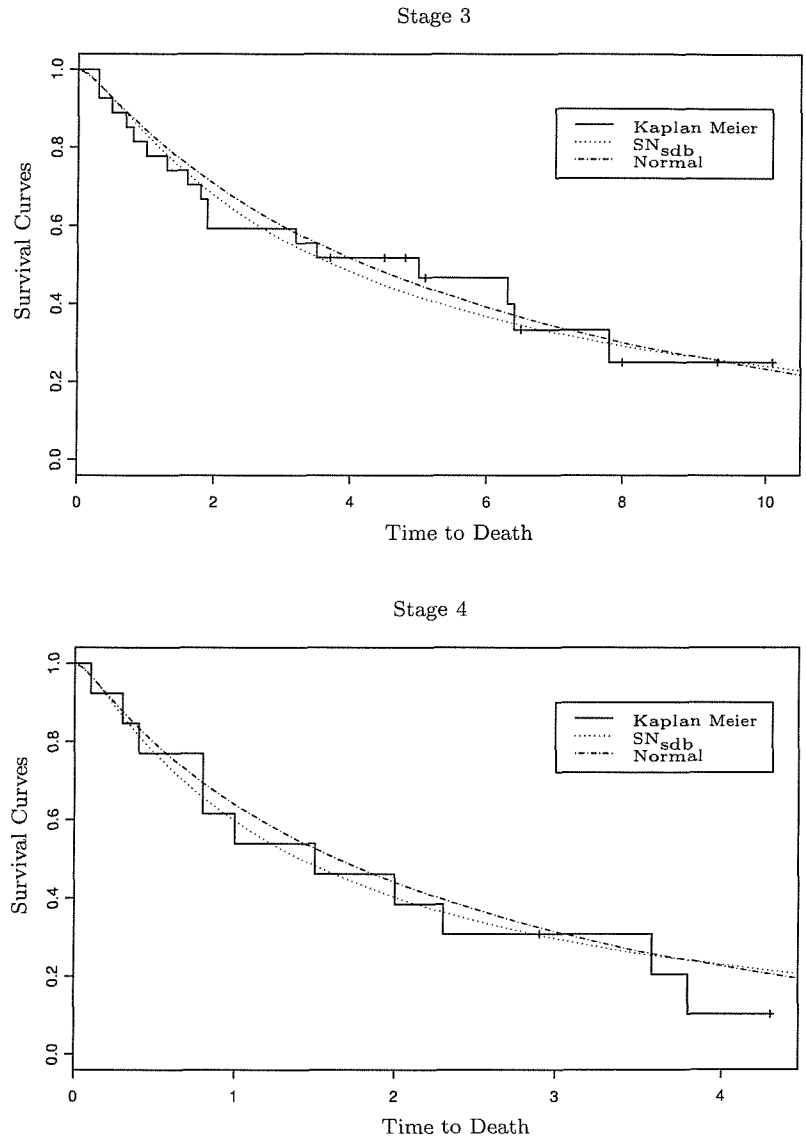

Figure 6.6: Predictive survival curves for larynx cancer patients in Stages 3 and 4.

stages 1 and 4 cancer patients. Specifically, the average survival time of a person in stage 1 is about 5.64 ($\approx$ exp(1.73)) times the same for a stage 4 disease individual. Patient's age at diagnosis also has some negative impact on time to death. Nevertheless, the effect seems not statistically significant since the 95% credible interval of β_4 includes the point zero

6.5 Summary and conclusions

In this chapter, we have proposed the use of log-skew-normal model for modeling univariate survival times. Our approach to analysis was illustrated in detail with the laryngeal cancer example. Although positive skewness is often an observed characteristic in time-to-event data, logarithmic responses can be symmetric or even negatively skewed. The new model offers flexibility to deal with possibly asymmetric error distribution. More importantly, it shows promise as a potential means of improving the overall fit to this type of data. Recall that attenuation on covariate effect(s) was also encountered in both the non-academic scores and the Martin Marietta data examples, refer to Chapter 4 for details. Whether or not this phenomenon is just a coincidence remains pending for further investigation. We would postulate that the underlying right censoring to be a possible cause for the anomalous posterior tail behavior of δ, but a full exploration in this area is necessary.

Chapter 7

Overall conclusions and future work

7.1 Conclusions

This thesis proposed a new class of multivariate skew elliptical distributions and examined its suitability in analyzing practical data sets. The family is a simple generalization of the proposal of Sahu, Dey and Branco (2003). Sahu *et al.* developed the m-variate asymmetric distributions by conditioning on m unobserved random variables. We have extended their results so that skewness is generated by $p \in \aleph$ (not necessarily equal to m) latent variables.

The primary focus of the current research is to concentrate on the particular case of the univariate skew normal distribution. We provided results for general p, but were usually most interested in $\mathrm{SN}_{\mathrm{new}}$ (density with $p = 2$). Some appealing features of $\mathrm{SN}_{\mathrm{new}}$ are as follows.

- It allows continuous variation from normality to non-normality.
- It acknowledges a wide range of indices of skewness, whilst offering flexibility to account for some non-normal peakedness.
- It is very easy to simulate observations from the distribution.
- Although the pertinent density function is rather intricate, empirical analysis based on MCMC methods is feasible.

Hence the distribution can be valuable for fitting data exhibiting skewness and non-normal peakedness.

There are, however, two major drawbacks to this $\mathrm{SN}_{\mathrm{new}}$ distribution. Firstly, the density range of achievable kurtosis is restricted to the interval $(0, 0.87)$. This means that $\mathrm{SN}_{\mathrm{new}}$ can only yield tails thinner than the normal ones, making it not appropriate for analyzing data with extreme observations. Secondly, skewness parameters of the distribution possess the same distributional role, which complicates both the issues of interpretation and estimation. Similar limitations and the aforesaid advantages, yet to be established formally, may be expected to hold for the general p setting.

In order to illustrate the potential of the skew normal distribution for data analysis, we have presented applications to regression, variance components and survival models. The methodologies were exemplified through computer generated and real data examples. Note that all the empirical studies in this thesis were within the Bayesian framework. In addition, we treated the response variable in its original units and assumed a hierarchical model with conjugate priors, the latter being typically arbitrarily vague. The much computational burden in Bayesian inference was resolved through the use of Gibbs sampling. We saw that the simulation technique is trivial to specify distributionally and to implement computationally. The calculations were greatly facilitated by the free software `WinBUGS`.

The main conclusion from our illustrative analysis is that asymmetry and non-normal peakedness in practical data can be captured by the $\mathrm{SN}_{\mathrm{new}}$ distribution. It has the capability of producing more precise inference than other existing skew normal densities. In spite of the unpleasant exchangeability feature of the skewness parameters, characteristic measures such as location and spread are still identifiable from the Gibbs outputs. Relevant questions on parsimony can be easily sorted out using Bayesian model selection criteria. Experience shows that $\mathrm{SN}_{\mathrm{new}}$ gains no advantage over a simpler model if its skewness parameters joint posterior distribution has appreciable probability mass near zero. Example 2 in Chapter 4 provides a case in point. Moreover, the new distribution is more suitable for large data sets.

7.2 Recommendations for future research

Our suggestions for future study fall into two categories, namely further developments and applications of the proposed distributions.

Among the further developments, one area that deserves specific exploration is the direct specification of location, scale, skewness and kurtosis factors in the $\mathrm{SN}_{\mathrm{new}}$ density formula. This re-parametrization can be beneficial in various aspects.

1. Distributional parameters now have an intuitive interpretation.
2. Prior independence between parameters becomes a more plausible assumption.
3. It would circumvent the identifiability problem of MCMC outputs.
4. Faster convergence may be anticipated in Gibbs sampling.

As the number of skewness parameters increases ($p \geq 3$), theoretical considerations might hint at extending the skew normal distribution in other directions. We would hope that the skewing proposals listed in Section 3.7 can generalize the class without discarding flexibility and interpretability, but this awaits detailed investigations.

Besides those discussed in Chapters 4 – 6, the skew normal family is potentially applicable to a broad range of other empirical problems. A natural aspect to consider is the appropriateness of the class in linear statistical methods outside normality. As an example, the density can be convenient for creating skewed link function in generalized linear models. The reader is referred to Albert and Chib (1993) and Chen, Dipak, and Shao (1999) for some related work in this direction. Another possibility that is worth exploring is the relevance of the skew normal distribution in modeling dependent data. Applications of multivariate skew normal distributions to financial data can be found in Adcock (2004) and references therein. It may also be meaningful to consider employing the proposed distributions as prior densities in Bayesian analysis, see O'Hagan and Leonard (1976). Many issues concerning the skew normal distribution still

need to be explored, and experience has to be accumulated even further with the methods considered in this thesis.

Appendix A

BUGS and CODA Softwares

BUGS, stands for Bayesian inference Using Gibbs Sampling, is an MCMC software developed at the MRC Biostatistics Unit at the University of Cambridge. It uses a set of S-like syntax for specifying sampling models, priors and the data. The program then converts this syntax into an internal direct acyclic graph and selects the corresponding sampling method. By successively samples from the nodes of the graph, a Markov chain of simulated values is produced and output to a file for subsequent analysis. Convergence diagnostic of a BUGS output can be performed through CODA, stands for Convergence Diagnosis and Output Analysis. It is an S-Plus package designed to investigate MCMC output that may but does not need to be used in conjunction with BUGS. A Windows version of BUGS is WinBUGS. It is more user friendly and includes many useful tools that are not part of BUGS. These programs are freely available over the web at

http://www.mrc-bsu.cam.ac.uk/bugs/

Appendix B

BUGS code for SN_{new} linear regression model

For the non-academic scores example:

```
model
{
    for(i in 1:N)
    {
        y[i] ~ dnorm(mu[i], tau)
        x5[i] <- x1[i]*x3[i]
        mu[i] <- beta0+beta1*(x1[i]-mean(x1[]))+beta2*x2[i]+
            beta3*(x3[i]-mean(x3[]))+beta4*(x4[i]-mean(x4[]))+
            beta5*(x5[i]-mean(x5[]))+delta1*z1[i]+delta2*z2[i]
        temp1[i] ~ dnorm(0,1)
        z1[i] <- abs(temp1[i])
        temp2[i] ~ dnorm(0,1)
        z2[i] <- abs(temp2[i])
    }
}
```

Appendix C

BUGS code for SN_{sdb} random effect model

For the simulated examples:

```
model
{
    for(i in 1:I)
    {
        m[i] ~ dnorm(theta[i], tau.g)
        for(i in 1:J) y[i,j] ~ dnorm(mu[i], tau.e)
        theta[i] <- mu+delta*z[i]
        temp[i] ~ dnorm(0,1)
        z[i] <- abs(temp2[i])
    }
}
```

References

[1] Adcock, C. (2004). Capital asset pricing for UK stocks under the multivariate skew-normal distribution. In *Skew-elliptical distributions and their applications: A journey beyond normality*. (Eds. M. G. Genton). *London: Chapman and Hall*, 191–204.

[2] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

[3] Arnold, B. C. and Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Sociedad de Estadística e Investigaciacuteon Operativa Test*, **11**, 7–54.

[4] Arnold, B. C. and Beaver, R. J. (2000). The skew-Cauchy distribution. *Statistics & Probability Letters*, **49**, 285–290.

[5] Arnold, B. C., Beaver, R. J., Groeneveld R. A. and Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, **58**, 471–488.

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

[7] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

[8] Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society*, B, **61**, 579–602.

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

[10] Box, G. E. P. and Tiao, G. (1973). Bayesian inference in statistical analysis. *Reading, Massachusetts: Addison-Wesley*.

[11] Branco, M. D. and Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99–113.

[12] Butler, R. J., McDonald, J. B., Nelson, R. D. and White, S. B. (1990). Robust and partially adaptive estimation of regression models. *The Review of Economics and Statistics*, **72**, 321–327.

[13] Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function. *Communications in Statistics - Theory and Methods*, **19**, 197–203.

[14] Chen, M. H., Dipak, K. D. and Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, **94**, 1172–1186.

[15] Chen, M. H., Shao, Q. M. and Ibrahim, J. G. (2000). Monte Carlo methods in Bayesian computation. *London: Springer-Verlag*.

[16] DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903–915.

[17] Fang, K. T., Kotz, S. and Ng, K. W. (1990). Symmetric multivariate and related distributions. *London: Chapman and Hall*.

[18] Fernandez, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, **93**, 359–371.

[19] Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

[20] Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov chain Monte Carlo in practice*. (Eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter). *London: Chapman and Hall*, 145–161.

[21] Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian statistics 4*. (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). *London: Oxford University Press*, 147–167.

[22] Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

[23] Gelman, A., Roberts, G. O. and Gilks, W. R. (1995). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*. (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). *London: Oxford University Press*.

[24] Genton, M. G. and Loperfido, N. (2001). Generalized skew-elliptical distributions and their quadratic forms. *Technical Report, North Carolina State University, USA*.

[25] Gibbons, J. F. and Mylroie, S. (1973). Estimation of impurity profiles in ion-implanted amorphous targets using joined half-Gaussian distributions. *Applied Physics Letters*, **22**, 568–572.

[26] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. *London: Chapman and Hall*.

[27] Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics*, **13**, 271–275.

[28] John, S. (1982). The three parameter two-piece normal family of distributions and its fitting. *Communications in Statistics – Theory and Methods*, **11**, 879–885.

[29] Jeffreys, H. (1961). Theory of probability. *London: Oxford University Press*.

[30] Jones, M. C. (2001). A skew t distribution. In *Recent advances in probability and statistics; in Honor of T. Cacoullos*. (Eds. C. A. Charalambides, M. V. Koutras and N. Balakrishnan). *London: Chapman and Hall*, 269–278.

[31] Jones, M. C. and Faddy, M. J. (2003). A skew extension of the t distribution. *Journal of the Royal Statistical Society*, B, **65**, 159–174.

[32] Kardaun, O. (1983). Statistical analysis of male larynx-cancer patients – a case study. *Statistical Nederlandica*, **37**, 103–126.

[33] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–7395.

[34] Kelker, D. (1970). Distribution thoery of spherical distributions and a location-scale parameter generalization. *Sankhyā*, **32**, 831–860.

[35] Kimber, A. C. (1985). Methods for the two-piece normal distribution. *Communications in Statistics – Theory and Methods*, **14**, 235–245.

[36] Kimber, A. C. and Jeynes, C. (1987). An application of the truncated two-piece normal distribution to the measurement of depths of arsenic implants in silicon. *Applied Statistics*, **36**, 352–357.

[37] Klein, J. P. and Moeschberger, M. L. (1997). Survival analysis: techniques for censored and truncated data. *New York: Springer-Verlag*.

[38] Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society*, B, **57**, 247–262.

[39] Loperfido, N. (2002). Statistical implications of selectively reported inferential results. *Statistics & Probability Letters*, **56**, 13–22.

[40] Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalising constants via a simple identity: a theoritical exploration. *Statistica Sinica*, **6**, 831–860.

[41] Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew-normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference*, **83**, 291–309.

[42] O'Hagan, A. and Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, **63**, 201–203.

[43] Runnenberg, J. T. (1978). Mean, median, mode. *Statistical Nederlandica*, **32**, 73–79.

[44] Sahu, S. K., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regressian models. *The Canadian Journal of Statistics*, **31**, 129–150.

[45] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B, **64**, 1–34.

[46] Theodossiou, P. (1998). Financial data and the skewed generalized T distribution. *Management Science*, **44**, 1650–1661.