Bayesian inference on the scalar skew-normal distribution

Stefano Cabras\textsuperscript{1}, Walter Racugno\textsuperscript{1} and Laura Ventura\textsuperscript{2}

\textsuperscript{1} Department of Mathematics, University of Cagliari, Italy, s.cabras@unica.it
\textsuperscript{2} Department of Statistics, University of Padova, Italy, ventura@stat.unipd.it

Summary. In this paper we discuss a Bayesian analysis of the scalar skew-normal model. This model defines a class of distributions that extends the Gaussian model by including a shape parameter. Although the skew-normal model has nice properties, it presents some problems with the estimation of the shape parameter. To avoid these drawbacks, we explore through some examples the use of Severini’s integrated likelihood function in Bayesian inference.

Key words: Jeffrey’s prior, Integrated likelihood, Nuisance parameter, Reparametrization.

1 Introduction

Consider observations $y = (y_1, \ldots, y_n)$ from a random variable of the form $y = \mu + \sigma z$, where $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter, and $z = (z_1, \ldots, z_n)$ is a random sample from the skew-normal (SN) model with shape parameter $k$, given by [Azz85, Azz05]

$$p(z; k) = 2 \varphi(z) \Phi(kz), \quad y \in \mathbb{R}, \quad k \in \mathbb{R},$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the standard Gaussian density and distribution function, respectively. The shape parameter $k$ is a real parameter such that when $k = 0$ one gets back to the standard normal distribution. Applications of model (1), which can be easily extended to scale and regression model, may be found in many areas, including for example stochastic frontiers models (\cite{Adc04} and references therein), analysis of sexual dimorphism due to asymmetry \cite{Adc04} and analysis of capital asset pricing \cite{Adc04}.

The SN distribution has appealing properties on the probabilistic side, extending in a natural way those of the Gaussian distribution, and it is quite useful in modelling real data sets \cite{AC99}. Although these remarkable properties, the connected inferential steps have some undesirable aspects. Indeed, for small or even moderate sample size, $n$, the likelihood can be monotone in $k$. Moreover, for $k > 20$, SN models are not very different from each other, because the SN distribution tends
to the unique Half Normal distribution as $k \to \infty$, and point estimates cannot be precise [?, Azz05].

To avoid these drawbacks, several proposals have been discussed. In the frequentist approach, the method proposed by [?] is based on an alternative use of a classical bias correction to maximum-likelihood estimation, which never produces boundary estimates. In the Bayesian approach, [?] show that the Jeffrey’s prior of $k$ is proper, a particular situation for this class of priors, given that the range of $k$ is unbounded. Hence the posterior distribution for $k$ is always a proper distribution and, on the contrary of the mean, its mode or median produces always finite estimates, which are shown to have good frequentist properties.

The aim of this paper is to explore the use of Severini’s integrated likelihood [?] for Bayesian inference on $k$, when the scale and location parameters are considered as nuisance parameters. This integrated likelihood is based on a suitable reparametrization of the nuisance parameters, so that the latter is unrelated to the shape parameter $k$. For a well-known dataset generated with $k = 5$ but with $k = \infty$, we compare Severini’s integrated likelihood with the reference integrated likelihood [?] and we show that, on the contrary to the latter, Severini’s integrated likelihood is not monotone and tend to zero as $|k| \to \infty$. Moreover, we use Severini’s integrated likelihood with the Jeffrey’s prior on $k$ to obtain the posterior distribution for $k$. In this situation, the posterior mean is closer to the true $k$ than the posterior median in [?]. Other simulated datasets of different sample sizes for shape parameter $k$ larger than 5 are also discussed. Due to the mathematical formulation of Severini’s integrated likelihood, Bayesian analyses are based on Monte Carlo (MC) algorithm and adaptive quadrature integration (AQI), bypassing Markov Chain Monte Carlo (MCMC) integrations.

## 2 Integrated likelihoods

Suppose attention is devoted only to the shape parameter $k$ of the SN model, and the remaining parameter $\lambda = (\mu, \sigma)$ plays the role of the nuisance parameter. We focus on integration methods, based on the elimination of the nuisance parameter by simple integration, resulting in an integrated likelihood $L_1(k) = \int L(k, \lambda)\pi(\lambda|k)d\lambda$, where $L(k, \lambda)$ is the likelihood function and $\pi(\lambda|k)$ is the conditional prior distribution of the nuisance parameter given the parameter of interest $k$.

### 2.1 Reference integrated likelihood

In the framework of composite transformation models, a compelling choice for $\pi(\lambda|k)$ is the induced right invariant Haar density for $\lambda$, given by the default improper prior $\pi(\lambda|k) \propto 1/\sigma$ [?]. Note that $\pi(\lambda|k)$ is actually the conditional reference prior for $\lambda$ given $k$ [?]. The resulting integrated likelihood

$$L_H(k) = \int_{\mathbb{R}^+} \int_{\mathbb{R}} \sigma^{-1}L(k, \mu, \sigma)d\mu d\sigma \tag{2}$$

has a frequentist interpretation. It is the marginal likelihood for $k$ based on the maximal invariant statistic with respect to scale and location transformations (see
e.g. [?], section 2.8). From a Bayesian perspective, [?] obtain for (2) the closed form expression

\[ L_H(k) \propto F_{\Omega_R} \left( k \frac{y - \overline{y} 1_n}{s} \right) , \tag{3} \]

where \( F_{\Omega_R}(\cdot) \) is the distribution function of a centered \( n \)-dimensional Student’s \( t \) distribution with \( n \)-degrees of freedom, scale matrix \( \Omega_R = I_n + (k^2/n)1_n 1_n^T \), and \( y, \, s^2 \) are sample mean and variance, respectively.

As pointed out in [?], direct calculation of (3) is difficult especially for large \( n \), and a direct numerical integration in (2) can be more convenient. However, this approach presents all the drawbacks of MCMC algorithms discussed in [?]. Finally, \( L_H(k) \) does not vanish for large values of \(|k|\) [?], and this phenomenon makes difficult a direct use of \( L_H(k) \) in order to draw inference about \( k \).

### 2.2 Severini’s integrated likelihood

Alternatively to (2), we propose to use Severini’s integrated likelihood. In particular, [?] suggests that the conditional prior density \( \pi(\lambda | k) \) should be chosen by finding a nuisance parameter \( \phi \) that is unrelated to \( k \), and then taking the prior density for \( \phi \) to be independent of \( k \).

The nuisance parameter \( \phi \) is determined requiring that the partial MLE of \( \phi \) given \( k \), \( \hat{\phi}_k \), does not depend on \( k \). Following [?], \( \phi \) can be defined by the equation

\[ E \left( \frac{\partial \ell(k, \hat{\lambda}_k)}{\partial \lambda^T}; \hat{k}, \phi \right) = 0 , \]

where \( \ell(k, \lambda) = \log L(k, \lambda) \). Such a parameter \( \phi \equiv \phi(k, \lambda, \hat{k}) \) has the property that, for any value of \( k \), \( \phi_k = \lambda + O_p(n^{-1/2}) \), and we refer to \( \phi \) as a globally-orthogonal nuisance parameter. An important feature of \( \phi \) is that it depends on the data.

An integrated likelihood may then be formed by integrating with respect to a prior density \( \pi(\phi) \) giving

\[ L_S(k) = \int_{\phi} \tilde{L}(k, \phi) \pi(\phi) d\phi , \tag{4} \]

where \( \tilde{L}(k, \phi) \) denotes the likelihood for \((k, \phi)\) with \( \phi \in \Phi \). The prior \( \pi(\phi) \) is related to \( \pi(\lambda | k) \) since \( \pi(\lambda | k) = \pi(\phi(k, \lambda, \hat{k})) | \partial \phi(k, \lambda, \hat{k}) / \partial \lambda | . \pi(\lambda | k) \) depends on the data only through \( \hat{k} \) and \( L_S(k) \) can be viewed as the result of a two-stage procedure, similar to hierarchical Bayes models [?].

Function (4) has many interesting properties. First of all, from the frequentist point of view, the integrated likelihood \( L_S(k) \) is closely related to the modified profile likelihood [?], and it is score unbiased and information unbiased to order \( O(n^{-1}) \). Moreover, from the Bayesian point of view, \( L_S(k) \) is nearly independent of the choice of the prior density, since \( L_S(k) \) approximately factorizes as \( L_0(k) L_1(\phi) \), where \( L_0(\cdot) \) and \( L_1(\cdot) \) are two suitable functions. Finally, from the computational point of view, following [?] Section 6, \( L_S(k) \) can be obtained using a MC integration, avoiding hence all the drawbacks of MCMC algorithms. For instance, the posterior distribution
\[ \pi(k|y) \propto \pi(k)L_S(k), \]

based on the reference prior \([?]\)

\[ \pi(k) \propto \int_{-\infty}^{\infty} 2t^2 \varphi(t) \frac{\varphi^2(kt)}{\Phi(kt)} dt, \]

can be obtained through AQI, as explained in the next section.

3 Examples

In this section, we discuss some applications of the integrated likelihood (4) for the shape parameter of the SN distribution. In particular, we show that \( L_S(k) \) is computationally practicable, even for large \( n \), and it allows a direct use to draw inference on \( k \). We focus on the Frontier dataset, available at http://azzalini.stat.unipd.it/SN/frontier.dat, which consists of \( n = 50 \) observations drawn from a SN model, with \( \mu = 0, \sigma = 1 \) and \( k = 5 \). We also simulate other datasets of different sample sizes for \( k > 5 \).

Integrated likelihood. In order to compute (4) when the MLE of \( k \) is infinite, we consider a variation of the importance sampling procedure suggested in [?], Section 6. The procedure steps with their comments are here illustrated.

Step 1. Calculate MLEs of the SN parameters using the central parametrization [Azz85] to estimate mean, standard deviation and asymmetry in \( y \). Then convert central parameters to the direct parameters \( \mu, \sigma \) and \( k \). For example, in the case of Frontier data, we obtain \( \hat{\lambda} = (\hat{\mu}, \hat{\sigma}) = (-0.11, 1.23) \) and \( \hat{k} = 1533.77 \). We acknowledge that, in the case of infinite MLE of \( k \), \( \hat{k} \) is not the MLE of \( k \) and the function \( L_S(k) \) we obtain is not exactly the integrated likelihood that appears in [?]. However, as noted above, differences between a SN model with \( k = \hat{k} \) or \( k = \infty \) are not appreciable and we can consider \( \hat{k} \) as a MLE of \( k \).

Step 2. Let \( se(\hat{\mu}) \) and \( se(\hat{\sigma}) \) denote the standard errors of \( \hat{\mu} \) and \( \hat{\sigma} \), respectively. Simulate the couple \( \phi_j = (\mu_j, \sigma_j) \), for \( j = 1, ..., M \), from the prior distribution \( \pi(\phi) \) implicitly defined by simulating \( \mu_j \sim \varphi((\mu - \hat{\mu})/se(\hat{\mu})) \) and \( \sigma_j \) from a Truncated Normal distribution with mean \( \hat{\sigma} \) and standard deviation \( se(\hat{\sigma}) \). As noted above \( L_S(k) \) is approximately independent of the choice of \( \pi(\phi) \), therefore we simplified the importance sampling by choosing \( \pi(\phi) \) as our sampling distribution.

Step 3. For each \( \phi_j \) we obtain \( \lambda_j(k) \) according to

\[ \lambda_j(k) = \arg \max_{\lambda} E \left[ \ell(k, \lambda) ; \hat{k}, \phi_j \right], \]

where \( E \left[ \ell(k, \lambda) ; \hat{k}, \phi_j \right] \) denotes the expected value of the loglikelihood function \( \ell(k, \lambda) \) with respect to the SN model, with parameters \( \mu_j, \sigma_j \) and \( \hat{k} \). We have

\[ E \left[ \ell(k, \lambda) ; \hat{k}, \phi_j \right] = E \left[ n \log 2 - n \log \sigma + \sum_{i=1}^{n} \log \varphi(Z_i) + \log \Phi(kZ_i) ; \hat{k}, \phi_j \right] = n \left( \log 2 - \log \sigma + E \left[ \log \varphi(Z_i) + \log \Phi(kZ_i) ; \hat{k}, \phi_j \right] \right). \]
The univariate integral involved in the former expectation has been computed using AQI. In this way, we have relative reliable approximation algorithm where the random part is limited to the simulation of $\phi_j$.

Step 4. Calculate
$$\hat{L}(k) = \frac{1}{M} \sum_{j=1}^{M} L(k, \lambda_j).$$

Step 5. Repeat Steps 2 - 4 for $k_1, \ldots, k_N$ obtaining $\hat{L}(k_1), \ldots, \hat{L}(k_N)$.

Step 6. Smoothing splines on $\hat{L}(k_1), \ldots, \hat{L}(k_N)$ are needed in order to achieve a good approximation of $L_S(k)$ by smoothing the variation due to the MC standard errors, whose magnitude do not allow a direct use of the values $\hat{L}(k_1), \ldots, \hat{L}(k_N)$.

Finally, we obtain $\pi(k \mid y)$ using AQI to calculate the normalization constant by evaluating $L_S(k)$ on the approximation obtained in Step 6. Results with $N = 100$ and $M = 500$ need penalizing times with a standard PC, which make, at the moment, prohibitive a comprehensive study of the frequentist properties of the obtained $L_S(k)$. For this reason we calculate $L_S(k)$ in some cases which we think enlighten properties of $L_S(k)$ in the SN model.

In the Frontier data, where the MLE of $k$ is infinite, [?] obtained a monotone $L_H(k)$ which does not allow to draw any concluding inference on $k$, [?] obtain a corrected MLE of $k$ equal to 9.14, and we obtained a similar result being the maximum of $L_S(k)$ close to 10 as shown in Figure 1.

![Fig. 1. $L_S(k)$ for the Frontier data.](image)

For $k = 7$ and $k = 15$ we may still draw good inference with $n = 50$ observations as shown in Figure 2. Note that for $k = 15$ we used only $M = 100$ MC addends in order to show that smoothing splines make a good job in approximating $L_S(k)$.

We finally consider larger samples ($n = 100$) with $k = 30$ and $k = 100$. In these cases, Figure 3 shows that inference with $L_S(k)$ is negative biased because $L_S(k)$ is concentrated around $k = 20$. This indicates that observations contain very weak information in order to distinguish SN models larger than $k > 20$. However, for the
generated datasets the fit of a SN model with \( k \) fixed to \( \arg \max_k L_S(k) \) is enough in order to capture the skewness in the data.

**Fully Bayesian analysis.** We finally consider \( L_S(k) \) for a fully Bayesian analysis for \( k \) for the Frontier data. The default prior \( \pi(k) \) makes a correction of \( L_S(k) \) by concentrating the posterior toward lower values as shown in Figure 1. In fact the posterior mean of \( k \) we obtain is about 8.2, which is the lowest point estimator ever obtained for this dataset. Moreover, the 95% equal tails credible region for \( k \) is (2.1, 19.7) which clearly indicates that the SN model have a positive asymmetry, but it is far from the Half Normal distribution.
4 Final remarks

The shape parameter $k$ of the SN model regulates, but differs from the skewness whose maximum value tends to 1 $(-1)$ as $k \to \infty (-\infty)$. Therefore information contained in observed data will be weak in order to distinguish among large values of $k$. If one wants to account for skewness larger than $|1|$, then other models should be chosen instead of the SN. On this purposes Skew-$t$ distribution [AC03] may be a good alternative (see also [?]), and the method discussed in this paper to eliminate nuisance parameters may also apply to this model. More generally, the method here discussed may also be used in composite scale and regression models, when inference on the shape parameter is of interest and the reference integrated likelihood (2) gives rise to non satisfactory results.

References


