

A note on Bayesian hypothesis testing for the scalar skew-normal distribution

Cinzia Carota and Caterina Conigliani

Abstract The skew-normal distribution is a class of densities that preserves some useful properties of the normal distribution while allowing a shape parameter to account for skewness. It has various remarkable properties in terms of mathematical tractability and turned out to be quite useful in modelling real data. However from an inferential point of view its use gives raise to many difficulties, that are intrinsically tied with the shape of the likelihood function. This fact suggests to solve the problem by calibrating the likelihood with a weight function, and perhaps the most intuitive calibration can be obtained in the Bayesian framework, where the prior distribution plays naturally the role of the weight function. Here we consider in details the problem of testing normality in the general skew-normal model, and solve it by means of different tools for hypothesis testing in the Bayesian framework, namely the Bayes factor and the Jeffreys divergence, pointing out benefits and problems of both approaches

Key words: Bayesian testing, model selection, skew-normal distribution

1 Introduction

The skew-normal (SN) distribution is a class of densities that preserves some useful properties of the normal distribution while extending the normal model by allowing a shape parameter to account for skewness. It appeared several times in the literature (Roberts, 1966; O'Hagan and Leonard, 1976), and was widely studied and given

Cinzia Carota

Dipartimento di Statistica e Matematica Applicata, Università di Torino, via Maria Vittoria 38, 10134, Torino, Italy, e-mail: carota@econ.unito.it

Caterina Conigliani

Dipartimento di Economia, Università Roma Tre, Via Silvio D'Amico, 77, 00145 Roma, Italy, e-mail: caterina.conigliani@eco.uniroma3.it

the present name by Azzalini (1985). Here we consider a specific testing problem in the scalar case; for an overview including inference problems consider for instance Dalla Valle (2004), while for generalizations to the multivariate case see Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (1999).

A random variable Z has a standard skew-normal distribution if its probability density function is given by $f_Z(z|\lambda) = 2\phi(z)\Phi(\lambda z)$, where $z \in (-\infty, \infty)$, ϕ and Φ denote respectively the probability density function and the distribution function of a standard normal, and λ is a real parameter controlling the skewness of the distribution. In particular, if $\lambda = 0$ one gets back to the normal case. If we then add a location parameter μ and a scale parameter σ , the random variable $Y = \mu + \sigma Z$ is said to have a $SN(\mu, \sigma, \lambda)$ distribution, with probability density function given by

$$f_Y(y|\mu, \sigma, \lambda) = \frac{2}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right),$$

where $\mu \in (-\infty, \infty)$, $\sigma > 0$, $\lambda \in (-\infty, \infty)$, and ϕ and Φ are defined as before.

Notice that the SN class of distributions has various remarkable properties in terms of mathematical tractability and turned out to be quite useful in modelling real data. However, as pointed out for instance in Azzalini and Capitanio (1999), in Chiogna (1998) and in Pewsey (2001), from an inferential point of view the use of the SN gives raise to many difficulties that cannot be solved simply by introducing an alternative parametrization of the SN model, or by using different standard estimators such as the one based on the method of moments. These difficulties are intrinsically tied with the shape of the likelihood function, and this fact suggests to solve the problem by calibrating the likelihood with a weight function; work along this line can be found for instance in Sartori (2006), but perhaps the most intuitive calibration can be obtained in the Bayesian framework, where the prior distribution plays naturally the role of the weight function. In this sense, it is particularly interesting to see the papers by Liseo and Loperfido (2006) and Bayes and Branco (2007), that look at a non-informative Bayesian analysis of the SN model.

In particular, Liseo and Loperfido (2006) consider an approach based on reference priors (Bernardo, 1979): the joint reference prior for the SN parameters μ, σ, λ factorises into the usual non-informative prior for location-scale parameters (that effectively play the role of nuisance parameters), *i.e.* $\pi(\mu, \sigma) \propto 1/\sigma$, and a default prior for λ which is proper (although the support of the parameter is unbounded) with tails of order $O(\lambda^{-3/2})$, and this compensates the fact that the likelihood of the SN might not vanish in the tails. The problem of introducing a sensible prior distribution for λ has also been considered in Bayes and Branco (2007). In particular, as the reference prior for λ in Liseo and Loperfido (2006) is not particularly easy to work with, Bayes and Branco (2007) find that a good approximation for it is given by a $t(0, \pi/2; 1/2)$, where $t(a, b; d)$ denotes a Student- t distribution centered in a , with scale b and d degrees of freedom. Moreover, they propose an alternative non-informative Bayesian analysis based on a uniform prior for a reparametrization of λ : if one assumes that $\delta = \lambda/\sqrt{1+\lambda^2} \sim U_{[-1,1]}$, then it is straightforward to prove that $\lambda \sim t(0, \sqrt{1/2}; 2)$; this is clearly not a vague distribution on λ , but Bayes

and Branco (2007) claim that it is objective, in the sense that it was not built using one's prior knowledge about λ .

It is interesting to note that both in Liseo and Loperfido (2006) and in Bayes and Branco (2007), the prior assumptions for λ are tested for different inferential procedures, but only for the standard case. In particular, both papers consider the problem of comparing the $SN(0, 1, 0)$ model against the $SN(0, 1, \lambda)$ model, *i.e.* the problem of testing $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ for the standard skew-normal model, and solve it by means of the Bayes factor, assuming that a Bayes factor less than 0.5 indicates strong evidence against the null model (and conversely a Bayes factor greater than 2 indicates strong evidence in favour of the null model). The results (in terms of a simulation study) were found to be encouraging, although under the uniform prior the Bayes factor seemed to have a tendency to reject the null model more often than other methods, even when it was the correct model.

Here we begin our analysis by considering again the problem of testing $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$, but for the general $SN(\mu, \sigma, \lambda)$ model. Section 2 shows the results of a simulation experiment similar to those of Liseo and Loperfido (2006) and Bayes and Branco (2007), that are not so encouraging when we assume that all three parameters of the SN are unknown. And if instead of fixing (arbitrary) thresholds for the Bayes factor, we assume that there is evidence in favour of the null model $SN(\mu, \sigma, 0)$ when a Bayes factor is greater than 1 and evidence against the null model when a Bayes factor is less than 1, which is the standard practice, the results get even worse, with a disastrous behaviour in the case of uniform prior. In Section 3 we consider a different tool for Bayesian hypothesis testing on the skewness parameter of the SN, which is based on the Jeffreys divergence (Jeffreys, 1948), and its behaviour is compared to that of both the Bayes factor and the classical test of Salvan (1986).

2 A Bayes factor approach

Here we consider the problem of comparing the $SN(\mu, \sigma, 0)$ model against the $SN(\mu, \sigma, \lambda)$ model, *i.e.* the problem of testing $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ for the general SN model, by means of the Bayes factor. Recall that given two models, say M_0 and M_1 , and letting $f_r(\mathbf{y} | \theta_r)$ and $\pi_r(\theta_r)$ be respectively the distribution of the data and the prior distribution of the parameters θ_r under model M_r ($r = 0, 1$), the Bayes factor for M_0 against M_1 is defined as

$$B_{01}(\mathbf{y}) = \frac{\int f_0(\mathbf{y} | \theta_0) \pi_0(\theta_0) d\theta_0}{\int f_1(\mathbf{y} | \theta_1) \pi_1(\theta_1) d\theta_1},$$

and represents the weight of evidence in the data in favour of M_0 against M_1 . Multiplying B_{01} by the prior odds, *i.e.* by the ratio of the prior model probabilities $\pi(M_0)$ and $\pi(M_1)$, we then obtain the posterior odds, *i.e.* the ratio of the poste-

rior model probabilities $\pi(M_0|\mathbf{y})$ and $\pi(M_1|\mathbf{y})$, which represents the key tool for Bayesian model selection.

Now, following what is standard practice, here we assume that there is evidence against the null model when $B_{01}(\mathbf{y}) < 1$ (and conversely there is evidence in favour of the null model when $B_{01}(\mathbf{y}) > 1$); this is equivalent to assuming equal prior model probabilities for the null and the alternative models, so that the Bayes factor is equal to the posterior odds. However, in order to compare our results with those of Liseo and Loperfido (2006) and Bayes and Branco (2007), we will also look at some sort of *strong evidence* against the null model, *i.e.* at the proportion of the times $B_{01}(\mathbf{y}) < 0.5$; it is important to stress that this is equivalent to assuming that the prior odds are equal to $\pi(M_0)/\pi(M_1) = 2$, and that any other possible threshold could be employed in the same way (and therefore in our opinion is rather arbitrary): if we believed that a priori M_0 is k times more likely than model M_1 , then we should reject the null model when $B_{01}(\mathbf{y}) < 1/k$.

As in Liseo and Loperfido (2006) and Bayes and Branco (2007), here the behaviour of the Bayes factor for choosing between the particular models under comparison is analysed in a simulation experiment. In particular, we simulated a large number of data sets of size n from a $SN(0, 1, \lambda^*)$ distribution, for given n and λ^* , and computed the proportion of the times the Bayes factor rejected the null model $SN(\mu, \sigma, 0)$. The results are shown in Table 1 for both the Jeffreys prior for λ and the uniform prior on δ , and are directly comparable with those of Table 1 in Liseo and Loperfido (2006) and with those of Table 8 in Bayes and Branco (2007).

Table 1 Proportion of the times $H_0 : \lambda = 0$ is rejected by the Bayes Factor when simulating 1000 data sets of size n from a $SN(0, 1, \lambda^*)$

n	λ^*	BF ₀₁ (Jeffreys prior)		BF ₀₁ (uniform prior)	
		evidence	strong evidence	evidence	strong evidence
10	0	0.92	0.14	1.00	0.02
	0.5	0.91	0.11	1.00	0.01
	1	0.92	0.14	1.00	0.01
	5	0.97	0.26	1.00	0.05
	10	0.97	0.36	1.00	0.08
50	0	0.27	0.07	0.99	0.09
	0.5	0.29	0.07	1.00	0.11
	1	0.28	0.09	1.00	0.13
	5	0.95	0.81	0.99	0.81
	10	0.99	0.94	0.99	0.92
100	0	0.16	0.04	1.00	0.07
	0.5	0.16	0.05	0.99	0.08
	1	0.23	0.09	1.00	0.11
	5	0.99	0.95	0.99	0.98
	10	1.00	1.00	1.00	0.99

First, if we look at the *strong evidence* against the null hypothesis, the fact that we are introducing uncertainty also with respect to μ and σ implies that the behaviour

of the Bayes factor is a lot less encouraging than in Liseo and Loperfido (2006) and Bayes and Branco (2007). In particular, and this is obvious, we see that here we need larger sample sizes than in the standard case: for instance 10 observations are not enough for an effective hypothesis testing on λ when all three parameters of the SN are unknown. Moreover, and this was also expected, even when n is large enough both under the Jeffreys and the uniform prior the power of the test is lower than in the standard case, with the proportion of correct rejections for instance for $n = 100$ reaching values close to 1 only for $\lambda^* = 10$ (as opposed to $\lambda^* = 1$ when μ and σ are known). It is also interesting to note that if we focus the attention on the evidence *in favour* of the null hypothesis, rather than *against* the null, then when μ and σ are also unknown the Bayes factor is hardly ever greater than 2, so that even when $\lambda^* = 0$ the evidence in favour of the normal model is never *strong* (in the sense advocated by Liseo and Loperfido, 2006, and Bayes and Branco, 2007).

If we now look at the proportion of the times the $SN(\mu, \sigma, 0)$ model is rejected for $B_{01}(\mathbf{y})$ being less than 1, *i.e.* at the standard evidence against the null hypothesis, the main point arising from Table 1 is that the behaviour of the Bayes factor under the uniform prior is disastrous, in that it always rejects the null, no matter what the true generating process is, and for all values of n ; this tendency has been already noted in Bayes and Branco (2007) for the standard case, but here it becomes really extreme. Instead, when we assume the Jeffreys prior and n is large enough, the behaviour of the Bayes factor is quite reasonable, except for the percentage of (wrong) rejections when $\lambda^* = 0$, that is definitely too high; it clearly improves as n gets bigger, but we found that at least 200 observations were required in order for this percentage to reach the nominal value 0.05.

Summarising the results of Table 1, the Bayes factor behaves well enough (although obviously not as well as in the standard case) under both priors if n is large enough and if, following Liseo and Loperfido (2006) and Bayes and Branco (2007), we look at some (arbitrary) concept of *strong evidence* against the null. Instead, if we refer to the standard notion of evidence against an hypothesis, then the behaviour of the Bayes factor in this context is clearly not very encouraging, especially as a tool for local testing, so that it is interesting to explore the performance of different tools for Bayesian hypothesis testing.

3 A divergence-based Bayesian approach

An alternative to the Bayes Factor for testing for normality in the Azzalini skew-normal distribution class is an approximate Bayesian test based on the Jeffreys divergence derived in Carota (2010). Here we are going to compare the performance of the Bayes factor considered in the previous section to that of this alternative testing tool, assuming again that μ and σ are nuisance parameters, distributed according to the non-informative prior $\pi(\mu, \sigma) \propto 1/\sigma$, and assuming an informative prior $\pi(\lambda)$ for the skewness parameter λ . If $\pi(\lambda)$ is taken to be sufficiently concentrated around the null value $\lambda = 0$, then it is possible to test the null hypothesis of

normality $H_0 : \lambda = 0$ versus the alternative $H_1 : \lambda \neq 0$ by computing a measure of divergence between $\pi(\lambda)$ and the corresponding posterior distribution $\pi(\lambda|\mathbf{y})$. The idea is that by means of the peakedness of $\pi(\lambda)$ around $\lambda = 0$ we express strong prior beliefs in the normal model and the null hypothesis of normality is confirmed by the data if the posterior distribution of λ turns out to be close to the prior, *i.e.* if a suitable measure of the divergence between $\pi(\lambda|\mathbf{y})$ and $\pi(\lambda)$ is sufficiently small. In particular, we adopt the Jeffreys divergence

$$J(\pi(\lambda|\mathbf{y}) : \pi(\lambda)) = \frac{1}{2} \left\{ \int_{-\infty}^{\infty} \pi(\lambda|\mathbf{y}) \log \left(\frac{\pi(\lambda|\mathbf{y})}{\pi(\lambda)} \right) d\lambda + \int_{-\infty}^{\infty} \pi(\lambda) \log \left(\frac{\pi(\lambda)}{\pi(\lambda|\mathbf{y})} \right) d\lambda \right\}$$

and the calibration suggested in McCulloch (1989). This is based on the comparison of two Bernoulli distributions with parameters 0.5 and p respectively, and here we assume that sample evidence in favour of H_0 is represented by $J(\pi(\lambda|\mathbf{y}) : \pi(\lambda))$ being less than 0.44 = $J(\text{Bernoulli}(0.5) : \text{Bernoulli}(p = 0.9))$. The meaning of this calibration in our testing problem can be described as following: we reject H_0 for $J \geq 0.44$ since predicting outcomes with the prior of λ (very concentrated around $\lambda = 0$), rather than with its posterior, is comparable to describing an unobserved event as having probability greater or equal to 0.9, when in fact the probability is 0.5. Moreover, by expanding the logarithmic function around the null value $\lambda = 0$, we approximate $J(\pi(\lambda|\mathbf{y}) : \pi(\lambda))$ by

$$J_T(\pi(\lambda|\mathbf{y}) : \pi(\lambda)) = \frac{1}{2} \left[E(\lambda^3|\mathbf{y}) - E(\lambda^3) \right] I_S$$

where I_S denotes the locally most powerful invariant test for $H_0 : \lambda = 0$ against the one-sided alternative $H_1 : \lambda > 0$ provided by Salvan (1986) (see also Azzalini, 2006). Notice that J_T represents the first non-zero term in the Taylor expansion of J , and that I_S is proportional to the sample skewness.

It is also important to note that in the divergence-based approach the role of the prior distribution for λ is significantly different from that considered in the approach of Section 2. In fact, as we already pointed out, the Bayes factor represents a tool for choosing between two competing models, in our case a $N(\mu, \sigma)$ and a $SN(\mu, \sigma, \lambda)$, so that $\pi(\lambda)$ represents one's prior beliefs on the skewness parameter under the alternative (in fact, $\pi(\lambda)$ could be any distribution on the real line, could be skew, could be multimodal, could represents strong prior beliefs about λ , or could be non-informative). Instead, under this divergence-based approach there is essentially one model under consideration, the $SN(\mu, \sigma, \lambda)$, while the model of interest, the one that we would like to adopt, is the normal one. In this sense, the peakedness of $\pi(\lambda)$ around $\lambda = 0$ under the $SN(\mu, \sigma, \lambda)$ is really part of the definition of model of interest, and the alternative SN model is just an elaboration of the model of interest used as a tool for model criticism that can handle surprising experimental results revealing departures from normality in the direction of skewness. For these reasons the only relevant aspect in the elicitation of the prior is the peakedness of $\pi(\lambda)$, which has to be commensurated with the sample size n in order for the comparison with the posterior to make sense. For instance if, for purposes of comparison with the prior

distributions assumed in Section 2, we assume a prior of the form $t(0, \sqrt{1/2a}; 2a)$ (for $a = 1$ this corresponds to the uniform prior on δ), then it is the coefficient a that controls the peakedness of $\pi(\lambda)$ around zero, and the strong relationship between n and a can be observed in Table 2 that shows the proportion of the times $H_0 : \lambda = 0$ is rejected by J_T (under this prior) when simulating 1000 data sets of size n from a $SN(0, 1, 0)$. In particular, the results in Table 2 point out that the value of a that leads to a proportion of (wrong) rejections close to the nominal value 0.05 varies with n , with larger values of n needing larger values of a . (Correspondingly the prior probability $Pr(|\lambda| < 0.5)$ increases from 0.45, for $a = 1$, to 0.85, for $a=5$.)

Table 2 Proportion of the times $H_0 : \lambda = 0$ is rejected by J_T under the $t(0, \sqrt{1/2a}; 2a)$ prior for λ when simulating 1000 data sets of size n from a $SN(0, 1, 0)$

n	a = 1	a = 2	a = 3	a = 4	a = 5
30	0.75	0.36	0.09	0.00	0.00
50	0.67	0.37	0.12	0.03	0.01
100	0.66	0.44	0.24	0.10	0.02
200	0.62	0.49	0.34	0.16	0.07

We can then look at the power of the test, and compare the results obtained with the Jeffrey's divergence to those obtained with other methods that are commonly employed in this context. In particular, Table 3 shows the proportion of the times the hypothesis of normality was rejected by J_T , by the Salvan test for normality and by the Bayes factor BF_{01} when simulating a large number of data sets of size $n = 50$ and $n = 100$ from a $SN(0, 1, \lambda^*)$; note that for computing both J_T and the Bayes factor we assumed the $t(0, \sqrt{1/2a}; 2a)$ prior for λ with $a = 3.5$ when $n = 50$, and $a = 5$ when $n = 100$.

Table 3 Proportion of the times $H_0 : \lambda = 0$ is rejected by I_S , BF_{01} and J_T under the $t(0, \sqrt{1/2a}; 2a)$ prior for λ , when simulating 1000 data sets of size n from a $SN(0, 1, \lambda^*)$

n	a		$\lambda^* = 0$	$\lambda^* = 0.5$	$\lambda^* = 1$	$\lambda^* = 5$	$\lambda^* = 10$
50	3.5	I_S	0.03	0.04	0.05	0.52	0.64
		J_T	0.06	0.08	0.10	0.70	0.82
		BF_{01}	0.94	0.95	0.95	0.95	0.96
100	5	I_S	0.04	0.04	0.06	0.89	0.95
		J_T	0.02	0.02	0.05	0.87	0.93
		BF_{01}	0.93	0.95	0.95	0.96	0.98

Notice, from Table 3, that while J_T has a behaviour comparable to that of I_S , the Bayes factor nearly always rejects the hypothesis of normality, even when $\lambda^* = 0$, similarly to what we noticed under the assumption of a uniform prior on δ in Section 2. We believe that the reason for this unusual behaviour is that the prior distributions that we are assuming for λ are very concentrated around zero in order for J_T to

work, but this represents a problem from the Bayes factor's point of view, since such peaked priors make the alternative $SN(\mu, \sigma, \lambda)$ model too close to the normal one. Then it is natural that when the Bayes factor is asked to choose between a normal model and a model which is just about normal, only possibly slightly asymmetric, it favours the alternative model since no simulated data sets are really symmetric. This seems to indicate that also the uniform prior on δ introduced by Bayes and Branco (2007), that as we already pointed out is not really non-informative on λ , is too concentrated on $\lambda = 0$ for the Bayes factor to work well, at least when the location and scale parameters of the skew-normal model are also unknown.

Summarising, the Bayes factor seems to work well provided the sample size is large enough and the prior on λ is sufficiently vague. Instead, J_T can be applied with smaller sample sizes, but the peakedness of the prior needs to be roughly commensurated with the sample size.

References

1. Azzalini, A.: A class of distributions which includes the normal ones. *Scand J Stat* **12**, 171-178 (1985)
2. Azzalini, A.: Normality, Salvati test for. In: *Encyclopedia of Statistical Science*. John Wiley and Sons, Inc. (2006)
3. Azzalini, A., Capitanio, A.: Statistical applications of the multivariate skew-normal distributions. *J R Stat Soc, Series B* **61**, 579-602 (1999)
4. Azzalini, A., Dalla Valle, A.: The multivariate skew-normal distributions. *Biometrika* **83**, 715-726 (1996)
5. Bayes, C.L., Branco D'Elia, M.: Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian J Prob Stat* **21**, 141-163 (2007)
6. Bernardo, J.M.: Reference posterior distributions for Bayes inference. *J R Stat Soc, Series B* **41**, 113-147 (1979)
7. Carota, C.: Tests for normality in classes of skew-t alternatives. *Stat Prob Lett* **80**, 1-8 (2010)
8. Chiogna, M.: Some results on the scalar skew-normal distribution. *J It Stat Soc* **7**, 1-13 (1998)
9. Dalla Valle, A.: The skew-normal distribution. In: Genton, M.G. (Ed.) *Skew-Elliptical Distributions and their Applications: a Journey beyond Normality*, pp.3-24. Chapman and Hall/CRC, Boca Raton, FL (2004).
10. Jeffreys, H.: *Theory of Probability*. Clarendon Press, Oxford (1948)
11. Liseo, B.: La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. *Statistica* **50**, 59-70 (1990)
12. Liseo, B., Loperfido, N.: A note on reference priors for the scalar skew-normal distribution. *J Stat Plan Inference* **136**, 373-389 (2006)
13. O'Hagan, A., Leonard, T.: Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201-202 (1976)
14. McCulloch, R.E.: Local model influence. *J Am Stat Ass* **84**, 473-478 (1989)
15. Pewsey, A.: Problems of inference for Azzalini's skew-normal distribution. *J Appl Stat* **27**, 859-870 (2001)
16. Roberts, C.: A correlation model useful in the study of twins. *J Am Stat Ass* **61**, 1184-1190 (1966)
17. Salvati, A.: Test localmente più potenti tra gli invarianti per la verifica dell'ipotesi di normalità. In: *Società Italiana di Statistica (Ed.) Atti della XXXIII Riunione della Società Italiana di Statistica* 2, pp.173-179. Cacucci, Bari (1986).
18. Sartori, N.: Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J Stat Plan Inference* **136**, 4259-4275 (2006).