Christophe Ley and Davy Paindaveine

Université libre de Bruxelles, ECARES and Département de Mathématique
Campus Plaine, CP210, Boulevard du Triomphe, B-1050, Brussels, Belgium
chrisley@ulb.ac.be, dpaindav@ulb.ac.be

First of all, we would like to congratulate the author for an extremely interesting and insightful paper. In times where the need for flexible models has become obvious in view of the non-normal nature of most real data sets (be it in finance, biosciences, engineering, climatology, etc.) and where, consequently, more and more "new" distributions are popping up in the literature, a survey paper reviewing the most important approaches is certainly highly welcomed. For researchers and — perhaps even more importantly — for practitioners, a comparison between these approaches, as undertaken by the author, is crucial in order to be able to decide which flexible family to use in which situation. The author's contribution is an essential first step, that should later be complemented with similar comparisons in the multivariate case and with practical guidelines on how to choose the best family for a given data set.

In this discussion, we shall comment on the 10 criteria chosen by the author to compare Families 1-4 (Section 1), present a new criterion based on fixing the median of the initial distribution (Section 2), discuss, in a generic family, the relevance of parameter orthogonality when testing for symmetry (Section 3), and conclude by expressing the need for goodness-of-fit tests allowing to decide which family best fits the data set at hand (Section 4).

1 On the ten criteria to compare the four flexible families

For the sake of readability, we will follow the author's structure and grouping of criteria in our discussion. We write Ci for the i^{th} criterion, $i = 1, \ldots, 10$.

C1-2: In our view, it seems unnatural that bimodality is obtained when applying a skewing mechanism to a symmetric unimodal density, particularly so in the vicinity of symmetry. We thus agree that unimodality is an important requirement. The property of allowing for an explicit expression of the mode is slightly less important in comparison to C1, but of course happens to be a nice side-product in Family 3, where the mode is fixed. This has led us to propose, in Section 2, a new criterion based on fixing the median of the original density g.

C3: Skewness ordering allows to increase interpretability of the skewness parameter, hence is certainly a desirable property. In the same order of ideas, it would be highly interesting to measure the skewness and kurtosis that each flexible family does allow for, e.g. via the AG-skewness of Arnold & Groeneveld (1995) or skewness-invariant measures of kurtosis as proposed in Jones *et al.* (2011). This, however, is beyond the scope of our discussion, and we therefore leave this as an open problem.

C4: This criterion is somehow striking, as it underlines the fact that, despite the announcement of shape families, the author has mainly presented skewness families (which by no means is meant to be a criticism!). We point out that explicit links between highest finite moments of the original symmetric density g and those of the corresponding skewed densities f_g have been worked out in Ley & Paindaveine (2010a, Theorem 3.2) and in Ferreira & Steel (2006, Theorems 3-6) for Family 2 and 4, respectively.

C5-7: Tractability certainly is an important issue, too, but mostly for researchers, as practitioners will not so much care about these criteria. C7 happens to be very helpful for the inferential purposes we discuss in Section 3.

C8: We attach particular importance to this criterion, as flexible modelling is useless without good parameter estimation, mostly via maximum likelihood estimation. We tend to agree with the author that the possibility of having an infinite value of the skewness parameter in Family 1 is not a very severe issue, although the remedy of Azzalini & Arellano-Valle (2013) in form of a penalized maximum likelihood estimation presents a nice way-out of the problem. We also seize the opportunity to point out a further notorious problem of Family 1, more precisely, of the skew-normal distribution: the profile log-likelihood function for skewness admits a stationary point in the vicinity of symmetry. This issue has been treated in detail in Ley & Paindaveine (2010b, c), where it is shown that this undesirable feature fortunately is not present in other multivariate skew-symmetric distributions of the form $2g(\boldsymbol{x}-\boldsymbol{\mu})w(\boldsymbol{\delta}'(\boldsymbol{x}-\boldsymbol{\mu}))$.

C9: Parameter orthogonality, and the ensuing simple structure of the Fisher information matrix, is essential when it comes to hypothesis testing in the presence of nuisance parameters. We discuss this matter in Section 3.

C10: Nice as they are, "physical" mechanisms, in our opinion, are not so relevant to compare flexible models; they are rather a bonus that may provide such models a certain justification in given situations, but not more.

2 A new criterion based on median-invariance

Family 3 is the only one where the unique mode is explicitly available. More specifically, conditions for mode-invariance have been put forward (see Fujisawa & Abe 2014), from which it can be deduced that Family 3A is mode-invariant (a mode-preserving property for Family 4 can be found in Ferreira & Steel 2006, Theorem 2). This mode-invariance property is a nice feature, which, e.g., greatly simplifies the AG-skewness calculation. Such invariance is desirable since skewness and location are of a different — in some sense, even "orthogonal" — nature. It then seems natural to impose that location remains unchanged when introducing skewness.

The mode is, however, only one of the many location functionals that may be used to express this orthogonality between location and skewness. Another obvious choice is the mean, which has the disadvantage to require moment conditions (a particularly unpleasant feature when flexible modelling leads to considering heavy tails). We tend to think that the median is an appropriate choice. First, it is, in the absolutely continuous context considered, always unambiguously defined, irrespective of any moment or unimodality assumptions. Second, when considered as location functionals over the space of densities, the median, unlike the mode, is continuous with respect to the $\mathcal{L}_\infty$-norm, which is a desirable condition when aiming at efficient inference about location. This leads us to add a further criterion to the list provided by the author:

C11: Median-invariance.

Families 1 and 3, except for very peculiar cases, do not enjoy median-invariance (for the sake of simplicity, we speak in what follows of "fixing the median"). In the case of skew-symmetric densities of the form $2g(x)w(x)$, it is easily seen that C11 fails to hold when w is monotone increasing, hence in particular when $w(x) = D(\lambda x)$ for some cumulative distribution function (cdf) D over $[0, 1]$. As regards Family 3, Theorem 4.2 of Fujisawa & Abe (2014) establishes how the median changes with the chosen transformation of scale; the median-variation is especially obvious in Family 3A.

Regarding Family 2, we have derived in Ley & Paindaveine (2010a, Theorem 3.1) a very simple condition for checking whether $g(W^{-1}(x))/w(W^{-1}(x))$ fixes the median: the median is fixed iff $W(0) = 0$. This condition is for instance satisfied by Tukey's g-and-h transformation, but it is violated by the sinh-arcsinh transformation of Jones & Pewsey (2009).

The convenient form of the distribution function $w(G(\cdot))$ in Family 4 leads to a straightforward condition for median-invariance: the median is fixed iff $w(1/2) = 1/2$. Except for particular choices of the parameter values, this condition is not fulfilled when w is the beta cdf; the proportional odds choice inspired by Marshall & Olkin (1997) also violates this condition. Of course, what happens for all other models belonging to Family 4 can be investigated in a similar way.

Thus, to conclude, C11 is a natural requirement, that provides an alternative to the less stable mode-invariance. While Condition C11 is not satisfied by Families 1 and 3, it may hold for Families 2 and 4 (and extremely simple conditions to check whether or not C11 holds there are available).

3 Parameter orthogonality and inference

Let $X_1, \ldots, X_n$ be independent univariate observations with a common four-parameter density $f_{\mu,\sigma,\delta,\eta}(\cdot)$, where $\mu(\in \mathbb{R})$, $\sigma(> 0)$, $\delta(\in \mathbb{R})$, and $\eta(> 0)$ are location, scale, skewness, and tail-weight parameters, respectively. Suppose we are interested in testing $\mathcal{H}_0 : \delta = \delta_0$ for some $\delta_0 \in \mathbb{R}$ versus $\mathcal{H}_1 : \delta \neq \delta_0$, all other parameters remaining unspecified (sometimes the location parameter can be fixed, but fixing scale and tail-weight parameters would be highly unrealistic). Efficient testing procedures are based on the score function

$$\boldsymbol{\ell}(x) = \begin{pmatrix} \ell_\mu(x) \\ \ell_\sigma(x) \\ \ell_{\delta_0}(x) \\ \ell_\eta(x) \end{pmatrix} := \begin{pmatrix} \frac{\partial_\mu f_{\mu,\sigma,\delta_0,\eta}(x)}{f_{\mu,\sigma,\delta_0,\eta}(x)} \\ \frac{\partial_\sigma f_{\mu,\sigma,\delta_0,\eta}(x)}{f_{\mu,\sigma,\delta_0,\eta}(x)} \\ \frac{\partial_\delta f_{\mu,\sigma,\delta,\eta}(x)|_{\delta=\delta_0}}{f_{\mu,\sigma,\delta_0,\eta}(x)} \\ \frac{\partial_\eta f_{\mu,\sigma,\delta_0,\eta}(x)}{f_{\mu,\sigma,\delta_0,\eta}(x)} \end{pmatrix}$$

and the corresponding Fisher information matrix $\boldsymbol{\Gamma} := \int_{\mathbb{R}} \boldsymbol{\ell}(x)\boldsymbol{\ell}'(x) f_{\mu,\sigma,\delta_0,\eta}(x)dx$. This is the case for, e.g., the Rao score (or Lagrange multiplier) test and for the tests obtained from Le Cam's theory of asymptotic experiments.

If the information matrix $\boldsymbol{\Gamma} = (\Gamma_{ij})$ is block-diagonal with respect to δ (that is, if $\Gamma_{3j} = 0$ for $j = 1, 2, 4$), then an optimal test at asymptotic level α is the test that rejects the null hypothesis whenever

$$T^{(n)} := \frac{|n^{-1/2} \sum_{i=1}^n \ell_{\delta_0}(X_i)|}{\Gamma_{33}^{1/2}} > \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big), \tag{3.1}$$

where Φ stands for the cdf of the standard normal distribution and where root-n consistent (and, in principle, locally asymptotically discrete) estimators were substituted for μ, σ, and η. Under the assumed block-diagonality, this substitution has no effect on the null asymptotic behavior of $T^{(n)}$ (hence, from contiguity, also on its asymptotic behavior under local alternatives of the form $\mathcal{H}_1^{(n)} : \delta = \delta_0 + n^{-1/2}\tau$). The non-specification of location, scale, and tail-weight then has no "cost" in terms of asymptotic (local) power.

If the above block-diagonality does not hold, then estimating nuisance parameters will have a cost in terms of asymptotic power. Worse: the test in (3.1) will not have null asymptotic size α, since the test statistic will not be insensitive (in the asymptotic sense described above) to estimation of nuisance parameters. A suitable test statistic may be obtained by replacing the score function for skewness $\ell_{\delta_0}(x)$ with

$$\ell^*_{\delta_0}(x) := \ell_{\delta_0}(x) - (\Gamma_{13}, \Gamma_{23}, \Gamma_{34}) \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{14} \\ \Gamma_{12} & \Gamma_{22} & \Gamma_{24} \\ \Gamma_{14} & \Gamma_{24} & \Gamma_{44} \end{pmatrix}^{-1} \begin{pmatrix} \ell_\mu(x) \\ \ell_\sigma(x) \\ \ell_\eta(x) \end{pmatrix},$$

that is, with the orthogonal projection of ℓ_{δ_0} onto the subspace orthogonal to the other score functions (in the metric associated with $\mathbf{\Gamma}$). The new score $\ell^*_{\delta_0}$ is called the *efficient score function*. The corresponding efficient information for skewness is given by

$$\Gamma^*_{33} := \Gamma_{33} - (\Gamma_{13}, \Gamma_{23}, \Gamma_{34}) \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{14} \\ \Gamma_{12} & \Gamma_{22} & \Gamma_{24} \\ \Gamma_{14} & \Gamma_{24} & \Gamma_{44} \end{pmatrix}^{-1} \begin{pmatrix} \Gamma_{13} \\ \Gamma_{23} \\ \Gamma_{34} \end{pmatrix}.$$

The encurred loss of power can be characterized by the difference $\Gamma_{33} - \Gamma^*_{33}$. As we have seen, no loss of power is obtained when ℓ_{δ_0} is "orthogonal" to the other scores (still in the metric associated with $\mathbf{\Gamma}$). The further we move away from orthogonality, the larger that loss. In the extreme case where ℓ_{δ_0} is a linear combination of ℓ_μ, ℓ_σ and ℓ_η, both the efficient score $\ell^*_{\delta_0}$ and the efficient information Γ^*_{33} are equal to zero, implying that the optimal test for testing $\delta = \delta_0$ against $\delta \neq \delta_0$ coincides with the trivial α-test (that is, the test that rejects the null whenever an independent Bernoulli random variable with parameter α takes on value 1). In other words, the performance would be so poor that the most efficient test has asymptotic power equal to the nominal level α.

This, in our opinion, underlines the importance of parameter-orthogonality, and hence of Criterion C9. A famous example falling into the category of extreme confounding is the skew-normal distribution when testing for symmetry about an unspecified location. Further details about this issue in skew-symmetric families are stated in Ley & Paindaveine (2010b, Section 3) or Ley & Paindaveine (2010c, Section 5).

In the case of testing for symmetry about a specified location, parameter orthogonality in Family 4 has been put to use in Ley & Paindaveine (2009) where efficient tests under unknown scale have been derived, whereas Cassart *et al.* (2008) took advantage of the parameter-orthogonality structure in Family 3A to build optimal tests for symmetry about an unknown center against Fechner-type alternatives (see the author's Equation (7)).

Summing up, parameter-orthogonality is a clear plus for Family 3 (particularly 3A), especially compared to Family 1. On the other hand, one has to acknowledge that the Fisher information singularity issue has become part of the fame of skew-symmetric models, and has been much discussed in the literature. In particular, alternative parameterizations have been proposed to remove this singularity, such as the centred parameterization of

Azzalini (1985) as well as the reparameterization given in Hallin & Ley (2014) (see also Rotnitzky *et al.* 2000 for a general discussion about such singularities).

4 Which family fits the data at hand best?

The comparison proposed by the author is an important step in order to provide information on the pros and cons of each construction. One point, that is of paramount importance for practitioners, is missing, though: given a certain data set, which construction is most appropriate?

The problem is obviously difficult, as testing that a given family is suitable for the data at hand involves many nuisance parameters, the most severe being the original symmetric density g. In some families, it is possible to estimate g relatively easily. For instance, in Family 1, a kernel density estimator of g may be based on the fact that, if X has density $x \mapsto f_g(x)$, then $|X|$ has density $x \mapsto 2g(x)I_{[x>0]}$. However, knowing g (or being able to estimate g in a natural way) does not imply that one can test that a given family is suitable. In particular, despite the above possibility to estimate g, it seems to be tricky to test the null that the underlying density belongs to Family 1.

This suggests an additional criterion, call it

C12: Testability,

that requires that natural or satisfactory goodness-of-fit tests for the considered family can be defined. It is beyond the scope of the present discussion to investigate whether or not the various families identified by the author do satisfy C12. Here, we only settle the case of the Fechner family — that is, of the subfamily of Family 3 described by the author's Equation (7) — by showing that this family satisfies C12.

The Fechner family is characterized by the fact that the conditional distribution given that the observation is positive only differs by a scale factor from the conditional distribution given that the observation is negative. This suggests performing a symmetry test on observations that have been transformed to correct for this scale factor. To be more specific, a possible test that the underlying distribution is of the Fechner type rejects the null for large values of

$$Q^{(n)}(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} \Big\{ F_n(x) + F_n(-x) - 1 \Big\}^2 \, dF_n(x),$$

where, denoting by Med_+ and Med_- the median of the positive X_i's and minus the median of the negative X_i's, $F_n(\cdot)$ stands for the cdf of the transformed observations

$$Y_i = \begin{cases} X_i/\mathrm{Med}_+ & \text{if } X_i \geq 0 \\ X_i/\mathrm{Med}_- & \text{if } X_i < 0. \end{cases}$$

This test statistic is inspired by the test for symmetry of Rothman & Woodroofe (1972). It is not clear how to extend such a test to the whole of Family 3.

References

[1] Arnold, B.C. & Groeneveld, R.A. (1995) Measuring skewness with respect to the mode. *Amer. Statist.* **49**, 34–38.

[2] Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171–178.

[3] Azzalini, A. & Arellano-Valle, R.B. (2013) Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *J. Statist. Plann. Infer.* **143**, 419–433.

[4] Cassart, D., Hallin, M. & Paindaveine, D. (2008) Optimal detection of Fechner-asymmetry. *J. Statist. Plann. Infer.* **138**, 2499–2525.

[5] Ferreira, J.T.A.S. & Steel, M.F.J. (2006) A constructive representation of univariate skewed distributions. *J. Amer. Statist. Assoc.* **101**, 823–829.

[6] Fujisawa, H. & Abe, T. (2014) A family of multivariate skew distributions with monotonicity of skewness. Manuscript.

[7] Hallin, M. & Ley, C. (2014) Skew-symmetric distributions and Fisher information: the double sin of the skew-normal. *Bernoulli* **20**, 1432–1453.

[8] Jones, M.C. & Pewsey, A. (2009) Sinh-arcsinh distributions. *Biometrika* **96**, 761–780.

[9] Jones, M.C., Rosco, J.F. & Pewsey, A. (2011) Skewness-invariant measures of kurtosis. *Amer. Statist.* **65**, 89–95.

[10] Ley, C. & Paindaveine, D. (2009) Le Cam optimal tests for symmetry against Ferreira and Steel's general skewed distributions. *J. Nonparametr. Statist.* **21**, 943–967.

[11] Ley, C. & Paindaveine, D. (2010a) Multivariate skewing mechanisms: a unified perspective based on the transformation approach. *Stat. Probab. Lett.* **80**, 1685–1694.

[12] Ley, C. & Paindaveine, D. (2010b) On the singularity of multivariate skew-symmetric models. *J. Multivariate Anal.* **101**, 1434–1444.

[13] Ley, C. & Paindaveine, D. (2010c) On Fisher information matrices and profile log-likelihood functions in generalized skew-elliptical models. *Metron* **68**, Special Issue on "Skew-symmetric and flexible distributions", 235–250.

[14] Marshall, A.W. & Olkin, I. (1997) A new method of adding a parameter to a family of distributions with applications to the exponential and Weibull families. *Biometrika* **84**, 641–652.

[15] Rothman, E.D. & Woodroofe, M. (1972) A Cramér von-Mises type statistic for testing symmetry. *Ann. Math. Statist.* **43**, 2035–2038.

[16] Rotnitzky, A., Cox, D.R., Bottai, M. & Roberts, J. (2000) Likelihood-based inference with singular information matrix. *Bernoulli* **6**, 243–284.