Modulated symmetry in non-standard settings

Adelchi Azzalini

Department of Statistical Sciences, University of Padua, Italy

1 Standard setting

Much work in recent yeas has dealt with with the following construction, which we refer to as the 'standard setting' in its basic form. If $f_0(x)$ denotes a d-dimensional density centrally symmetric about 0, that is, $f_0(x) = f_0(-x)$, and $G(x)$ denotes a real-valued function such that

$$G(x) \geq 0, \qquad G(x) + G(-x) = 1 \tag{1}$$

then

$$f(x) = 2\, f_0(x)\, G(x)\,, \qquad x \in \mathbb{R}^d, \tag{2}$$

is a density function. Here a symmetric *base* density function is *perturbed* or *modulated* by the factor $G(x)$ which can be chosen quite freely, since condition (1) is not restrictive; it is quite remarkable that the normalizing constant is always 2.

To construct a function G which fulfils (1) it is convenient to express it in the form

$$G(x) = G_0\{w(x)\} \tag{3}$$

where G_0 is a univariate distribution function of a continuous variable symmetric about 0 and $w(-x) = -w(x)$ for all $x \in \mathbb{R}^d$. It can be shown that each function (3) is of type (1) and that each G of type (1) can be written is form (3). The latter representation is not unique; however, the set of densities generated by these two forms are coincident.

A prominent example of this construction is the multivariate skew-normal distribution, whose density in the case without location and scale parameters is

$$2\, \phi_d(x; \bar{\Omega})\, \Phi(\alpha^\top x), \qquad x \in \mathbb{R}^d, \tag{4}$$

where $\bar{\Omega}$ is a positive-definite correlation matrix and α is a d-vector of shape parameters. This family of distributions, once supplemented with location and scale parameters, enjoys a number of formal properties which justify to consider it a convincing extension of the normal family.

An important aspect of formulation (2)–(3) is the existence of the following two stochastic representations. If T is a univariate random variable with distribution function G_0 and Z_0 is an independent d-dimensional variable with density f_0, then both

$$Z' = (Z_0 | T \leq w(Z_0)), \qquad Z'' = \mathrm{sign}(T - w(Z_0))\, Z_0 \tag{5}$$

have density f. A corollary of the second representation is that a variable Z with density function f satisfies

$$t(Z) \stackrel{d}{=} t(Z_0) \tag{6}$$

for any even q-dimensional function $t(\cdot)$. For instance, if Z has distribution (4), then $Z^\top \bar{\Omega}^{-1} Z \sim \chi^2_d$.

Clearly, there are very many other density functions besides (4) which can be obtained from (2), usually obtained in combination with (3). Moreover, the basic form (2) can be extended

further to even more general constructions. These other extensions, however, retain certain characteristics of the original construction, in one way or another. A systematic fact is to start from a symmetric base density function f_0, to which a modulation action is applied. In the remaining pages, we summarize various explorations in other directions, where f_0 is not a symmetric density. A number of these other constructions pertain to non-Euclidean spaces.

2 Distributions on the simplex

Compositional data arise when we record the proportions of constituents of certain specified types to form a whole. A typical example is represented by the geochemical composition of rocks or other material, such as the proportions of sand, silt and clay in sediments. If there are D different constituents, each observed unit produces a D-part composition represented by the proportions $p = (p_1, p_2, \ldots, p_D)$ such that

$$p_1 > 0, \ldots, p_D > 0, \qquad \sum_{i=1}^{D} p_i = 1 \tag{7}$$

where strict inequalities are indicated, instead of the more general $p_j \geq 0$, in the light of what follows. The sum constraint implies what p is essentially a d-dimensional entity where $d = D - 1$. The geometrical object formed by all points satisfying conditions (7) is the standard d-simplex in $\mathbb{R}^D$, denoted $\mathbb{S}^d$.

A standard way to handle compositional data, extensively examined by Aitchison (1986), is via a suitable transformation from $\mathbb{S}^d$ to $\mathbb{R}^d$ followed by the introduction of a suitable probability distribution on $\mathbb{R}^d$. This distribution, however, is required to enjoy various formal properties so to ensure that certain operations on the original simplex space can be translated into corresponding operations on the Euclidean space, and *vice versa*. These requirement explains why the multivariate normal distribution has played a prominent role in this context. However, Mateu-Figueras *et al.* (2005, 2007) have shown how the formal properties of (4) allow to use it as a replacement of the normal family in this context with enhanced flexibility thanks to the extra parameter which regulates shape, and still retain a great deal of the features of the original formulation linked to on the normal distribution.

3 Distributions on the circle

Circular data, which arise when observations represent angles, are the simplest form of directional data. Since the origin is arbitrary, treatment of circular data requires to introduce probability distributions which are periodic; in the continuous case, this means that the density $f(\theta)$ at angle θ satisfies $f(\theta + 2\pi) = f(\theta)$, for all θ.

A great deal of work on circular data has adopted a symmetric density $f(\theta)$, as it apparent from a standard source like Mardia and Jupp (1999). Proposals exist to handle skewed data distributions, and the above-described formulation provides a viable route in this direction. A wrapped skew-normal distribution on the circle has been studied by Arthur Pewsey in a series of papers; see specifically Pewsey (2006). A related bimodal distribution has been considered by Hernández-Sánchez and Scarpa (2012).

A somewhat difference route has been followed by Umbach and Jammalamadaka (2009, 2010) who examine distributions of type (2)–(3) taking f_0 and $g_0 = G_0'$ to be symmetric densities on the unit circle. These conditions lead to a direct extension on the unit circle of the stochastic representations (5) and the property of perturbation invariance (6).

4 Generalized symmetry

The above-summarized standard setting hinges on the requirement of symmetry. This condition can however be relaxed, introducing the following requirements. Assume that there exists an invertible transformation $R(x)$ on $\mathbb{R}^d$ such that, for all $x \in \mathbb{R}^d$,

$$f_0(x) = f_0(R(x)), \qquad |\mathrm{det}R'(x)| = 1, \qquad w(R(x)) = -w(x) \tag{8}$$

where $R'(x)$ denotes the Jacobian matrix of partial derivatives. When $R(x) = -x$ we are back to the construction of Section 1. It can be shown that, under conditions (8), the first of which represents a form of *generalized symmetry*, the key facts of Section 1 still hold, with appropriate adaptions (Azzalini, 2012). A specifically relevant fact is that the perturbation invariance property (6) holds if $t(\cdot)$ satisfies $t(x) = t(R^{-1}(x))$.

As as example, consider a perturbed form of the bivariate normal density $\phi_2(x; \bar{\Omega})$ given by

$$2\, \phi_2(x; \bar{\Omega})\, \Phi(\alpha(x_1^2 - x_2^2)), \qquad x = (x_1, x_2) \in \mathbb{R}^2,$$

where $\alpha \in \mathbb{R}$ is a parameter. In this case the perturbation factor does not satisfy (1), but conditions (8) hold with

$$R(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x \tag{9}$$

implying that 2 is the appropriate normalization factor, and a χ^2 property like for (4) holds here too.

5 Discrete distributions

The final part of the talk summarizes work in progress, jointly with Giuliana Regoli, where a construction similar to (2)–(3) is adopted in the case of a discrete symmetric distribution base f_0. Various results hold, matching many of those for the continuous context. Finally, a specific formulation of this family is developed for fitting the score difference of football matches, where the regulating parameters reflect strength and other characteristics of the competing teams in a given tournament. The aims is to construct predictive distributions of future matches tuned for the specific teams in play.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, Chapman & Hall.

Azzalini, A. (2012). Selection models under generalized symmetry settings *Ann. Inst. Stat. Math.*, **64**, 737–750.

Hernández-Sánchez, E., and Scarpa, B. (2012). A wrapped flexible generalized skew-normal model for a bimodal circular distribution of wind directions. *Chil. J. Statist.*, **3**, 131–143.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Barceló-Vidal, C. (2005). Additive logistic skew-normal on the simplex. *Stochastic Environmental Research and Risk Assessment*, **19**, 205-–214.

Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2007). The skew-normal distribution on the simplex. *Commun. Statist. Theory Methods*, **36**, 1787-–1802.

Mardia, K. V., and Jupp, P. E. (1999). *Directional Statistics*. J. Wiley & Sons.

Pewsey, A. (2006). Modelling asymmetrically distributed circular data using the wrapped skew-normal distribution. *Environmental & Ecological Statistics*, **13**, 257–269.

Umbach, D., and Jammalamadaka, S. R. (2009). Building asymmetry into circular distributions. *Statist. Probab. Lett.*, **79**, 659—663.

Umbach, D., and Jammalamadaka, S. R. (2010). Some moment properties of skew-symmetric circular distributions. *Metron*, **LXVIII**, 265–273.