How to sample from the SN and related distributions

when we want to fix skewness and other cumulants

Adelchi Azzalini
adelchi.azzalini@unipd.it

18th September 2015, updated 15th October 2015

A recurring question in my e-mail and also elsewhere is the following: ‘I do not grasp how to sample data from a skew-normal distribution with a certain skewness parameter — can you help?’. A variant form of the question has ‘skew-*t*’ (ST) in place of ‘skew-normal’ (SN). These pages address the above question in practice, using the R package sn. Background theoretical aspects are kept as limited as possible, but cannot be avoided altogether. It is assumed that the command

```
library(sn)
```

has been successfully issued in your R session.

1 Parameterizations

There are various parameterization is use in the package sn; the main ones are DP (direct parameterization) and CP (centred parameterization). Some points to bear in mind are as follows.

1. DP refers to the parameters appearing in the expression of the densities, but they are not so directly interpretable as the CP components. For instance, the first CP component is the mean value of the distribution, which is a standard measure of location; the first DP component, ξ, is instead not individually related to the mean or the median or other measures of location.

2. We can move back and forth between the two sets, from DP to CP using the function dp2cp and *vice versa* from CP to DP using the function cp2dp. There is a difference however: while any choice of the DP components is admissible and dp2cp will always lend an answer, the same is not true for CP; if we have picked up an inadmissible set of CP components, cp2dp will produce nothing.

One more word is appropriate about ‘the skewness’. Skewness or asymmetry is a qualitative feature of a distribution; there are many ways to measure it. Among these measures, the more common one is the ‘coefficient of skewness’ represented by the third standardized cumulant, denoted γ_1, which constitutes the third component of the CP set, in the univariate case; in the multivariate case we consider a vector of such univariate values. Since, for the univariate SN distribution there is a one-to-one correspondence between the γ_1 and α, the matching component of the DP set, then also α represents skewness, but it operates on a different scale. In other cases, namely for the ST distribution and especially in multivariate contest, the situation is slightly more complex and the analogous correspondence involves more components of the parameter sets.

2 How to sample

2.1 Univariate distributions

Consider a probability distribution with mean value 3, standard deviation 1.2, skewness (γ_1) 0.8, for which we write

```
    cp <- c(mean=3, s.d.=1.2, gamma1=0.8)
```

or simply

```
    cp <- c(3, 1.2, 0.8)
```

and assume that they refer to some member of the SN family. From here we get the corresponding DP vector with

```
    dp <- cp2dp(cp, family="SN")
```

If we print the content of `dp`, this is the outcome

```
    R> dp
        xi omega alpha
     1.523 1.903 4.190
```

Now we can sample 20, say, random values from this distribution with

```
    y <- rsn(20, dp=dp)
```

We could write more compactly on one line

```
    y <- rsn(20, dp=cp2dp(cp, family="SN"))
```

but this is *not* recommended for general usage: in many cases the `rsn(20, ...)` line will be within a loop to be executed many times and it is advisable to keep the conversion step `cp2dp(cp, family="SN")` outside that loop. In the present example, the saving in execution time will be limited (unless the loop has very, very many cycles), but in more complex situations to be described later, the saving can be somewhat more appreciable.

The feasible range for γ_1 of the SN distribution is quite restricted; it is required that $|\gamma_1| < 0.99527\ldots$. Therefore the code

```
    cp <- c(3, 1.2, 1.6)
    dp <- cp2dp(cp, family="SN")
```

will produce the message

```
    gamma1 outside admissible range
```

To overcome this limitation, we must consider a wider family di distributions, such as the ST family. Using a wider family implies to specify an additional parameter component. In this case, the extra parameter is constituted by the coefficient of kurtosis, denoted γ_2, that is, the fourth standardized cumulant. Hence we could write for instance:

```
cpST <- c(3, 1.2, 1.6, 6.1)
dpST <- cp2dp(cpST, family="ST")
y2 <- rst(20, dp=dpST)
```

The range of feasible (γ_1,γ_2) pairs for the ST distribution is quite large, but still there are limitations. So we may incur in the message 'CP outside admissible range'. A pictorial representation of the feasible (γ_1,γ_2) area in provided in Figure 1 of Arellano-Valle & Azzalini (2013); see also p. 104 of Azzalini & Capitanio (2014).

Notice that we can select extra ST distributions, not in that area, for instance by writing

```
param <- c(xi=1, omega=2, alpha=3, nu=3.3)
y3 <- rst(10, dp=param)
```

and get a legitimate ST sample. The explanation is that this DP vector has no corresponding CP vector, since γ_2 does not exist in this case; therefore the command

```
dp2cp(param, family="ST")
```

will not return a CP vector, but still the distribution is legitimate.

To summarize, with the CP vector there are parameter restrictions; with the DP vector you can pick-up any values you like, provided omega>0 and nu>0. This is the key reason why DP is preferred to CP when we need to specify one of these distributions, although CP is constituted by more familiar quantities.

2.2 Multivariate distributions

For the multivariate SN distribution, individual components of the DP and CP sets are replaced by vectors, except the scale parameter which is now a positive-definite symmetric matrix. The whole set of ingredients is then encapsulated in a R list; in the multivariate setting, the list elements *must* be named.

As an example of a bi-dimensional SN distribution, consider

```
Sigma <- matrix(c(1, 0.4, 0.4, 0.6), 2, 2)
cpM <- list(mean=c(0,-1), var.cov=Sigma, gamma1=c(0, -0.6))
dpM <- cp2dp(cpM, family="SN")
yM <- rmsn(10, dp=dpM)
```

where mu and gamma1 are formed by individual components which have the same meaning as in the univariate case and var.cov is the (co)variance matrix of the distribution. The component-wise interpretation of the elements of gamma1 does *not* hold instead for the elements of the matching DP element alpha; the same applies to the other elements of DP.

The earlier points about parameter restrictions in the univariate case carry on here. Analogously to the univariate case, we move on to the multivariate ST distribution to achieve a wider range of skewness and kurtosis. This is accomplished by introducing an extra element, gamma2M, which represents the Mardia's measure of multivariate kurtosis; note that gamma2M is a single value, not a vector. Hence the revised version of the above chunk of code is

```
Sigma <- matrix(c(1, 0.4, 0.4, 0.6), 2, 2)
```

```
cpM <- list(mean=c(0,-1), var.cov=Sigma, gamma1=c(0, -0.6), gamma2M=10)
dpM <- cp2dp(cpM, family="ST")
yM <- rmst(10, dp=dpM)
```

The step of parameter conversion using `cp2dp` for the ST distribution involves the solution of non-linear equations. Therefore it is more involved than for the SN distribution, especially so in the multivariate case. Hence here the earlier cautionary note against unnecessary parameter conversions becomes more relevant.

References

Arellano-Valle, R. B., and Azzalini, A. (2013). The centred parameterization and related quantities of the skew-t distribution. *J. Multiv. Anal.*, **113**, 73–90. Available online 12 June 2011.

Azzalini, A. with the collaboration of Capitanio, A. (2014). *The Skew-Normal and Related Families*. IMS Monographs series, Cambridge University Press.