Environmental Modelling & Software 90 (2017) 27–33

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

# A finite mixture approach to univariate data simulation with moment matching

Earl Bardsley

*Faculty of Science and Engineering, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand*

ARTICLE INFO

*Article history:*
Received 31 July 2016
Received in revised form
17 November 2016
Accepted 17 November 2016

*Keywords:*
Moment matching
Finite mixture distribution
Data simulation
Beta distribution

ABSTRACT

Environmental data simulation is carried out for various purposes and is most simply achieved when recorded data can be regarded as a sequence of independent random variables. Simulating from such data involves generating random variables from some fitted probability distribution, preserving the main statistical characteristics. The particular case of data simulation with moment matching is revisited, with a proposal that data be simulated from many-component finite mixture distributions. The simulation procedure matches data moments while also giving greater flexibility for data approximation. In contrast, a single parametric distribution fitted to irregular data will result in data simulations more representative of the fitted distribution than the original data. The flexible finite mixture approach therefore has potential to replace parametric univariate data simulation generally. The method is illustrated with simulations from finite mixtures of generalised beta distributions with matching of data mean, variance, skewness and kurtosis.



## 1. Introduction

Environmental data simulation has many practical applications and a range of simulation methodologies have been proposed which encapsulate the often complex temporal structure of many environmental time series. Recent publications include Wasko et al. (2015), Koutsoyiannis (2016), Tsekouras and Koutsoyiannis (2014), Efstratiadis et al. (2014), and Srivastav and Simonovic (2014).

Despite the ubiquitous nature of persistence and other effects, some environmental time series and other recorded data sets may be regarded to a first approximation as a sequence of independent random variables from an unknown univariate distribution. Data simulation then reduces to the classic model of generating random variables from some specified probability distribution with similar statistical properties to the recorded data (Kuhl et al., 2010). This method is revisited here for the specific situation where it is desired that simulations match the moments of the original data, which is a helpful way of quantifying statistical connection between recorded and simulated data. That is, the probability distribution generating the simulations must have the same first moments as the recorded data.

The simplest approach to moment matching is bootstrap assignment of selection probability $1/N$ to each member of a given set of $N$ recorded data values. Simulations from such bootstrap distributions preserve data moment properties but have the disadvantage that only previously recorded values can be simulated. Generating different-valued data while maintaining moment matching can be achieved by simulating from a fitted standard probability distribution which has the same first few moments as the data. However, even the more flexible parametric distributions are not able to approximate recorded data which has some degree of irregularity. Simulated data in such cases are likely to be more reflective of the fitted distribution than the original data, despite matching of moments.

Non-parametric kernel-based distributions have flexibility to approximate even the most irregular data but will not generally match the moments or other specific statistical characteristics of the recorded data. See, for example, Howe (2013) and cited references. Taylor and Thompson (1986) describe a data-based simulation approach which has some similarities to the method to be presented here. However, their method is essentially a kernel approach which does not seek to match moments.

This brief communication is concerned with presenting an alternative method of moment-matching simulation from univariate data. It has the particular advantage of improved flexibility to simulate data characterised by some degree of irregularity.

The proposed simulation method is simple, comprising repeats of the 3-step process: (i) sample $m$ data values at random and without replacement from some existing set of $N$ recorded data

*E-mail address:* earl.bardsley@waikato.ac.nz.

http://dx.doi.org/10.1016/j.envsoft.2016.11.019
1364-8152/

values, (ii) specify the parameter values of a parametric distribution such that its first $k$ moments are the same as the first $k$ moments of the $m$ data values, (iii) generate a random variable from that distribution.

The "without replacement" aspect here is only with respect to selecting a given set of $m$ data values. The $m$ data values are conceptually returned to the data and some may be reselected in a later sample draw.

## 2. Simulating from finite mixture distributions

The simulation process as described equates to generating random variables from a finite mixture distribution made up of $J = C(N, m)$ component parametric distributions, each one of which has the same first $k$ moments as its corresponding set of $m$ data values. For example, $N = 100$ and $m = 4$ creates a finite mixture distribution made up of $C\ (100,4) = 3{,}921{,}225$ component parametric distributions because there are 3,921,225 possible subsets of 4 data values obtainable from 100 data values.

Any finite mixture distribution so defined will have the same first $k$ central moments as the data set of $N$ values, as is noted below.

The equality of central moments of the data and a finite mixture distribution follows from the expression for $\mu'_i$, denoting the $i$ th moment about zero of an unweighted finite mixture distribution:

$$\mu'_i = J^{-1} \sum_{j=1}^{J} \mu'_{ij} \quad i \le k \tag{1}$$

where $\mu'_{ij}$ is the $i$ th moment about zero for the $j$ th component distribution of the finite mixture distribution.

Eq. (1) also provides the $i$ th moment about zero of the $N$ data values, given that the $i$ th moments about zero of the $j$ th component distribution are equal to the $i$ th moments about zero of the $j$ th data subset. That is, there is equality of the moments about zero for the finite mixture distribution and the data set.

If any distribution's first $k$ moments about zero are known, then the first $k$ central moments must also be known because central moments can be expressed as functions of moments about zero. Therefore, if any two distributions both have the same first $k$ moments about zero then they must also have the same first $k$ central moments. It follows that a finite mixture distribution as previously defined must have the same first $k$ central moments as the $N$ data values. Given this linkage of moments, the first $k$ central moments of the simulated data will tend toward the first $k$ central moments of the $N$ data values as the number of simulations increases.

Different choices of component distributions will give rise to different finite mixture distributions. However, all possible finite mixture distributions as defined here will always have the same first $k$ central moments as the $N$ data values.

A specific requirement of the procedure is that each of the $C(N,m)$ component distributions must have parameters which give the same first $k$ distribution moments as apply for the respective data samples of size $m$. It could happen that for some data samples it will not be possible to assign the required parameter values. For example, if all the $m$ sample values happen to be the same. In such situations the simulated variable is simply a bootstrap selection of one of the $m$ data values. That is, some of the component distributions in the finite mixture may be discrete bootstrap distributions. However, this does not affect the equivalence of the central moments of the finite mixture distribution and those of the $N$ data values.

The finite mixture distributions proposed here are simply a means to a convenient algorithm for simulating data with moment matching. There is no suggestion that a many-component finite mixture density function might serve to estimate the density function of the unknown distribution generating the recorded data. Indeed, the mixture density functions in their fine detail will often be characterised by numerous singularities and probability spikes, as well as possibly including some discrete component distributions. However, this is not an issue for data simulation purposes, which requires only that the distribution function of a finite mixture distribution gives a good smoothed approximation of the data empirical distribution function. This will also be evident as strong similarity between histograms of data and simulated data. Because $C(N,m)$ increases rapidly with $N$, it would be expected that distribution function approximations would hold even for relatively small $N$.

For example, suppose there are 10 recorded data values and $m$ is set to 4. This yields $C\ (10,4) = 210$ component distributions of the finite mixture distribution. This is illustrated in Fig. 1 for a finite mixture distribution function comprised of 210 four-parameter beta component distributions, with parameter values determined from all possible combinations of four values drawn from the ten data values: −1, −0.5, 0, 1, 2, 8, 9, 10.5, 11, 12. There is no suggestion of course that data sets of ten values should be used for practical data simulation. Details of setting up beta component distributions are given in Section 3.

Considering now the effect of the magnitude of $m$ on the finite mixture distributions, larger $m$ values will not in general yield finite mixture distributions which better approximate the original data. This applies particularly if there is some degree of data irregularity. The smallest possible $m$, corresponding to the number of distribution parameters, results in a higher probability that all the $m$ values of a given sample will be located near a mode in the data distribution. This creates a higher frequency of simulated values near data modes. When $m$ is large the sample values will be more widely spread and give a greater probability of the associated parametric distributions having a larger standard deviation and therefore simulating data less reflective of data modes.

It is evident that $m$ has similarities to the bandwidth parameter in kernel-based distribution estimation, with larger values of $m$ resulting in less sensitivity to irregular data variation.

## 3. Simulating from generalised beta distribution mixtures

Kuhl et al. (2010) point to the utility of fitting a single generalised (4-parameter) beta distribution to data as a basis for data simulation. A many-component finite mixture of generalised beta distributions therefore has attraction for even more flexible approximation to data while still matching the mean, variance, skewness, and kurtosis of the data values.

With respect symbolism, the 4-parameter beta probability density function is defined:

$$f(x;a,b,p,q) = c(x-a)^{p-1}(b-x)^{q-1} \quad a<x<b;\ \ p,q>0 \tag{2}$$

where $p$ and $q$ are shape parameters, $a$ and $b$ define the distribution end points, and $c$ is a positive term dependent on the parameter values.

There is restriction on the choice of $m$ values with beta finite mixture distributions. This is because as $m$ increases it is possible that the $m$ data values in a sample could have a configuration such that the sample skewness or kurtosis is outside the beta distribution region on the Pearson beta plane. Generating a random variable from a beta distribution would not then be possible. In such situations a random variable would need to be simulated from some other applicable parametric distribution, or a bootstrap simulation from the $m$ sample values would be utilised.

The beta plane region for $m = 4$ falls entirely within the beta