4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

GENERATION OF CORRELATED NON-GAUSSIAN RANDOM VARIABLES FROM INDEPENDENT COMPONENTS

Juha Karvanen

Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000, FIN-02015 HUT, Finland
juha.karvanen@hut.fi

Laboratory for Advanced Brain Signal Processing
Brain Science Institute, Riken
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

ABSTRACT

Simulations are often needed when the performance of new methods is evaluated. If the method is designed to be blind or robust, simulation studies must cover the whole range of potential random input. It follows that there is a need for advanced tools of data generation. The purpose of this paper is to introduce a technique for the generation of correlated multivariate random data with non-Gaussian marginal distributions. The output random variables are obtained as linear combinations of independent components. The covariance matrix and the first four moments of the output variables may be freely chosen. Moreover, the output variables may be filtered in order to add autocorrelation. Extended Generalized Lambda Distribution (EGLD) is proposed as a distribution for the independent components. Examples demonstrate the diversity of data structures that can be generated.

1. INTRODUCTION

The assumption on the Gaussianity of the observed data is needed for several fundamental theoretical results in various fields of science. For instance, in signal processing, Wiener and Kalman filters are optimal for Gaussian noise. In communications, the channel capacities are derived for channels with additive white Gaussian noise (AWGN). In statistics, all traditional methods, such as regression analysis, analysis of variance and factor analysis, are based on the Gaussianity assumption. The central limit theorem provides the theoretical foundation for the use of Gaussian distribution. Another explanation for the dominance of the Gaussianity is that closed form solutions are often available only for the Gaussian case.

In practice, non-Gaussian data is encountered in numerous applications. The beginning of the list could be, for instance, speech signals, urban channel noise, fading channels, biomedical signals, image data, etc. The existence of independent component analysis (ICA) as a research topic is also an evidence of the importance of non-Gaussianity. ICA, as well many other methods in modern statistics and neural computation, combines non-Gaussianity and numerical optimization. This means often that it is impossible to derive analytical results. Consequently, the performance of the methods must be evaluated in simulations.

It is usually difficult to exhaustively evaluate the performance of a new method in simulations. Good results from a single test do not guarantee good performance in general. A large variety of tests is especially important if the tested method is claimed to be blind or robust. If artificial data is used, the statistical properties of the data must be controlled and alternated. This leads us to the problem addressed in this paper. In the non-Gaussian case, it is nontrivial task to explicitly define a multivariate distribution with an arbitrary dependence structure. The purpose of this paper is to introduce a technique for the generation of correlated multivariate random data with non-Gaussian marginal distributions. The marginal distributions are characterized by first four moments and the dependence is characterized by a covariance matrix. Both the covariance matrix and the first four moments of the marginal distributions may be freely chosen. The ICA model is employed in a straightforward manner in the generation. The output random variables are obtained as linear combinations of independent components. Moreover, the output variables may be filtered in order to add autocorrelation.

In Section 2 the mixing matrix and the moments of the independent sources are derived from the given correlation matrix and output moments. In Section 3 the EGLD is proposed for the generation of the independent sources. Examples are provided in Section 4. Finally, the benefits and alternatives of the proposed technique are discussed in Section 5

2. MODEL FOR DATA GENERATION

The basic model used in data generation is the ICA model with instantaneous mixing

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where the observations $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ are generated from the mutually independent sources $\mathbf{s} = [s_1, s_2, \ldots, s_m]^T$. Matrix $\mathbf{A}_{m\times m}$ is called as a mixing matrix.

The problem of the data generation may be defined as follows: Given $\mathbf{\Sigma_x}$, the covariance matrix of $\mathbf{x}$ and the first four central moments $\mu(x_i), \mu_2(x_i), \mu_3(x_i), \mu_4(x_i)$ for each component of $\mathbf{x}$, find such mixing matrix $\mathbf{A}$ and such distributions for the components of $\mathbf{s}$ that the model (1) holds.

First we concentrate on finding the mixing matrix $\mathbf{A}$ and assume that we can generate independent univariate random variables with the given first four moments as their theoretical moments. A solution for the actual generation of such variables is presented in Section 3.

A solution for the mixing matrix $\mathbf{A}$ is obtained using the covariance matrix $\mathbf{\Sigma_x}$. The mixing matrix $\mathbf{A}$ is obtained e.g. as the real part square root of $\mathbf{\Sigma_x}$ or via Cholesky factorization. This corresponds to the standard whitening in ICA but because we are dealing with the theoretical covariance matrix all results are exact. Since correlation is unaffected in a rotation, the mixing matrix $\mathbf{A}$ can be multiplied by an arbitrary orthogonal matrix $\mathbf{U}$. It turns out that the matrix $\mathbf{U}$ is a parameter that has an effect on the fifth and higher order moments of output distributions. This is demonstrated in Section 4.

The rotation (together with the source distributions) defines the higher order dependence structure. It is possible to fix the rotation e.g. choosing the desired fourth-order cross-cumulants and then finding a fourth-order factorization similar to the eigenvalue decomposition of the covariance matrix [3, 4, 2]. However, this approach is not suggested because the interpretation of the fourth-order cross-cumulant matrices is not as intuitive as the interpretation of the covariance matrix.

After the mixing matrix is fixed, it is possible to calculate the moments of the sources from the desired moments of the outputs. Using the properties of the moments in the case of independent random variables we obtain

$$\mu(x_i) = \sum_{j=1}^{m} a_{ij}\mu(s_i) \tag{2}$$

$$\mu_2(x_i) = \sum_{j=1}^{m} a_{ij}^2\mu_2(s_i) \tag{3}$$

$$\mu_3(x_i) = \sum_{j=1}^{m} a_{ij}^3\mu_3(s_i) \tag{4}$$

$$\mu_4(x_i) = \sum_{j=1}^{m} a_{ij}^4(\mu_4(s_i) - 3\mu_2^2(s_i)) + 3\mu_2^2(x_i), \tag{5}$$

where $\mathbf{A} = [a_{ij}]$. If we define $A_q = [a_{ij}^q]$ and $\mu_q(\mathbf{x}) = [\mu_q(x_1), \mu_q(x_2), \ldots, \mu_q(x_m)]^T$ we can write the solution for the moments of the sources as follows

$$\mu(\mathbf{s}) = A^{-1}\mu(\mathbf{x}) \tag{6}$$

$$\mu_2(\mathbf{s}) = A_2^{-1}\mu_2(\mathbf{x}) \tag{7}$$

$$\mu_3(\mathbf{s}) = A_3^{-1}\mu_3(\mathbf{x}) \tag{8}$$

$$\mu_4(\mathbf{s}) = A_4^{-1}(\mu_4(\mathbf{x}) - 3\mu_2^2(\mathbf{x})) + 3\mu_2^2(\mathbf{s}). \tag{9}$$

Obviously, some combinations of the correlation and the moments are theoretically impossible. The covariance matrix must be positive definite. According a result from statistics [5] a necessary condition for the central moments is

$$\begin{vmatrix} 1 & 0 & \ldots & \mu_k \\ 0 & \mu_2 & \ldots & \mu_{k+1} \\ \vdots & & \ddots & \vdots \\ \mu_k & \mu_{k+1} & \ldots & \mu_{2k} \end{vmatrix} \geq 0 \tag{10}$$

for all integers $k \geq 1$. The validity of the moments and the covariance matrix of $\mathbf{x}$ may be checked calculating first the moments of $\mathbf{s}$ from (6) and then applying the condition (10) separately for the moments of each component of $\mathbf{s}$. Because we are dealing only with first four moments, only cases $k = 1$ and $k = 2$ need to be checked. If the condition (10) does not hold for the moments of some component s_i, it is not possible to generate observations $\mathbf{x}$ with the desired moments from the independent components with the chosen rotation matrix $\mathbf{U}$.

The generating model (1) may be generalized in order to allow autocorrelated output. A temporal structure may be added defining

$$v_i = b_i(z)x_i, \tag{11}$$

where $b_i(z) = b_{i0} + b_{i1}z^{-1} + \ldots + b_{il}z^{-l}$ is the transfer function of a finite impulse response (FIR) filter and x_i is a linear mixture of independent sources defined in (1). The model (11) can be seen as a FIR-MIMO (multiple input - multiple output) model where the structure of the transfer

functions is restricted. The model is convolutive but the delayed cross-correlations are complete defined by the instantaneous covariance matrix and the component-wise transfer functions. The system combining instantaneous mixing (1) and FIR-filtering (11) is illustrated in Figure 1.

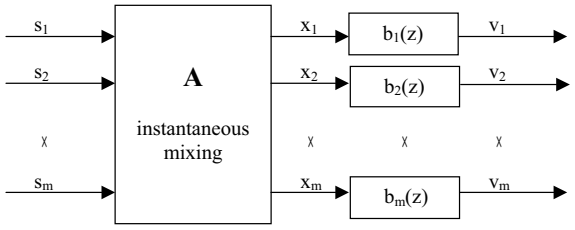

Fig. 1. A system with instantaneous mixing and FIR-filters $b_1, b_2, \ldots, b_m$. Output variables $v_1, v_2, \ldots, v_m$ with user-defined statistical properties are generated from independent sources $s_1, s_2, \ldots, s_m$

Autocorrelated random variables $\mathbf{v} = [v_1, v_2, \ldots, v_m]^T$ with given covariance matrix $\mathbf{\Sigma_v}$, given first four moments and given transfer functions $b_1, b_2, \ldots, b_m$, are generated from independent sources $\mathbf{s} = [s_1, s_2, \ldots, s_m]^T$. The procedure of generating $\mathbf{x}$ from the sources $\mathbf{s}$ is presented above. The covariance matrix and moments of $\mathbf{x}$ can be solved from the covariance matrix and moments of $\mathbf{v}$ as follows

$$\mu(x_i) = \frac{\mu(v_i)}{\sum_{j=1}^{l} b_{ij}} \tag{12}$$

$$\mu_2(x_i) = \frac{\mu_2(v_i)}{\sum_{j=1}^{l} b_{ij}^2} \tag{13}$$

$$\mu_3(x_i) = \frac{\mu_3(v_i)}{\sum_{j=1}^{l} b_{ij}^3} \tag{14}$$

$$\mu_4(x_i) = \frac{\mu_4(v_i) - 3\mu_2^2(v_i)}{\sum_{j=1}^{l} b_{ij}^4} + 3\mu_2^2(x_i) \tag{15}$$

$$\mathbf{\Sigma_x} = \left[\sqrt{\mu_2(x_i)\mu_2(x_j)}\right] = \left[\frac{\sqrt{\mu_2(v_i)\mu_2(v_j)}}{\sum_{k=1}^{l} b_{ik}b_{jk}}\right]. \tag{16}$$

3. EXTENDED GENERALIZED LAMBDA DISTRIBUTION (EGLD) AS A DISTRIBUTION OF THE SOURCES

In the previous section the relation between the moments of the output and the moments of the sources was established. What remains to be solved is the generation of univariate random variables with the given moments. We propose the EGLD to be used as the distribution of the sources. In the EGLD family there exist a distribution for every set of first four moments that is theoretically possible. The EGLD is a combination of Generalized Lambda Distribution (GLD) and Generalized Beta Distribution (GBD). The distribution was presented in [15], generalized in [13, 12] and extended in [8]. In ICA, the EGLD has been applied as a model of the output distributions [6, 9].

The GLD is defined by the inverse distribution function

$$F^{-1}(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \tag{17}$$

where $0 \leq p \leq 1$ and λ_1, λ_2, λ_3 and λ_4 are the parameters of the distribution. Observations are easily generated from the GLD employing the inverse distribution function. If p has Uniform(0,1) distribution then the random variable $y = F^{-1}(p)$ has a distribution with the cumulative distribution function F. In the case of the GLD, the inverse distribution function is available in the closed form, which is not true with most of standard distributions. This makes the GLD especially useful in the random variable generation.

Next, we need a mapping from the moments to the GLD parameters. The relationship between parameters λ_1, λ_2, λ_3 and λ_4 and moments μ, μ_2, μ_3 and μ_4 is established by four non-linear equations that can be solved numerically [8]. However, due to the intricacy of the computational process, the parameters λ_1, λ_2, λ_3 and λ_4 are tabulated in [7] as functions of μ_3 and μ_4 for the case where $\mu = 0$ and $\mu_2 = 1$.

It is shown [8] that (17) is a valid distribution if and only if

$$\frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}} \geq 0. \tag{18}$$

If $\lambda_3 > 0$ and $\lambda_4 > 0$, it follows from (17) that the distribution is defined in the finite interval $[\lambda_1 - \frac{1}{\lambda_2}, \lambda_1 + \frac{1}{\lambda_2}]$. In many cases, these distributions may successfully approximate also distributions defined in infinite domain. For instance, the distribution function with the parameter values $\lambda_1 = 0$, $\lambda_2 = 0.1975$, $\lambda_3 = \lambda_4 = 0.1349$ differs from N(0,1) distribution function by at most 0.002 [5]. If $\lambda_3 < 0$ and $\lambda_4 < 0$ the GLD has infinite domain $(-\infty, \infty)$. For the tabulated parameter values it holds $\lambda_3 \lambda_4 > 0$. The coverage of the GLD in (μ_3^2, μ_4)-space (with $\mu = 0$ and $\mu_2 = 1$) is illustrated in Figure 2.

For the moment values not covered by the GLD, the GBD extension is used. The GBD is characterized by the density function

$$f(x) = C\beta_2^{-(\beta_3+\beta_4+1)}(x-\beta_1)^{\beta_3}(\beta_1+\beta_2-x)^{\beta_4} \tag{19}$$

on the interval $[\beta_1, \beta_1 + \beta_2]$ and zero elsewhere. The β_1 β_2 β_3 and β_4 are the parameters of the distribution and C is a constant. The area that the GBD covers is given in terms of moments

$$1 + \mu_3^2 < \mu_4 < 3 + 2\mu_3^2. \tag{20}$$

The coverage of the GBD in (μ_3^2, μ_4)-space (with $\mu = 0$ and $\mu_2 = 1$) is illustrated in Figure 3.

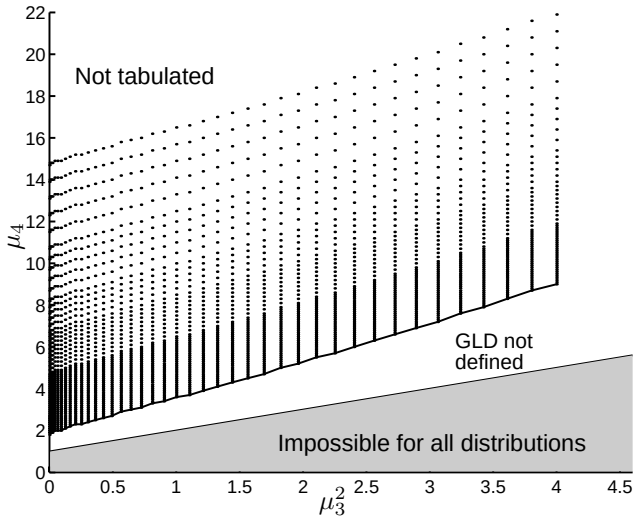

Fig. 2. Illustration of the GLD in (μ_3^2, μ_4)-space (with $\mu = 0$ and $\mu_2 = 1$). The GLD tables [7] give a solution for the points marked by dots. For other suitable (μ_3^2, μ_4) values, the lambda parameters may be interpolated. The tables are not available for very high values of the kurtosis even if the GLD family covers also such values. The GBD extension is needed for the low-kurtosis region where the GLD is not defined.

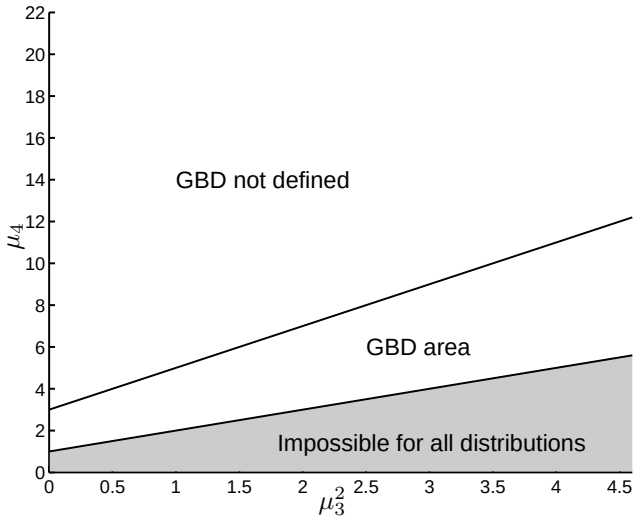

Fig. 3. Illustration of the GBD in (μ_3^2, μ_4)-space (with $\mu = 0$ and $\mu_2 = 1$). The GBD extension is needed for the low-kurtosis region where the GLD is not defined.

Tables are also provided [7] for the GBD parameters β_3 and β_4 as functions of the central moments μ_3 and μ_4. However, the parameters can be obtained directly by solving the moment equations

$$\mu_3 = \frac{2(\beta_4 - \beta_3)\sqrt{C_3}}{C_4\sqrt{(\beta_3+1)(\beta_4+1)}} \tag{21}$$

$$\mu_4 = \frac{3C_3(\beta_3\beta_4 C_2 + 3\beta_3^2 + 5\beta_3 + 3\beta_4^2 + 5\beta_4 + 4)}{C_4 C_5(\beta_3+1)(\beta_4+1)}, \tag{22}$$

where $C_k = \beta_3 + \beta_4 + k$ for $k = 1, \ldots, 5$. Then the parameters β_1 and β_2 are given by

$$\beta_2 = (\beta_3 + \beta_4 + 2)\sqrt{\frac{(\beta_3 + \beta_4 + 3)\mu_2}{(\beta_3+1)(\beta_4+1)}} \tag{23}$$

$$\beta_1 = \mu - \frac{\beta_2(\beta_3+1)}{\beta_3 + \beta_4 + 2}. \tag{24}$$

A beta distributed random variable can be generated from two Gamma distributed variables. The method is described in [5] and implemented e.g. in Matlab.

4. EXAMPLES

The first example illustrates the diversity of correlation structures and marginal distributions that can be generated using the proposed procedure. The pairwise scatter plots of five random variables are presented in Figure 4. 200 observations are generated from the multivariate distribution with the following user defined properties: The matrix of the correlation coefficients is

$$\begin{pmatrix} 1 & 0.7 & 0.3 & -0.2 & -0.2 \\ 0.7 & 1 & 0.2 & -0.2 & -0.2 \\ 0.3 & 0.2 & 1 & -0.3 & -0.2 \\ -0.2 & -0.2 & -0.3 & 1 & 0.4 \\ -0.2 & -0.2 & -0.2 & 0.4 & 1 \end{pmatrix}.$$

All marginal distributions have zero mean and unit variance. The theoretical values for the third moment are 0.6, -0.4, 0, 0.4, 0.2 and for the fourth moment 5, 4, 3, 2, 1.5.

The effect of rotation is studied in the second example. In the bivariate case, an orthogonal matrix is characterized by a rotation angle θ. In the example the covariance matrix and the first four moments are kept constant while the rotation is changed. The resulting scatter plots are presented in Figure 5. Information on the theoretical moments is given in Table 1. It can be seen that the effect of the rotation remarkable even if correlation and the first four moments are same in every subfigure. The source distributions change when rotation is changed. This has an effect on the fifth and higher order moments of the output.

5. CONCLUSION

In this paper a technique for multivariate data generation is introduced. The benefits of the proposed technique are

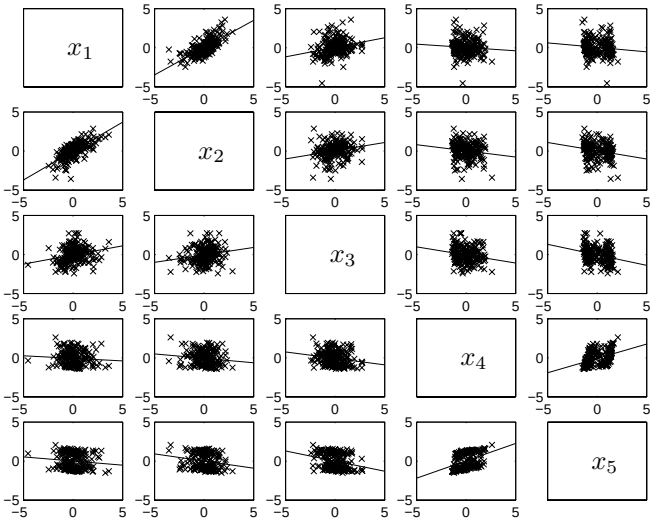

Fig. 4. A scatter plot matrix of 200 observations generated from a five-variate distribution with predefined first four moments and covariance matrix. The least-square regression lines are shown in every subplot. The marginal distributions include symmetric (x_3), skewed (x_1, x_2 and x_4) and bimodal (x_5) distributions.

- The random variable generation can be implemented combining existing building blocks.
- The dependence structure and marginal distributions are defined in terms of covariance matrix and moments which are easy to interpret.
- The generated random variables may be autocorrelated.
- The covariance matrix and the first four moments may be flexibly chosen resulting a large variety of multivariate structures.

Many multivariate distributions proposed in literature do not share these properties. For example, the Dirichlet distribution and the bivariate Pearson system have rather restricted dependence structure and marginal distributions [14]. Koehler-Symanowski distribution family [10] is an example of multivariate distribution with flexible marginals but it does not provide any direct way to generate observations or control the correlation structure [1]. Mixtures of densities (Note that this is the original use of word 'mixture' even if in ICA, the observed linear combinations are often also called as mixtures.) allow great flexibility for the marginal distributions but defining the multivariate dependence may be problematic. A general framework for multivariate density mixtures is provided in [11].

Instead of the EGLD family, other families of distributions may be utilized as the source generating distribution. There are two requirements for the family of distributions. First, the distribution family must be parameterized such a way that every set of the first four moments can be explicitly mapped to some parameter values. Second, it must be possible to generate observations from the distribution. In the case of the EGLD, tables are provided to map the moments to the parameters. It is particularly easy to generate observations from the GLD. Generating observations from the GBD is not as straightforward but still possible. Other potential choices for the source generating distribution are, for instance, the Pearson system, the Johnson systems and the Burr system [14, 5].

Using only four moments to characterize the output distributions may be considered as a drawback. On the other hand, the completely defined marginal distributions restrict the possible dependence structures. In some cases it might also be useful to have several distributions that have the same moments up to order four. These can be easily generated changing the rotation matrix.

Besides introducing a tool for the data generation, this paper demonstrates that even the simplest linear ICA model allows versatile data structures. This supports the use of ICA in various applications.

6. REFERENCES

[1] A. Caputo. Some properties of the family of Koehler and Symanowski distributions. Discussion Paper 103, University of Munich, 1998.

[2] J.-F. Cardoso. Fourth-order cumulant structure forcing. application to blind array processing. In *Proc. 6th*

SSAP workshop on statistical signal and array processing, pages 136–139, 1992.

[3] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[4] J.-F. Cardoso and P. Comon. Independent component analysis, a survey of some algebraic methods. In *Proc. ISCAS'96*, volume 2, pages 93–96, 1996.

[5] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[6] J. Eriksson, J. Karvanen, and V. Koivunen. Source distribution adaptive maximum likelihood estimation of ICA model. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2000*, pages 227–232, 2000.

[7] Z. A. Karian and E. J. Dudewicz. The extended generalized lambda distribution (EGLD) system for fitting distributions to data with moments, II: Tables. *American Journal of Mathematical and Management Sciences*, 1996.

[8] Z. A. Karian, E. J. Dudewicz, and P. McDonald. The extended generalized lambda distribution system for fitting distributions to data: History, completion of theory, tables, applications, the "final word" on moment fits. *Communications in Statistics: Simulation and Computation*, 25(3):611–642, 1996.

[9] J. Karvanen, J. Eriksson, and V. Koivunen. Adaptive score functions for maximum likelihood ICA. *Journal of VLSI Signal Processing*, 32:83–92, 2002.

[10] K. Koehler and J. Symanowski. Constructing multivariate distributions with specific marginal distributions. *Journal of Multivariate Analysis*, 55:261–282, 1995.

[11] A. W. Marshall and I. Olkin. Families of multivariate distributions. *Journal of American Statistical Association*, 83(834–841), 1988.

[12] J. S. Ramberg, E. J. Dudewicz, P. R. Tadikamalla, and E. F. Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21:201–204, 1979.

[13] J. S. Ramberg and B. W. Schmeiser. An approximate method for generating asymmetric random variables. *Communications of Association for Computing Machinery*, 17:78–82, 1974.

[14] A. Stuart and J. K. Ord. *Kendall's Advanced Theory of Statistics: Distribution Theory*, volume 1. Edward Arnold, sixth edition, 1994.

[15] J. W. Tukey. The practical relationship between the common transformations of percentages of counts and of amounts. Technical Report 36, Statistical Techniques Research Group, Princeton University, 1960.

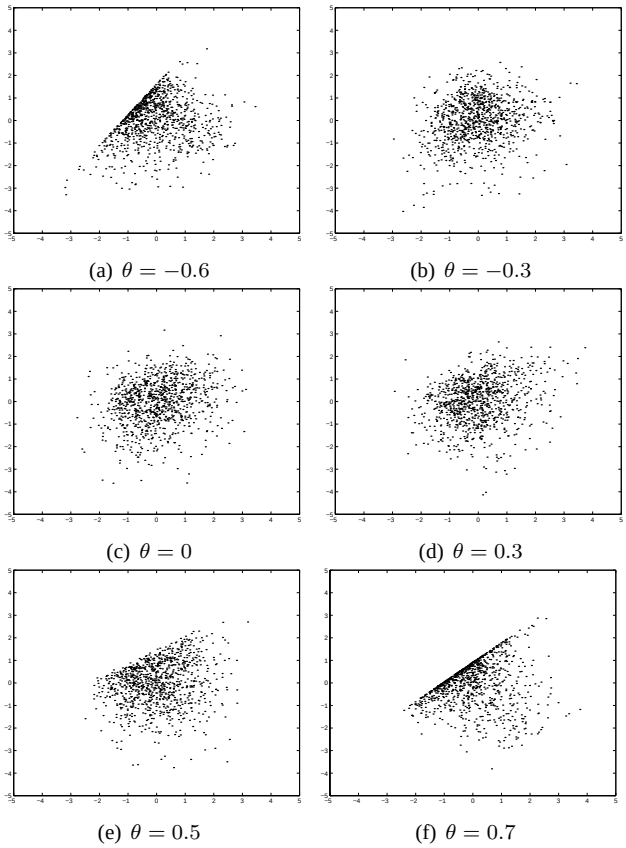

Fig. 5. An illustration of the effect of rotation. In all subfigures the distributions have the same theoretical moments up to order four and the same correlation 0.2. θ is the rotation angle compared to the original moments whitening. When the rotation is changed the source distributions change. As a result the output distributions also change. Even if the first four moments are the same, the scatter plots clearly differ from each other. Information on the moments is given in Table 1.

	Component 1		Component 2	
	μ_3	μ_4	μ_3	μ_4
output (a)-(f)	0.3	3.2	-0.5	3.4
sources (a)	1.02	3.26	-0.57	3.65
sources (b)	0.42	3.26	-0.53	3.43
sources (c)	0.31	3.21	-0.51	3.41
sources (d)	0.31	3.21	-0.66	3.55
sources (e)	0.31	3.25	-0.99	3.81
sources (f)	0.04	3.08	-1.53	4.61

Table 1. The theoretical third and fourth moment of the output and the sources from Figure 5.