Nonlin. Processes Geophys., 15, 159–167, 2008
www.nonlin-processes-geophys.net/15/159/2008/

Nonlinear Processes in Geophysics

Detecting spatial patterns with the cumulant function – Part 1: The theory

A. Bernacchia[1] **and P. Naveau**[2]

[1]Dipartimento di Fisica E.Fermi, Universita' La Sapienza, Roma, Italy
[2]Laboratoire des Sciences du Climat et de l'Environnement, IPSL-CNRS, France

Received: 5 July 2007 – Revised: 3 September 2007 – Accepted: 17 September 2007 – Published: 19 February 2008

Abstract. In climate studies, detecting spatial patterns that largely deviate from the sample mean still remains a statistical challenge. Although a Principal Component Analysis (PCA), or equivalently a Empirical Orthogonal Functions (EOF) decomposition, is often applied for this purpose, it provides meaningful results only if the underlying multivariate distribution is Gaussian. Indeed, PCA is based on optimizing second order moments, and the covariance matrix captures the full dependence structure of multivariate Gaussian vectors. Whenever the application at hand can not satisfy this normality hypothesis (e.g. precipitation data), alternatives and/or improvements to PCA have to be developed and studied.

To go beyond this second order statistics constraint, that limits the applicability of the PCA, we take advantage of the cumulant function that can produce higher order moments information. The cumulant function, well-known in the statistical literature, allows us to propose a new, simple and fast procedure to identify spatial patterns for non-Gaussian data. Our algorithm consists in maximizing the cumulant function. Three families of multivariate random vectors, for which explicit computations are obtained, are implemented to illustrate our approach. In addition, we show that our algorithm corresponds to selecting the directions along which projected data display the largest spread over the marginal probability density tails.

1 Introduction

In geosciences, Principal Component Analysis (PCA) has been an essential and powerful tool at detecting spatial structures amongst time series recorded at different locations. PCA is a dimensionality reduction technique that extracts relevant components of data, those responsible for the largest proportion of variability (Rencher, 1998). PCA builds decorrelated components of the data and it finds the spatial patterns that maximize the variance. Hence, second order moments are the foundation of PCA. But relying exclusively on second moments implies that PCA is only optimal when applied to multivariate Gaussian vectors. Although rarely stated and even more rarely checked, this underlined normality assumption is not always satisfied in practice.

Recently, different approaches have been tested to extend the applicability of PCA in geosciences. NonLinear PCA (NLPCA) has been applied to several geophysical datasets (e.g. Hsieh, 2004; Monahan et al., 2001). In the NLPCA algorithm, data are considered as the input of an auto-associative neural network with five layers, with a bottleneck in the third layer (Kramer, 1991). Through the minimization of a cost function, the output is forced to be as close as possible to the input, and the bottleneck layer is a low dimensional representation of the input. Since this neural network is nonlinear, NLPCA goes automatically beyond correlations. However, NLPCA suffers the intrinsic limitations of multilayered networks (e.g. Christiansen, 2005; Malthouse, 1998): it is computationally expensive and does not always converge to a global solution. Independent Component Analysis (ICA) has been also applied in the geosciences (Aires et al., 2000). ICA builds independent (rather than uncorrelated) components of the data, if any exist, by minimizing the entropy of the marginal density in the general non-Gaussian case (Bell and Sejnowski, 1995; Hyvarinen and Oja, 2000). Here we pursue a less ambitious aim: instead of trying to find decompositions that can explain the entire body of data with respect to a criterion, we focus on the part of data responsible for large anomalous behaviors.

In contrast to PCA, our approach tends to give maximal weight to data points which largely deviate from the mean, and to find the corresponding representative spatial patterns, i.e. the directions along which such points are prominently

Correspondence to: A. Bernacchia
(a.bernacchia@gmail.com)

distributed. The key element in our procedure is the expansion of the cumulant function that can provide information beyond the first two moments. By maximizing the cumulant function (Kenney and Keeping, 1951) over growing hyperspheres in the data space, a set of components can be derived. We first illustrate our procedure by the application to three synthetic types of multivariate random vectors (Normal, Skew-Normal, Gamma). Then we demonstrate that, for any multivariate distribution, a larger cumulant function along a given direction is implied by a fatter tail of the corresponding marginal probability density. Hence, our algorithm provide the directions along which anomalies are mostly expected.

We show that the first principal component of PCA is a special case of our approach whenever the Gaussian assumption is satisfied. Besides the Gaussian case we show that, for any probability density, the solution derived from the cumulant function can be transformed to the first principal component by decreasing the radius of the hypersphere. Other principal components could be found as well, by a generalization of the proposed method. Our method is computationally cheap, and the solutions are found in the form of unit (normalized) vectors, as in the case of PCA, allowing a unidimensional projection with an easy geometrical interpretation.

In summary, this paper focuses on the problem of characterizing spatial patterns associated to large anomalies, i.e. large deviations from the sample mean, when the data set under study cannot be assumed to be normally distributed.

2 Maximizing the cumulant function

In the univariate case, the cumulant function of the random variable X with finite moments is defined as the following scalar function

$$\log\left\{\mathbb{E}\big[\exp(sX)\big]\right\} = \sum_{n=1}^{\infty} \kappa_n \frac{s^n}{n!},$$

where $s \in \mathbb{R}$, $\mathbb{E}(.)$ represents the mean function and the scalar κ_n corresponds to the n^{th} cumulant of X.

The first two cumulants κ_1 and κ_2 are simply the mean and the variance of X, respectively. The third and fourth cumulants are classically called the skewness and the kurtosis parameters. Concerning the existence of cumulants, we assume in this paper that all the cumulant coefficients are finite and that the cumulant function is always well defined. The cumulant function and its coefficients have many interesting properties. For example, if X and Y are two independent random variables, then the n-th cumulant of the sum $X+Y$ is equal to the sum of the n-th cumulant of X and the n-th cumulant of Y for any integers n. If X follows a Gaussian distribution, then all but the first two cumulants are equal to zero. In a multivariate framework, the cumulant function of the random vector $\mathbf{X}=(X_1, \ldots, X_m)^t$ is simply defined as

$$\log\left\{\mathbb{E}\big[\exp(\mathbf{s}^t\mathbf{X})\big]\right\}, \text{ for all } \mathbf{s}^t = (s_1, \ldots, s_m) \in \mathbb{R}^m. \quad (1)$$

As in the univariate case, the linear and Gaussian properties associated to the cumulant function defined by Eq. (1) still hold, but the cumulant coefficients formulas are more cumbersome to write down in a multivariate framework. For more information about cumulants, we refer the reader to Kenney and Keeping (1951).

To identify possible favorite projection directions with respect to the multivariate cumulant function, we first rewrite the vector $\mathbf{s}=(s_1, \ldots, s_m)^t$ in Eq. (1) as the product $\mathbf{s}=|\mathbf{s}|\times\boldsymbol{\theta}$, where $|\mathbf{s}|^2=\sum s_i^2$, the scalar $|\mathbf{s}|$ represents the norm ("radius") of $\mathbf{s}$ and $\boldsymbol{\theta}$ is the unit "angular/direction" vector defined as $\theta_i=s_i/|\mathbf{s}|$ (note that $\boldsymbol{\theta}^t\boldsymbol{\theta}=1$). Secondly, the cumulant function for our vector $\mathbf{X}=(X_1, \ldots, X_m)^t$ projected along the direction vector $\boldsymbol{\theta}$ is introduced as

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) = \log\left\{\mathbb{E}\big[\exp(|\mathbf{s}|\ \boldsymbol{\theta}^t\mathbf{X})\big]\right\} = \sum_{n=1}^{\infty} k_n(\boldsymbol{\theta}) \frac{|\mathbf{s}|^n}{n!}. \quad (2)$$

Our algorithmic strategy is to maximize the cumulant function, at fixed non-small $|\mathbf{s}|$, with respect to the angular component $\boldsymbol{\theta}$ that varies over an unit hypersphere. Practically, we have to find the optimal $\boldsymbol{\theta}_s$ directions defined by

$$\boldsymbol{\theta}_s = \text{argmax}\left[G_{|\mathbf{s}|}(\boldsymbol{\theta}) \text{ such that } \boldsymbol{\theta}^t\boldsymbol{\theta} = 1\right] \quad (3)$$

for a fixed value of $|\mathbf{s}|$. If the radius $|\mathbf{s}|$ is small enough, and the mean k_1 is zero, the variance $k_2(\boldsymbol{\theta})$ dominates the cumulant function and the contributions of other cumulants can be neglected. In this situation, finding the first PCA component can be viewed as a special case of this optimization procedure, because maximizing the cumulant function for small $|\mathbf{s}|$ is equivalent to maximize the variance. As the value of $|\mathbf{s}|$ grows, higher and higher order cumulants become more dominant. Our main goal is to find $\boldsymbol{\theta}_s$ in Eq. (3) for the largest admissible $|\mathbf{s}|$ and study their properties. We will call the solutions of such an optimization scheme the "Maxima of the Cumulant Function" (MCF) directions.

We anticipate that, since the scalar product $\theta^t\mathbf{X}$ is invariant under orthogonal transformations, the cumulant function is invariant as well. Given that the unit hypersphere is also invariant, our algorithm is symmetric with respect to orthogonal transformations. For instance, if data vectors are rotated by a given angle, the solutions of the algorithm, in terms of $\boldsymbol{\theta}$, are rotated by the same amount. This symmetry implies that if the probability density is isotropic, then the cumulant function is isotropic as well, and no relative maxima of G exist on the unit hypersphere. In that case no directions are selected: for a rotationally symmetric distribution there is indeed no preferred direction along which anomalies are prominent, they are distributed uniformly over all angles.

To illustrate our optimization procedure, we derive in Sect. 3 explicit cumulant maximization schemes for three special cases of multivariate family distributions. Finally, in Sect. 4 we demonstrate that, in the general case, if a direction exists for which the marginal probability density of projected data display a larger tail than in other directions, our procedure is able to select that direction, which corresponds to the maximum of the cumulant function. Hence, our algorithm provide the directions along which anomalies are mostly expected. For assessing the outputs of our algorithm when applied to real data, we refer the reader to the second part of this paper (Bernacchia et al., 2008).

3 Theoretical examples

To study the properties of the maximization method developed in Sect. 2, three examples of distribution functions are considered in this paper. We choose these three families because explicit results can be derived and they have been classically used in the statistical modeling of temperatures and precipitation data.

Without loss of generality, some relevant matrices are diagonal thereafter. However, since the solutions covary with orthogonal transformations, they may be rotated along with the corresponding coordinate change. Analytical calculations are performed for the general multivariate case, while figures are given for the bivariate case.

3.1 Multivariate Gaussian vectors

Suppose that the data at hand can be appropriately fitted by a multivariate Gaussian vector. We assume that the observations have been centered (zero) mean and we denote the covariance matrix as Σ. The cumulant function of the centered Gaussian vector (e.g. Kenney and Keeping, 1951) is equal to

$$\log\left\{\mathbb{E}\left[\exp(\mathbf{s}^t\mathbf{X})\right]\right\} = \frac{1}{2}\mathbf{s}^t\Sigma\mathbf{s}.$$

Hence, it is easy to show that all cumulants but the second are equal to zero. The decomposition in Eq. (2), $\mathbf{s}=|\mathbf{s}|\times\boldsymbol{\theta}$, implies that the cumulant function becomes

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) = \frac{|\mathbf{s}|^2}{2}k_2(\boldsymbol{\theta}) = \frac{|\mathbf{s}|^2}{2}\,\boldsymbol{\theta}^t\Sigma\boldsymbol{\theta}$$

Our optimization problem is to maximize $G_{|\mathbf{s}|}(\boldsymbol{\theta})$ under the constraint $\boldsymbol{\theta}^t\boldsymbol{\theta}=1$. To find the optimal $\boldsymbol{\theta}_s$ defined by Eq. (3), we introduce a function L to be maximized, constrained by the Lagrange multiplier λ, as

$$L(\lambda,\boldsymbol{\theta}) = \frac{|\mathbf{s}|^2}{2}\,\boldsymbol{\theta}^t\Sigma\boldsymbol{\theta} - \lambda(\boldsymbol{\theta}^t\boldsymbol{\theta}-1).$$

Setting the derivative of L with respect to λ to zero gives the constraint $\boldsymbol{\theta}^t\boldsymbol{\theta}=1$, while setting the gradient with respect to $\boldsymbol{\theta}$ to zero gives

$$\nabla_{\boldsymbol{\theta}}L = |\mathbf{s}|^2\Sigma\boldsymbol{\theta} - 2\lambda\boldsymbol{\theta} = 0. \quad (4)$$

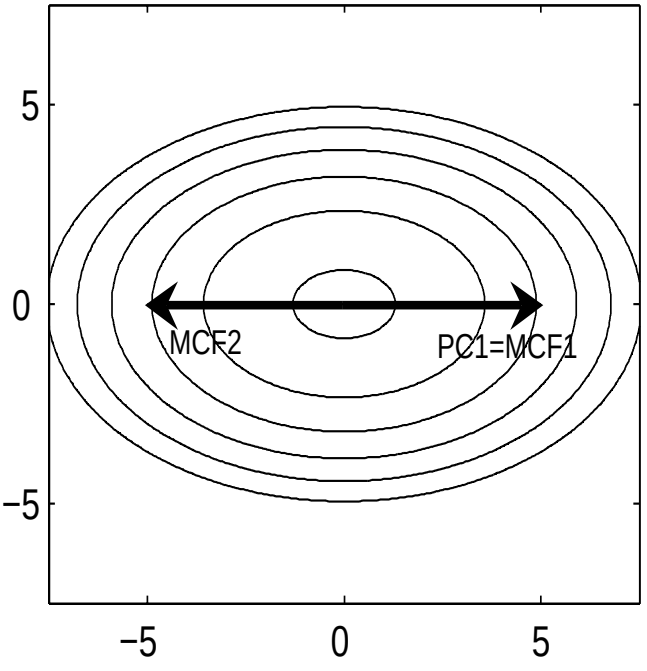

Fig. 1. Isoprobability contours of the bivariate Gaussian distribution, with zero mean and variances 1.2 and 0.5143 (entries of the diagonal covariance matrix). The first principal component is shown (PC1), together with the two (opposite) maxima of the cumulant function (MCF1 and MCF2), for any value of $|\mathbf{s}|$. All vectors are in arbitrary scale. The maxima of cumulant function are parallel to the first principal component, all pointing towards the large anomalies, in terms of high probability (at fixed vector norm). Probability contours are 10^{-1}, 10^{-3}, $10^{-5}\ldots10^{-11}$.

Let λ_Σ be the largest eigenvalue of the covariance matrix Σ and $\boldsymbol{\theta}_\Sigma$ its associated eigenvector. Introducing $\lambda_s=\frac{|\mathbf{s}|^2}{2}\lambda_\Sigma$, we can write $|\mathbf{s}|^2\Sigma\boldsymbol{\theta}_\Sigma=2\lambda_s\boldsymbol{\theta}_\Sigma$. Since the eigenvalue λ_s is the largest one, for a fixed $|\mathbf{s}|$, $\boldsymbol{\theta}_\Sigma$ is the maxima of the cumulant function. Note that $\boldsymbol{\theta}_\Sigma$ depends on Σ but not on $|\mathbf{s}|$. The optimal direction, for the Gaussian case, is $\boldsymbol{\theta}_s=\boldsymbol{\theta}_\Sigma$, and it corresponds to the classical first principal component.

To illustrate this result, a bivariate vector of the normal distribution is presented in Fig. 1 (in which a contour plot is drawn in logarithmic scale). The matrix Σ is assumed to be diagonal, with entries 1.2 and 0.5143. The first principal component (PC1), corresponding to the eigenvalue 1.2 is horizontal, while the second principal component, corresponding to the eigenvalue 0.5143 is vertical, and is not displayed. The two maxima of the cumulant function (for any value of $|\mathbf{s}|$), MCF1 and MCF2, are just the positive and negative part of PC1. Indeed, both $\boldsymbol{\theta}_\Sigma$ and $-\boldsymbol{\theta}_\Sigma$ are solutions of Eq. (4). While this is always true for PCA, the maxima of the cumulant function may in general neither be parallel nor orthogonal.

PC1 is indeed the direction along which large anomalies are distributed in the Gaussian case. In order to derive PC2

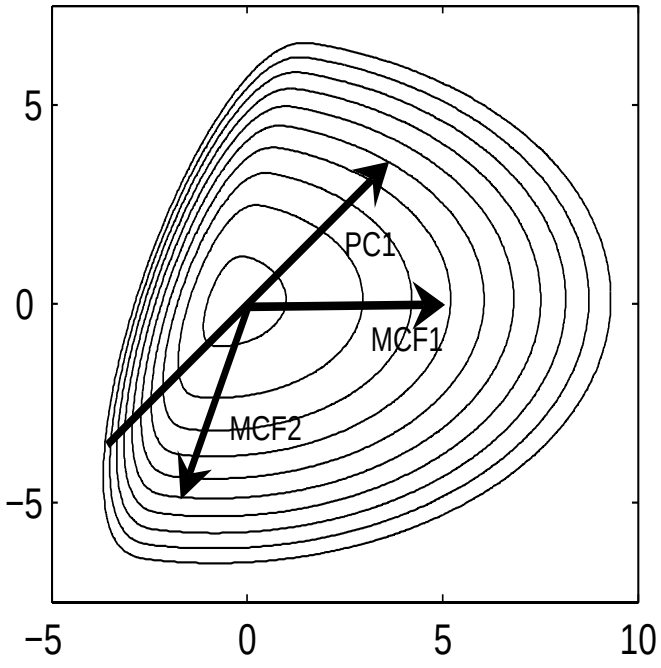

Fig. 2. Isoprobability contours of the bivariate Skew-Normal distribution, with parameters α=(4.365, −1.455) and the matrix Σ is chosen to be diagonal with entries 1.2 and 0.5143. The first principal component is shown (PC1), together with the two maxima of the cumulant function (MCF1 and MCF2), for large $|\mathbf{s}|$. All vectors are in arbitrary scale. In this case, the two maxima of the cumulant function do not correspond to the first principal component, they are neither parallel nor orthogonal, and point towards the (local) large anomalies, in terms of high probability (at fixed, and large, vector norm). Probability contours are 10^{-1}, 10^{-3}, $10^{-5} \ldots 10^{-19}$.

and other higher principal components from the cumulant function, one would have to determine not only its maxima, but also its minima and saddle points.

3.2 Multivariate Skew-Normal vectors

To introduce skewness to the Gaussian density, while keeping some of valuable properties of the normal distribution, Azzalini and his co-authors have extended the normal density to a larger class, called the Skew-Normal (SN) density (e.g. Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999; Gonzalez-Farias et al., 2004), that is defined as

$$f(\mathbf{x}) = 2\phi_{\Sigma}(\mathbf{x})\Phi(\boldsymbol{\alpha}^t\mathbf{x}) \tag{5}$$

where ϕ is a multivariate Normal probability density function with zero mean and covariance matrix Σ, and Φ is the cumulative density function of an univariate Gaussian random variable with zero mean and unit variance. The vector $\boldsymbol{\alpha}$ corresponds to the degree of skewness. When $\boldsymbol{\alpha}=\mathbf{0}$, there is no skewness, and the SN distribution reduces to the Gaussian case. From Eq. (5), it is possible to derive the cumulant function (e.g Azzalini and Dalla Valle, 1996) of a SN vector

$$\log\Big\{\mathbb{E}\big[\exp(\mathbf{s}^t\mathbf{X})\big]\Big\} = \frac{1}{2}\mathbf{s}^t\Sigma\mathbf{s} + \log\big[2\Phi(\sqrt{\tfrac{\pi}{2}}\boldsymbol{\mu}^t\mathbf{s})\big].$$

where $\boldsymbol{\mu}$ represents the mean vector, and is equal to

$$\boldsymbol{\mu} = \frac{\Sigma\boldsymbol{\alpha}}{\sqrt{\frac{\pi}{2}(1+\boldsymbol{\alpha}^t\Sigma\boldsymbol{\alpha})}}.$$

Note that the covariance matrix of the SN distribution is not Σ but $\Sigma-\boldsymbol{\mu}\boldsymbol{\mu}^t$ (Azzalini and Capitanio, 1999).

A bivariate example of the SN distribution is presented in Fig. 2, in which a contour plot is drawn in logarithmic scale. The matrix Σ is chosen to be diagonal with entries 1.2 and 0.5143. The skewness vector α is taken to be equal to (4.365, −1.455). The distribution in Fig. 2 has been centered such that the mean vector is zero, i.e. the original distribution is translated by $\boldsymbol{\mu}$, which is equal to (0.8367, −0.1195).

From the SN cumulant function (Azzalini and Capitanio, 1999), we can write the cumulant function of the centered vector $(\mathbf{X}-\boldsymbol{\mu})$ as

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) = -|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta} + \frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t\Sigma\boldsymbol{\theta} + \log\Big[2\Phi(\sqrt{\tfrac{\pi}{2}}|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta})\Big] \tag{6}$$

While it is not possible to find explicit solutions of the maximization problem defined by Eq. (3) for Eq. (6), one can provide valuable approximated solutions for both small and large $|\mathbf{s}|$. In the former case, the following two Taylor expansions are the key elements to derive our results

$$\log(1+s) = 1 + s - \frac{s^2}{2} + o(s^3) \text{ and} \tag{7}$$

$$1 + \mathrm{erf}(s) = 1 + \frac{2s}{\sqrt{\pi}} + o(s^3) \tag{8}$$

where erf corresponds to the error function defined by

$$2\Phi(s) = 1 + \mathrm{erf}\Big(\frac{s}{\sqrt{2}}\Big).$$

Then we can write the following approximation

$$\begin{aligned}\log\Big[2\Phi(\sqrt{\tfrac{\pi}{2}}|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta})\Big] &= \log\Big[1+\mathrm{erf}(\tfrac{\sqrt{\pi}}{2}|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta})\Big] \\ &= \log\Big[1+|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta}+o(|\mathbf{s}|^3)\Big], \text{ by (7)}, \\ &= |\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta} - \frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t\boldsymbol{\mu}\boldsymbol{\mu}^t\boldsymbol{\theta} + o(|\mathbf{s}|^3), \text{ by (8)}.\end{aligned}$$

From Eq. (6), it follows that the cumulant function $G_{|\mathbf{s}|}(\boldsymbol{\theta})$ is approximately equal to

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) \simeq \frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t(\Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^t)\boldsymbol{\theta} \quad \text{, for small } |\mathbf{s}|. \tag{9}$$

As previously noticed, the matrix $\Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^t$ represents the covariance of the SN distribution (Azzalini and Capitanio, 1999). Hence, the maximization of the right hand side of Eq. (9) is equivalent to solving the system defined by Eq. (4), but instead of working with Σ, we just need to replace Σ by $\Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^t$ in Eq. (4). Consequently, the solution to maximize the SN cumulant function in the neighborhood of zero is the largest eigenvector of the matrix $\Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^t$, i.e. the PC1 of the SN covariance matrix.

For large $|\mathbf{s}|$, this result does not hold and different directions are obtained. We need to recall the asymptotic expansion of the error function

$$1 + \mathrm{erf}(s) \simeq \begin{cases} -\frac{1}{s\sqrt{\pi}} \exp(-s^2)\Big[1 + o(s^{-3})\Big] & , \text{as } s \downarrow -\infty, \\ 2 & , \text{as } s \uparrow +\infty. \end{cases}$$

If $\boldsymbol{\mu}^t\boldsymbol{\theta} < 0$, the logarithm in Eq. (6) can be expanded as

$$\log\Big[2\Phi(\sqrt{\tfrac{\pi}{2}}|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta})\Big] = \log\Big[1 + \mathrm{erf}(\tfrac{\sqrt{\pi}}{2}|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta})\Big] \simeq -\frac{\pi}{2}\frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t\boldsymbol{\mu}\boldsymbol{\mu}^t\boldsymbol{\theta} + O\big(\log(|\mathbf{s}|^{-1})\big).$$

In this case, we have $G_{|\mathbf{s}|}(\boldsymbol{\theta}) \simeq \frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t\Big(\Sigma - \frac{\pi}{2}\boldsymbol{\mu}\boldsymbol{\mu}^t\Big)\boldsymbol{\theta}$. The direction that maximizes $G_{|\mathbf{s}|}(\boldsymbol{\theta})$ is an eigenvector of the matrix $\Sigma - \frac{\pi}{2}\boldsymbol{\mu}\boldsymbol{\mu}^t$, pointing towards $\boldsymbol{\mu}^t\boldsymbol{\theta} < 0$.

If $\boldsymbol{\mu}^t\boldsymbol{\theta} \geq 0$, we have

$$\log\Big[2\Phi(\sqrt{\tfrac{\pi}{2}}|\mathbf{s}|\boldsymbol{\mu}^t\boldsymbol{\theta})\Big] \simeq \begin{cases} \log 2 & , \text{if } \boldsymbol{\mu}^t\boldsymbol{\theta} > 0, \\ 0 & , \text{if } \boldsymbol{\mu}^t\boldsymbol{\theta} = 0. \end{cases}$$

The cumulant function $G_{|\mathbf{s}|}(\boldsymbol{\theta})$ can be then approximated by $G_{|\mathbf{s}|}(\boldsymbol{\theta}) \simeq \frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t\Sigma\boldsymbol{\theta}$; it is maximized by the largest eigenvector of Σ, pointing towards $\boldsymbol{\mu}^t\boldsymbol{\theta} \geq 0$. In summary, depending on the size of $|\mathbf{s}|$ (small or large) and the sign $\boldsymbol{\mu}^t\boldsymbol{\theta}$ (positive or negative), the solutions of Eq. (3) for the SN distribution can be viewed as the largest eigenvectors of each of three different matrices, $\Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^t$, $\Sigma - \frac{\pi}{2}\boldsymbol{\mu}\boldsymbol{\mu}^t$ and Σ.

For the bivariate example of Fig. 2, the PC1 is shown (in arbitrary scale), explaining 60% of the variance. The second PC (40% of the variance) is orthogonal to PC1 and is not displayed. Both maxima of the cumulant function for large $|\mathbf{s}|$, denoted as MCF1 and MCF2 (respectively for $\boldsymbol{\mu}^t\boldsymbol{\theta} \geq 0$ and $\mu^t\boldsymbol{\theta} < 0$), are presented in Fig. 2 for the bivariate example (same scale as PC1). The two local maxima point towards the large anomalies of the distribution: this can be seen by noting that a point at the upper-right end of PC1 corresponds to a small probability, and hence is less likely to be found, than a point at the right end of MCF1 (the two points being of equal norm). Similarly, a point at the down-left end of PC1 is less likely, in probability, than a point at the down end of MCF2.

3.3 Multivariate Gamma vectors

This section investigates the multivariate Gamma distribution defined by Cheriyan and Ramabhadran (see Kotz et al., 1998). Each component of the data vector $\mathbf{X}$ is distributed following a Gamma distribution, and the components depend each other by means of an auxiliary variable z. The joint distribution is

$$f(\mathbf{x}) = \int g(z;\alpha_0)\prod_{i=1}^{n} g(x_i - z;\alpha_i)dz \qquad (10)$$

where g is a gamma distribution, i.e.

$$g(z,\alpha) = \frac{e^{-z}z^{\alpha-1}}{\Gamma(\alpha)}$$

for $z \geq 0$, equal to zero otherwise. A bivariate example is presented in Fig. 3, with $n=2$, $\alpha_0=2$, $\alpha_1=0.5$ and $\alpha_2=4$ in Eq. (10).

The cumulant function for this multivariate Gamma distribution can be written as (Kotz et al., 1998)

$$\log\Big\{\mathbb{E}\big[\exp(\mathbf{s}^t\mathbf{X})\big]\Big\} = -\alpha_0\log\Big(1 - \sum_{i=1}^{n} s_i\Big) - \sum_{i=1}^{n}\alpha_i\log(1 - s_i)$$

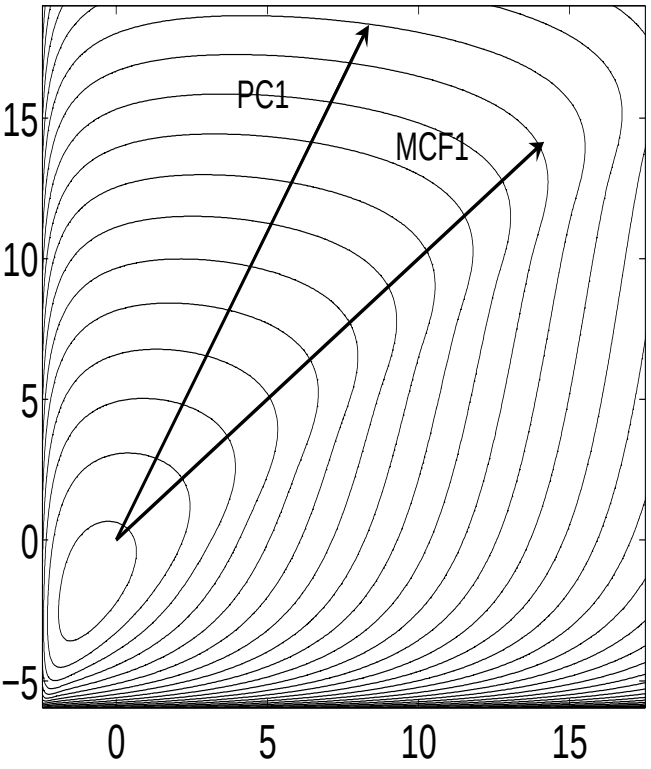

Fig. 3. Isoprobability contours of the bivariate Gamma distribution, with parameters α_0=2, α_1=0.5, α_2=4. The first principal component is shown (PC1), together with the maximum of the cumulant function (MCF1), for the largest admissible value of $|\mathbf{s}|=1/\sqrt{2}$. All vectors are in arbitrary scale. Again, the maximum of the cumulant function is different from the first principal component, and points towards the large anomalies, in terms of high probability (at fixed, and large, vector norm). Probability contours are $10^{-1/2}$, 10^{-1}, $10^{-3/2}\ldots$.

and the mean of the i-th component is $\mu_i=\alpha_0+\alpha_i$. By replacing $\mathbf{s}$ by $|\mathbf{s}|\times\boldsymbol{\theta}$, the cumulant function of the centered vector $(\mathbf{X}-\boldsymbol{\mu})$ can be written as

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) = -\alpha_0 \log\Big(1-|\mathbf{s}|\sum_{i=1}^{n}\theta_i\Big) - \sum_{i=1}^{n}\alpha_i\log(1-|\mathbf{s}|\theta_i) - |\mathbf{s}|\sum_{i=1}^{n}(\alpha_0+\alpha_i)\theta_i. \quad (11)$$

For small $|\mathbf{s}|$, the logarithms are approximated by the truncated Taylor expansions, i.e.

$$\sum_{i=1}^{n}\alpha_i\log\big(1-|\mathbf{s}|\theta_i\big) \simeq -|\mathbf{s}|\sum_{i=1}^{n}\alpha_i\theta_i + \frac{|\mathbf{s}|^2}{2}\sum_{i=1}^{n}\alpha_i\theta_i^2$$

and

$$\alpha_0\log\big(1-|\mathbf{s}|\sum_{i=1}^{n}\theta_i\big) \simeq -|\mathbf{s}|\sum_{i=1}^{n}\alpha_0\theta_i + \frac{|\mathbf{s}|^2}{2}\alpha_0\Big(\sum_{i=1}^{n}\theta_i\Big)^2.$$

It follows that the cumulant function can be approximated by

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) \simeq \frac{|\mathbf{s}|^2}{2}\boldsymbol{\theta}^t\mathbf{C}\boldsymbol{\theta}, \quad (12)$$

where the covariance matrix $\mathbf{C}$ is defined by (Kotz et al., 1998)

$$\mathbf{C} = \begin{pmatrix} \alpha_0+\alpha_1 & & \alpha_0 \\ & \ddots & \\ \alpha_0 & & \alpha_0+\alpha_n \end{pmatrix}.$$

Hence, for small $|\mathbf{s}|$, the solution of Eq. (3) is the classical PC1 eigenvector of the covariance matrix C associated with the largest eigenvalue. It is plotted in Fig. 3 for the bivariate example. The negative part of PC1 is not displayed, as well as the second principal component which is just orthogonal to the first.

For large $|\mathbf{s}|$, the cumulant function is not defined, since the logarithms in Eq. (11) must have positive arguments, for all unit vectors $\boldsymbol{\theta}$. In particular, the following two inequalities have to be satisfied

$$|\mathbf{s}|\theta_i < 1 \quad \text{and} \quad |\mathbf{s}|\sum_{i=1}^{n}\theta_i < 1.$$

In other words,

$$|\mathbf{s}| < \min\left(\min\left(\frac{1}{\theta_i}\right), \min\left(\frac{1}{\sum_i\theta_i}\right)\right).$$

However, we take the largest allowed value of $|\mathbf{s}|$, which is considered as a valid limit. Since $\boldsymbol{\theta}$ is a unit vector, the maximum of each component θ_i is 1, which holds when all the other components are zero, while the maximum for the sum $\sum_i\theta_i$ is $\sqrt{n}$, which holds when $\theta_1=\theta_2=\ldots=\theta_n=1/\sqrt{n}$. The largest allowed value of $|\mathbf{s}|$ is then $1/\sqrt{n}$: in that case G remains finite for all $\boldsymbol{\theta}$'s, except for $\theta_1=\theta_2=\ldots=\theta_n=1/\sqrt{n}$, where it diverges due to the first logarithm of Eq. (11) (all the others remain finite). For larger values of $|\mathbf{s}|$, G diverges over subspaces larger than a single point. Hence the boundary case $|\mathbf{s}|=1/\sqrt{n}$ is taken as representative for a "large" $|\mathbf{s}|$ limit, and the point $\theta_1=\theta_2=\ldots=\theta_n=1/\sqrt{n}$ is taken as the maximum of the cumulant function.

For the bivariate example the largest admissible value is $|\mathbf{s}|=1/\sqrt{2}$, and the maximum of the cumulant function is plotted in Fig. 3, denoted as MCF1, and scaled to PC1. Note that for high values of the probability distribution (e.g. the inner contour), PC1 seems a representative direction of the egg-like shape of the distribution, but for low probabilities it becomes clear that MCF1 is responsible for the large deviations. A point at the end of PC1 is indeed less probable than a point at the end of MCF1.

This result is understood by noting that the joint density (10) is the probability distribution of variables x_i defined as $x_i=z_0+z_i$, where the variables $z_0, z_1, \ldots, z_n$ are independent and Gamma distributed with parameters $\alpha_0, \alpha_1, \ldots, \alpha_n$. Hence, a large deviation of z_0, which occurs independently on others z's, corresponds to a large deviation of $\mathbf{x}$ which is placed on average along the line $x_1=x_2=\ldots=x_n$ (that corresponds to $\theta_1=\theta_2=\ldots=\theta_n$ for the cumulant function).

4 Maximizing the marginal density tail

We have seen that the multivariate cumulant function reduces to the variance if the radius $|\mathbf{s}|$ tends to zero. In that case, its maxima corresponds to the first principal component of data set. If $|\mathbf{s}|$ grows, higher order cumulants come into play, but is not clear what the corresponding maxima represent. In order to clarify this point, we rewrite the cumulant function defined by Eq. (2) in terms of the explicit integral over the probability density

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) = \log\int_{\mathbb{R}^n} f(\mathbf{x})\exp(|\mathbf{s}|\,\boldsymbol{\theta}^t\mathbf{x})d\mathbf{x}$$

This expression can be reduced to an unidimensional integral, by defining the projected data as $Z=\boldsymbol{\theta}^t\mathbf{X}$, and the marginal probability density of the projected data, $f_{\boldsymbol{\theta}}(z)$. The cumulant function is then

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) = \log\int_{-\infty}^{+\infty} f_{\boldsymbol{\theta}}(z)\exp(|\mathbf{s}|z)dz, \quad (13)$$

which corresponds to the cumulant function of the univariate vector $Z=\boldsymbol{\theta}^t\mathbf{X}$ with distribution density $f_{\boldsymbol{\theta}}(z)$. In the light of this representation, our maximization procedure is better understood: we are looking for directions $\boldsymbol{\theta}$ that correspond to a marginal probability density $f_{\boldsymbol{\theta}}(z)$ displaying maximal cumulant function, at fixed $|\mathbf{s}|$. We want to demonstrate that if $|\mathbf{s}|$ grows, a larger cumulant function corresponds to

a marginal density with a fatter tail. Hence our procedure selects the directions corresponding to the marginal densities with fatter tails, where the anomalous behaviour is expected.

Specifically, consider two different directions $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$: we want to demonstrate that if the marginal distribution along $\boldsymbol{\theta}$ has a fatter tail than the distribution along $\boldsymbol{\theta}'$, then the cumulant function has also a fatter tail along $\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}'$. More formally, we have the following theorem.

Theorem 1. *Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be two directions. If there exists a real z^* such that the density distribution $f_{\boldsymbol{\theta}}(z)$ of the random variable $\boldsymbol{\theta}^t\mathbf{X}$ is strictly larger than the density $f_{\boldsymbol{\theta}'}(z)$ of the random variable $\boldsymbol{\theta}'^t\mathbf{X}$, for all $z > z^*$, i.e.*

$$f_{\boldsymbol{\theta}}(z) > f_{\boldsymbol{\theta}'}(z), \quad \textit{for all } z > z^*,$$

then there exists a radius $|\mathbf{s}|^*$ such the cumulant function of $\boldsymbol{\theta}^t\mathbf{X}$ and $\boldsymbol{\theta}'^t\mathbf{X}$ satisfies

$$G_{|\mathbf{s}|}(\boldsymbol{\theta}) > G_{|\mathbf{s}|}(\boldsymbol{\theta}'), \quad \textit{for all } |\mathbf{s}| > |\mathbf{s}|^*.$$

Proof. In order to prove the result, we start by noting that the following inequality holds

$$\exp(|\mathbf{s}|z)\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big] > \exp(|\mathbf{s}|z^*)\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big]$$

for all $z>z^*$, where we have replaced the exponential function with its minimum value in the interval $z\in(z^*, +\infty)$. The inequality holds because the density difference $f_{\boldsymbol{\theta}}(z)-f_{\boldsymbol{\theta}'}(z)$ is positive in this interval, by assumption. Since the above inequality holds in the whole interval $z \in (z^*, +\infty)$, it can be integrated over, i.e.

$$\int_{z^*}^{+\infty} \exp(|\mathbf{s}|z)\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big]dz > \exp(|\mathbf{s}|z^*)\int_{z^*}^{+\infty}\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big]dz. \quad (14)$$

The two densities are normalized, i.e.

$$\int_{-\infty}^{+\infty} f_{\boldsymbol{\theta}}(z)dz = \int_{-\infty}^{+\infty} f_{\boldsymbol{\theta}'}(z)dz = 1$$

Splitting the integrals by z^*, and rearranging terms, the normalization condition is rewritten as

$$\int_{z^*}^{+\infty}\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big]dz = \int_{-\infty}^{z^*}\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz \quad (15)$$

Note that since the left hand side (l.h.s.) is positive, by assumption, the right hand side (r.h.s.) is positive as well. Equation (15) can be substituted in the r.h.s. of the inequality (14), giving

$$\int_{z^*}^{+\infty} \exp(|\mathbf{s}|z)\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big]dz > \exp(|\mathbf{s}|z^*)\int_{-\infty}^{z^*}\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz$$

A lower bound for the r.h.s. can be found, depending on the value of $|\mathbf{s}|$: in the following, we demonstrate that it exists an $|\mathbf{s}|^*$ such that, for all $|\mathbf{s}|>|\mathbf{s}|^*$,

$$\exp(|\mathbf{s}|z^*)\int_{-\infty}^{z^*}\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz > \int_{-\infty}^{z^*}\exp(|\mathbf{s}|z)\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz \quad (16)$$

where the value of $|\mathbf{s}|^*$ must be determined. The theorem is proven once we have demonstrated the last inequality (16), since then we have, for all $|\mathbf{s}|>|\mathbf{s}|^*$,

$$\int_{z^*}^{+\infty}\exp(|\mathbf{s}|z)\big[f_{\boldsymbol{\theta}}(z) - f_{\boldsymbol{\theta}'}(z)\big]dz > \int_{-\infty}^{z^*}\exp(|\mathbf{s}|z)\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz.$$

By rearranging terms, this is equivalent to $G_{|\mathbf{s}|}(\boldsymbol{\theta})>G_{|\mathbf{s}|}(\boldsymbol{\theta}')$. We rewrite the inequality (16) as

$$\int_{-\infty}^{z^*}\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz > \int_{-\infty}^{z^*}\exp\big(|\mathbf{s}|(z-z^*)\big)\big[f_{\boldsymbol{\theta}'}(z) - f_{\boldsymbol{\theta}}(z)\big]dz$$

for all $|\mathbf{s}|>|\mathbf{s}|^*$. Note that even if the integral in the l.h.s. is positive, its integrand, the density difference $f_{\boldsymbol{\theta}'}(z)-f_{\boldsymbol{\theta}}(z)$, is not guaranteed to be positive for all $z\in(-\infty, z^*)$. If it was positive as well, the inequality would hold trivially for all values of $|\mathbf{s}|$, because the exponential in the r.h.s. is smaller then one. This corresponds to the case where additionally to $f_{\boldsymbol{\theta}}(z)>f_{\boldsymbol{\theta}'}(z) \quad \forall z>z^*$, we assume $f_{\boldsymbol{\theta}}(z)<f_{\boldsymbol{\theta}'}(z) \quad \forall z<z^*$. However, we do not need this additional request, and we just note that it would help our procedure by leaving the problem of using a large $|\mathbf{s}|$. The integral in the l.h.s. is independent on $|\mathbf{s}|$, while the integral in the r.h.s. converges to zero for $|\mathbf{s}| \longrightarrow +\infty$, as long as the density difference remains finite, because the exponential tends to zero in the whole interval $z\in(-\infty, z^*)$. If we define $|\mathbf{s}|^*$ as the largest possible value of $|\mathbf{s}|$ for which the two integrals are equal, then for all $|\mathbf{s}|>|\mathbf{s}|^*$ the integral in the l.h.s. is larger than that of r.h.s., and the theorem is proven.

□

5 Discussion

In this paper, we have introduced a novel method selecting the spatial patterns representative for the large deviations in the dataset. The method consists in finding the vectors in the space of data for which the cumulant function is maximal. As in the case of PCA, the spatial patterns are found as normalized directions in the space of data, and a linear projection can be performed, with an easy geometrical interpretation. However, while PCA accounts only for the mass of

the distribution, the cumulant function can give information also about its tails. If one is interested on the large deviations, the projection allows to safely perform Extreme Value Analysis (Coles, 2001). In both cases the subspaces are ordered: in PCA the order follows the fraction of variance of each subspace; The maxima of the cumulant function are ordered by the value of G, expressing the relative importance of each marginal density tail.

Principal components are always symmetric, while large anomalous patterns, if generated by nonlinear processes, are expected to be neither specular nor orthogonal. Accordingly, the maxima of the cumulant function are not necessarily symmetric, since they account for the whole structure of dependencies, and not only covariances. Vector solutions of other nonsymmetric techniques, such as oblique Varimax rotations (Horel, 1981), generally do not covary with the space of data under orthogonal transformations. Hence, solutions depend not only on the shape of the underlying probability distribution, but also on its orientation: an undesirable property for our purposes. The maxima of the cumulant function, instead, covary with the probability density whose shape is the only feature determining the maxima.

In the case of normally distributed data, the maximization of cumulant function yields the first principal component for all values of $|s|$: the elliptically symmetric distribution is characterized by the two tails along the major axis of the ellipse, i.e. the first principal component. When the method is applied to Skew-Normal and Gamma distributions, for non-small $|s|$, the maxima of the cumulant function determine large anomalies: high probability directions far from the center of mass. Note that the limit radius $|s|\rightarrow\infty$ is the innovative key from a technical point of view, allowing for analytical solutions. Using the limit, we were also able to demonstrate that the solutions of our algorithm correspond, in general, to the directions along which the marginal probability density display the fattest tails.

Using the cumulant function is computationally cheap, there is no free parameter, and has the advantage of searching for local solutions, all of which are of interest. When a solution is found, is always a valid local solution, in contrast with neural networks applications, where local solutions are not of interest. In real applications, the radius $|s|$ must be taken as large as possible, until the expected error in the estimate of the cumulant function, due to the finite sample, reach a tolerance value (see Bernacchia et al., 2008). This corresponds to maximize a combination of cumulants which is of the highest reliable order with the given amount of data, accounting for the available set of anomalies.

The solutions of our algorithm are expected to transform continuously as the radius $|s|$ varies. Hence, even if the limit $|s|\rightarrow\infty$ cannot be taken in practice, the solutions for a finite value of $|s|$ are expected to represent a substantial departure from the PCA solution, towards the formal solution at $|s|\rightarrow\infty$. From the theoretical point of view, future work could be devoted to study in detail the nature of solutions at varying $|s|$. For instance, one could attempt to find under which conditions and to what extent the solutions are in between the PCA solution and the formal solution at infinite $|s|$. From the applicative point of view, we expect several datasets to be fruitfully analyzed with our new method (e.g. Bernacchia et al., 2008).

Note that the logarithm is taken for illustrative purposes: the moment generating function could be used instead of the cumulant function, since the maximization is invariant under application of a monotonous function. The present definition is however confortable in avoiding extremely large numbers. Centering of data about the mean is also a practical step, related with the constraint of dealing with finite samples: if the limit $|s|\rightarrow\infty$ could be really taken, the mean would be irrelevant.

Results of our procedure are corrupted if variables are standardized by a rescaling, since the relative scale of different directions is the key in detecting anomalies and comparing the size of tails. If variables are standardized, our procedure reduces to a special case of Independent Component Analyisis (ICA, see Hyvarinen and Oja, 2000), detecting independent components rather than large anomalies. Results are also corrupted if we try to estimate the cumulant function when the underlying probability density decays slower than exponential. In that case, the cumulant function diverge, and the variance of the empirical estimate increases with the size of the sample (Sornette, 2000), implying that the estimate is always unreliable.

Acknowledgements. This work was supported by the european E2-C2 grant, the National Science Foundation (grant: NSF-GMC (ATM-0327936)), by The Weather and Climate Impact Assessment Science Initiative at the National Center for Atmospheric Research (NCAR) and the ANR-AssimilEx project. Finally, we would like to thank I. Bordi, M. Petitta and A. Sutera for valuable discussions, and P. Aires and an anonimous referee for improving the manuscript with their comments.

Edited by: H. Rust
Reviewed by: F. Aires and another anonymous referee

References

Aires, F., Chedin, A., and Nadal, J.: Independent component analysis of multivariate time series: Application to the tropical SST variability, J. Geophys. Res., 105(D13), 17 437–17 455, 2000.

Azzalini, A. and Capitanio, A.: Statistical applications of the multivariate skew-normal distribution, J. Roy. Stat. Soc., B61, 579–602, 1999.

Azzalini, A. and Dalla Valle, A.: The multivariate skew-normal distribution, Biometrika, 83, 715–726, 1996.

Bell, A. J. and Sejnowski, T. J.: An information-maximization approach to blind separation and blind deconvolution, Neural Computation, 7, 1129–1159, 1995.

Bernacchia, A., Naveau, P., Yiou, P., and Vrac, M.: Detecting spatial patterns with the cumulant function – Part 2: An application

to El Niño Nonlin. Processes Geophys., 15, 169–177, 2008, http://www.nonlin-processes-geophys.net/15/169/2008/.

Christiansen, B.: The shortcomings of nonlinear principal component analysis in identifying circulation regimes, J. Climate, 18, 4814–4823, 2005.

Coles, S.: An introduction to statistical modeling of extreme values, Springer, Berlin, 2001.

Gonzalez-Farias, G., Dominguez-Molina, A., and Gupta, A. K.: Additive properties of skew-normal random vectors, J. Stat. Plan. Infer., 126, 521–534, 2004.

Horel, J. D.: A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field, Mon. Weather Rev., 109, 2080, 2092–2092, 1981.

Hsieh, W. W.: Nonlinear multivariate and time series analysis by neural network methods, Rev. Geophys., 42, RG1003, doi:1010.1029/2002RG000112233-243, 2004.

Hyvarinen, A. and Oja, E.: Independent Component Analysis: algorithms and applications, Neural Networks, 13, 411–430, 2000.

Kenney, J. F. and Keeping, E. S.: Cumulants and the Cumulant-Generating Function, in: Mathematics of Statistics, Princeton, NJ, 1951.

Kotz, S., Balakrishnan, N., and Johnson, N. H.: Continuous multivariate distributions, John Wiley and Sons, New York, 1998.

Kramer, M. A.: Nonlinear principal component analysis using autoassociative neural networks, J. Amer. Inst. Chem. Eng., 37, 233–243, 1991.

Malthouse, E. C.: Limitations of nonlinear PCA as performed with generic neural networks, IEEE Trans.NN, 9, 165–173, 1998.

Monahan, A. H., Pandolfo, L., and Fyfe J. C.: The preferred structure of variability of the northern hemisphere atmospheric circulation, Geophys. Res. Lett., 28, 1019–1022, 2001.

Rencher, A. C.: Multivariate statistical inference and applications, John Wiley and Sons, New York, 1998.

Sornette, D.: Critical phenomena in natural sciences, Springer-Verlag, Berlin, 2000.