Nonlin. Processes Geophys., 16, 57–64, 2009
www.nonlin-processes-geophys.net/16/57/2009/

Nonlinear Processes in Geophysics

Information theoretic measures of dependence, compactness, and non-gaussianity for multivariate probability distributions

A. H. Monahan[1] **and T. DelSole**[2]

[1]School of Earth and Ocean Sciences, University of Victoria, Victoria, BC, Canada and Canadian Institute for Advanced Research Earth System Evolution Program, Canada
[2]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia, and Center for Ocean-Land-Atmosphere Studies, Calverton, Maryland, USA

Abstract. A basic task of exploratory data analysis is the characterisation of "structure" in multivariate datasets. For bivariate Gaussian distributions, natural measures of dependence (the predictive relationship between individual variables) and compactness (the degree of concentration of the probability density function (pdf) around a low-dimensional axis) are respectively provided by ordinary least-squares regression and Principal Component Analysis. This study considers general measures of structure for non-Gaussian distributions and demonstrates that these can be defined in terms of the information theoretic "distance" (as measured by relative entropy) between the given pdf and an appropriate "unstructured" pdf. The measure of dependence, mutual information, is well-known; it is shown that this is not a useful measure of compactness because it is not invariant under an orthogonal rotation of the variables. An appropriate rotationally invariant compactness measure is defined and shown to reduce to the equivalent PCA measure for bivariate Gaussian distributions. This compactness measure is shown to be naturally related to a standard information theoretic measure of non-Gaussianity. Finally, straightforward geometric interpretations of each of these measures in terms of "effective volume" of the pdf are presented.

1 Introduction

A fundamental question in exploratory data analysis is: given observations of two variables x_1 and x_2, to what extent is the joint distribution of these variables "interesting", in the sense that it is "structured"? Different kinds of structure can be considered, among which some of the most important are:

I *Dependence*: to what extent does knowledge of x_1 imply knowledge about x_2?

II *Compactness*: to what extent is variance shared between x_1 and x_2; that is, how tightly concentrated around a lower-dimensional surface is the joint probability density function (pdf) $p(x_1, x_2)$?

That these are distinct measures of structure is illustrated by the the pdfs displayed in Fig. 1, all of which by construction have the same total variance var(x_1)+var(x_2). The Gaussian pdf (a) describes uncorrelated variables x_1 and x_2 without any clustering around a lower-dimensional surface; it possesses no structure in either of the senses described above. In contrast, the Gaussian pdf (b) is characterised by more variance along one axis than the other. While x_1 and x_2 are independent, their joint pdf possesses compactness. The Gaussian pdf (c) is equally concentrated around a single axis as is (b), but in such a way that the variables x_1 and x_2 are correlated and thus also characterised by dependence.

For bivariate Gaussian distributions with pdf

$$p(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right) \tag{1}$$

(assumed without loss of generality to be mean zero), where $\mathbf{x}^T = (x_1, x_2)$ and the covariance matrix Σ is defined as

$$\Sigma = \int \mathbf{x}\mathbf{x}^T p(\mathbf{x}) d^2\mathbf{x} = \begin{pmatrix} \sigma_{x_1}^2 & \rho\sigma_{x_1}\sigma_{x_2} \\ \rho\sigma_{x_1}\sigma_{x_2} & \sigma_{x_2}^2 \end{pmatrix}, \tag{2}$$

then measures of structure associated with dependence and compactness are associated respectively with ordinary least-squares regression (OLS) and principal component analysis (PCA) (the second of which is closely related to orthogonal least squares regression). For OLS, the natural measure of structure is ρ^2, the fraction of variance "explained" by the

Correspondence to: A. H. Monahan
(monahana@uvic.ca)

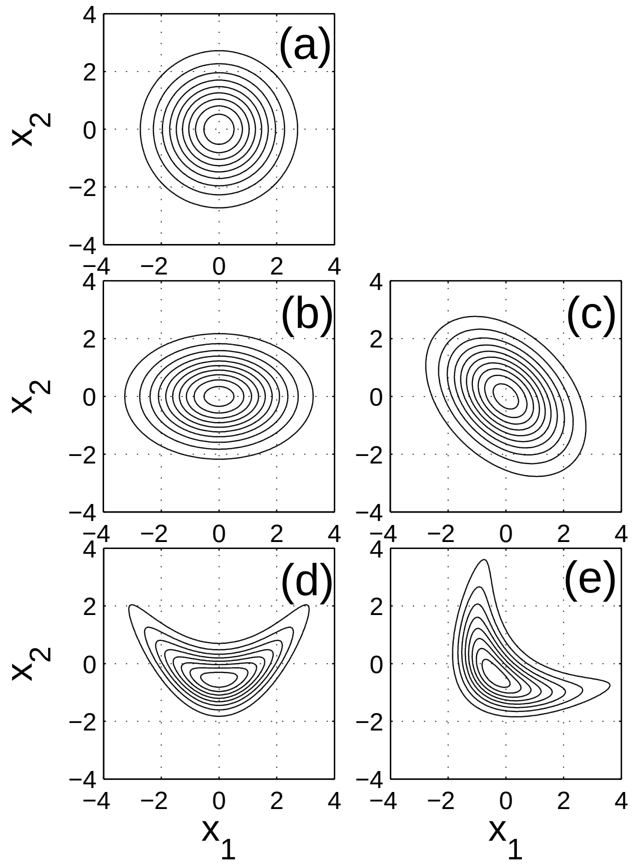

Fig. 1. Contours of joint pdfs of x_1, x_2 displaying different degrees of "structure". **(a)** Gaussian pdf with no compactness or dependence between x_1 and x_2. **(b)** Gaussian pdf displaying compactness but not dependence. **(c)** Gaussian pdf displaying both compactness and dependence. **(d)**, **(e)** non-Gaussian pdfs displaying both compactness and dependence. The variables x_1 and x_2 are uncorrelated in (d) and correlated in (e). Pdfs (b) and (d) have the same covariance matrix, and would not be distinguished by traditional linear measures of dependence and compactness; the same is true of (c) and (e). All pdfs have the same total variance $\text{var}(x_1)+\text{var}(x_2)=\text{Tr}\Sigma$.

Table 1. Measures of dependence, compactness, and non-Gaussianity for the distributions contoured in Fig. 1. ρ^2 is the fraction of variance explained by ordinary least-squares regression, $F(p)$ is the fraction of variance explained by the first PCA mode, $M(p)$ is the generalised measure of dependence (Eq. 16), $\mathcal{C}(p)$ is the generalised measure of compactness (Eq. 22), $S(p)$ is the compactness measure transformed to correspond to $F(p)$ for a bivariate Gaussian (Eq. 27), and $\nu(p)$ is the measure of non-Gaussianity (Eq. 31).

	ρ^2	$F(p)$	$M(p)$	$\mathcal{C}(p)$	$S(p)$	$\nu(p)$
(a)	0	0.5	0	0	0.5	0
(b)	0	0.69	0	0.15	0.69	0
(c)	0.15	0.69	0.15	0.15	0.69	0
(d)	0	0.69	0.55	0.69	0.91	0.63
(e)	0.15	0.69	0.37	0.69	0.91	0.63

regression; for PCA it is F, the fraction of variance explained by the first PCA mode:

$$F=\frac{\lambda_1}{\text{Tr}\Sigma}, \tag{3}$$

where λ_1 is the larger eigenvalue of Σ. Values of ρ^2 and F for pdfs (a)–(c) are given in Table 1.

When the joint distribution $p(x_1, x_2)$ is not Gaussian, the issue of characterising structure is more subtle. Panel (d) in Fig. 1 contours the pdf

$$p(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{x_1^2}{2\sigma_1^2}-\frac{(x_2-a(x_1^2-\sigma_1^2))^2}{2\sigma_2^2}\right) \tag{4}$$

(for which $\text{mean}(x_1)=\text{mean}(x_2)=0$, $\text{var}(x_1)=\sigma_1^2$, $\text{var}(x_2)=2a^2\sigma_1^4+\sigma_2^2$), where the parameters (a, σ_1, σ_2) have been chosen so that the pdfs in (b) and (d) have the same covariance matrices. It is evident that the pdf (d) is concentrated around a low-dimensional (nonlinear) curve, and therefore also is characterised by compactness. In fact, visual inspection suggests that the degree of concentration of the pdf (and therefore of compactness) is greater for (d) than for (b), but the traditional linear measure of compactness F would not distinguish between them. Furthermore, x_1 and x_2 in (d) are dependent, despite being uncorrelated: strongly positive and negative values of x_1 are associated with strongly positive values of x_2. The traditional linear measure of dependence ρ^2 does not characterise this structure. The compactness of pdf (d) has contributions from both the anisotropy of the covariance matrix (shared with pdf (b)) and from the degree of non-Gaussianity. In order to tease these apart, it is desirable to also define a third measure of "interesting" structure:

III *Non-Gaussianity*: to what extent does the joint distribution differ from a bivariate Gaussian?

Such a measure would allow the determination for a given pdf of the relative contribution to compactness of non-Gaussianity and covariance anisotropy.

The pdf (e) (constructed by rotating the pdf (d) through 45°) has the same covariance matrix as (c); again, these pdfs would not be distinguished by the measures ρ^2 and F. By inspection, the degree of compactness of pdf (e) is the same as that of pdf (d): the degree of concentration of a pdf around a lower-dimensional curve should not depend on its orientation. However, the degree of dependence between x_1 and x_2 has changed relative to (d) (for example, the conditional pdf $p(x_2|x_1)$ is much tighter for $x_1>0$ than for $x_1<0$). This example further illustrates the fact that the ideas of compactness and dependence are distinct. Finally, the fact that the discussion of dependence, compactness, and non-Gaussianity can be framed in terms of the plots in Fig. 1 suggests that measures of each of these should have straightforward geometrical interpretations.

The above discussion motivates the consideration of measures of compactness that are invariant under orthogonal rotations, and which reduce to PCA for the case of a bivariate Gaussian; of measures of dependence which are *not* invariant under rotation and which reduce to ordinary least-squares regression for the case of a bivariate Gaussian; and of measures of non-Gaussianity. The notion of "interesting" structure of course is a relative concept, and can only be measured relative to specified "uninteresting" distributions. In the construction of these measures, we are thus confronted with the need to measure the difference between two pdfs: one data-driven, the other some specified background reference. A natural framework for measuring the difference between two pdfs is provided by information theory (e.g. Cover and Thomas, 1991; Majda et al., 2005), through which such differences can be related to the new "information" provided by the data-driven pdf relative to the background pdf.

In fact, a well-known general measure of dependence is provided by information theory: this is multiinformation (which in two dimensions is also known as mutual information). Similarly, information theory provides a natural measure of non-Gaussianity known as negentropy. Less well-established is a general measure of compactness. Previous approaches to this problem, using tools such as Nonlinear Principal Component Analysis (e.g. Monahan et al., 2003), have been hampered by the lack of a rigorous theoretical framework and by the methodological difficulties of nonlinear nonparametric function estimation.

The goal of the present study is to further develop measures of "interesting" structure for general non-Gaussian pdfs that can provide a rigorous basis for non-Gaussian exploratory data analysis. In particular, we will propose a general measure of compactness with firm foundations in information theory and which reduces to PCA for bivariate Gaussians. This measure will be contrasted with the well-established measures of dependence and non-Gaussianity provided by mutual information and negentropy. The measure of compactness will be seen to be a combined measure of Gaussianity and covariance isotropy, and therefore to have a natural connection to the standard information theoretic measure of non-Gaussianity. This discussion presents a unifying notion of "structure" in probability distributions: each of the measures of dependence, compactness, and non-Gaussianity are defined in terms of the information theoretic "distance" (as measured by relative entropy) between the given pdf and the appropriate "unstructured" pdf. Finally, it will be shown that these measures have natural geometrical interpretations in terms of the "effective volumes" of the associated probability distributions. A similar measure of compactness was introduced in Peña and van der Linde (2007); the present study demonstrates the connection of the compactness measure to PCA, emphasizes the fundamental difference between it and mutual information as measures of structure, and illustrates how all of these measures of structure can be expressed as relative entropies. There has been considerable recent interest in information theoretic measures of predictability in geophysical systems (e.g. DelSole, 2004; Kleeman and Majda, 2005; DelSole and Tippett, 2007); the present study considers the applicability of these ideas to exploratory data analysis. This study does not address the problem of estimating these measures from finite datasets: since the proposed measures apply to non-Gaussian data the estimation problem is considerably more difficult than those of the corresponding Gaussian measures, as the underlying pdfs are not known a priori to be parameterised with a finite set of coefficients. Nevertheless, one must have a clear idea of what constitutes "interesting structure" without regard to estimation questions before complexities due to finite data can be addressed.

2 Information theoretic entropy

A natural starting point for the characterisation of the structure of the pdf $p(\mathbf{x})$ of an N-dimensional random variable $\mathbf{x}$ is the entropy

$$H(p) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \tag{5}$$

(e.g. Cover and Thomas, 1991). This quantity arises naturally as the measure of the "information content" of a pdf, and is characterised by the following properties relevant to the discussion of "structure":

1. Under a diffeomorphic coordinate transformation $\mathbf{x}\rightarrow\mathbf{x}'=\mathbf{G}(\mathbf{x})$,

   $$H(p(\mathbf{x}')) = H(p(\mathbf{x})) - \int p(\mathbf{x}) \ln \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{x}'} \right) \right| d\mathbf{x} \tag{6}$$

   (Majda et al., 2002). In particular, under a linear rescaling of each variable: $x_i'=a_i x_i$,

   $$H(p(\mathbf{x}')) = H(p(\mathbf{x})) + \sum_{i=1}^{N} \ln |a_i|, \tag{7}$$

   and under a unitary transformation $\mathbf{x}'=\mathcal{U}\mathbf{x}$, H is invariant

   $$H(p(\mathbf{x}')) = H(p(\mathbf{x})). \tag{8}$$

2. We have the inequality:

   $$H(p) \leq \frac{1}{2} \ln[(2\pi e)^N \det \Sigma] = H(p_G), \tag{9}$$

   where $H(p_G)$ is the entropy of a Gaussian random variable with the same covariance matrix as $p(\mathbf{x})$ (e.g.

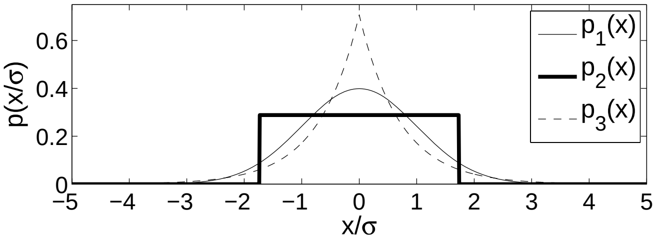

Fig. 2. Three pdfs of equal variance but differing entropy: $H(p_2)<H(p_3)<H(p_1)$. The Gaussian pdf has maximum entropy among all pdfs with the same variance.

Cover and Thomas, 1991). Thus, of all distributions with the same covariance matrix, the Gaussian has the largest entropy (and so is minimally "informative"). This point deserves further comment. The three univariate pdfs (illustrated in Fig. 2)

$$p_1(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{10}$$

$$p_2(x) = \begin{cases} \frac{1}{2\sqrt{3}\sigma} & |x| \le \sqrt{3}\sigma \\ 0 & |x| > \sqrt{3}\sigma \end{cases} \tag{11}$$

$$p_3(x) = \frac{1}{\sqrt{2}\sigma}\exp\left(-\frac{\sqrt{2}|x|}{\sigma}\right) \tag{12}$$

are each of variance σ^2 but of respective entropies: $H(p_1)=\ln(\sqrt{2\pi e}\sigma)$, $H(p_2)=\ln(\sqrt{12}\sigma)$, and $H(p_3)=\ln(\sqrt{2e^2}\sigma)$.

The entropy of the Gaussian distribution $p_1(x)$ is larger than those of the other two distributions, and so it is less "informative" than either: the boxcar distribution $p_2(x)$ because it does not display long tails, and the exponential distribution $p_3(x)$ because it is sharply peaked around $x=0$. The Gaussian distribution combines sufficient flatness around its median value with sufficiently thick tails to be maximally entropic (that is, minimally informative).

3. A pdf is said to be sphered if all of the eigenvalues of its covariance matrix are equal; that is, if it its covariance matrix is proportional to the identity matrix (note that while the covariance matrix of a sphered pdf is invariant under rotation, the pdf itself is not necessarily isotropic). Given a pdf $p(\mathbf{x}')$ with covariance Σ obtained from the sphered pdf $p_S(\mathbf{x})$ by a linear rescaling of the coordinate axes such that the total variance $\mathrm{Tr}\Sigma$ is fixed, then $H(p_S)\ge H(p)$. That is, the entropy is maximised by the sphered pdf among all pdfs related by linear rescalings of the axes such that the total variance is maintained. To see this result, fix the matrix Σ (with eigenvalues λ_i) and consider the sphered pdf $p_S(\mathbf{x})$ with covariance matrix $(\mathrm{Tr}\Sigma/N)\mathsf{I}_N$, where I_N is the identity matrix. The pdf $p(\mathbf{x}')$ with covariance matrix Σ is obtained through a linear transformation $x_i'=\sqrt{\gamma_i}x_i$, where $\gamma_i=N\lambda_i/\mathrm{Tr}\Sigma$ and the x_i are aligned along the eigenvectors of Σ. From Eq. (7), it follows that

$$\begin{aligned} H(p) &= H(p_S) + \frac{1}{2}\sum_{i=1}^{N}\ln\frac{N\lambda_i}{\mathrm{Tr}\Sigma} \\ &= H(p_S) + \frac{N}{2}\ln\left[\frac{N}{\mathrm{Tr}\Sigma}(\det\Sigma)^{1/N}\right] \\ &\le H(p_S), \end{aligned} \tag{13}$$

where the desired result (given by the final inequality) follows from the arithmetic-geometric inequality:

$$\frac{\mathrm{Tr}\Sigma}{N} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i \ge \left(\prod_{i=1}^{N}\lambda_i\right)^{1/N} = (\det\Sigma)^{1/N}, \tag{14}$$

(where equality holds when all λ_i are equal). Thus, if a pdf is stretched along some axes and compressed along others such that the total variance is unchanged, then the pdf with maximum entropy arises when all axes carry equal variance.

Note that it follows from this and the previous property that of all pdfs with the same total variance, the entropy is maximised by a sphered Gaussian. This fact can be proved directly using standard maximum entropy methods (e.g. Cover and Thomas, 1991); properties 2 and 3 have been presented separately in order to highlight the distinction between spheredness and Gaussianity in the context of maximum entropy distributions.

A natural measure of the difference between two pdfs $p(\mathbf{x})$ and $q(\mathbf{x})$ is the relative entropy:

$$D(p||q) = \int p(\mathbf{x})\ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)d\mathbf{x} \tag{15}$$

(Cover and Thomas, 1991). This quantity is non-negative (taking the value of zero only if $p=q$) and is invariant under an arbitrary invertible coordinate transformation $\mathbf{x}\Rightarrow\mathbf{x}'=\mathbf{G}(\mathbf{x})$. While relative entropy is not a Euclidean distance measure (in particular, it is not symmetric: $D(p||q)\neq D(q||p)$), it is a useful measure of the difference between two pdfs. The measures of dependence, compactness, and non-Gaussianity to which we now turn will each be defined in terms of the relative entropy between the given pdf and an appropriate "unstructured" pdf.

3 Measures of structure: dependence

By definition, the components of the random variable $\mathbf{x}$ are independent if and only if their joint distribution factors as the product of the marginals: $p(\mathbf{x}) = \prod_{i=1}^N p_{x_i}(x_i)$. The well-known result follows that a natural measure of *dependence* in a multivariate pdf is the *multiinformation* (Schneidman et al., 2003)

$$\begin{aligned} I(\mathbf{x}) &= D(p(\mathbf{x}) || \prod_{i=1}^N p_{x_i}(x_i)) \\ &= \sum_{i=1}^N H(p_{x_i}(x_i)) - H(p(\mathbf{x})), \end{aligned} \tag{16}$$

where the second equality follows from the definitions of marginal distributions and of entropy. It follows that the quantity

$$\begin{aligned} M(p) &= 1 - e^{-2I(\mathbf{x})} \\ &= 1 - \left(\exp\left[H(p(\mathbf{x})) - \sum_{i=1}^N H(p_{x_i}(x_i)) \right] \right)^2 \\ &= 1 - \left(\frac{\exp(H(p(\mathbf{x})))}{\prod_{i=1}^N \exp(H(p_{x_i}(x_i)))} \right)^2 \end{aligned} \tag{17}$$

is a measure taking values between 0 and 1, with $M(\mathbf{x})=0$ when the x_i are mutually independent and $M(\mathbf{x})=1$ when at least two variables are fully dependent (that is, $x_j = f(x_1, ..., x_{j-1}, x_{j+1}, ..., x_N)$ for some j). For the measure of dependence, the "unstructured" pdf against which the given pdf is compared is given by the product of the marginals along each x_i which by construction has no dependence among any of the variables.

For the bivariate case ($N=2$), $I(\mathbf{x})$ is known as the mutual information (e.g. Cover and Thomas, 1991). For a bivariate Gaussian, it is well known that

$$I(x_1, x_2) = -(1/2)\ln(1-\rho^2), \tag{18}$$

where ρ is the correlation coefficient between x_1 and x_2, from which it follows that

$$M(p) = 1 - \exp(-2I(x_1, x_2)) = \rho^2 \tag{19}$$

In the limit that $p(x_1, x_2)$ is bivariate Gaussian, then, $M(p)$ corresponds to the fraction of variance accounted for by an ordinary least-squares regression between x_1 and x_2.

The integral (16) defining mutual information is invariant under an arbitrary coordinate transformation, and therefore might be considered to also be a natural general measure of compactness. In fact, the mutual information is *not* invariant under an orthogonal rotation of $\mathbf{x}$. This is most easily seen in the context of a Gaussian distribution, for which the correlation coefficient ρ^2 is not invariant under rotations: in particular, under a rotation of (x_1, x_2) such that the coordinate axes are aligned with the principal component axes, the correlation coefficient vanishes. The resolution of this apparent paradox is that while the integral in Eq. (16) is invariant under the unitary transformation $\mathbf{x} \rightarrow \mathbf{x}' = \mathcal{U}\mathbf{x}$, the integral does not retain its identity as mutual information. This is because under the rotation the product of the marginal distributions of the original variables, $p_{x_1}(x_1)p_{x_2}(x_2)$ is not transformed into the product of the marginals of the rotated variables, $p_{x'_1}(x'_1)p_{x'_2}(x'_2)$ (a detailed discussion of this point in the context of a bivariate Gaussian distribution is presented in Appendix A). Like ρ^2, mutual information is not invariant under a unitary transformation that mixes the two variables: in general, $M(p(\mathbf{x}')) \neq M(p(\mathbf{x}))$. Mutual information (and more generally multiinformation) therefore does not provide the desired compactness measure, to which we now turn.

4 Measures of structure: compactness

As was discussed in the Introduction, we seek a measure of compactness of multivariate distributions; that is, a measure of the extent to which the full distribution is concentrated around a lower-dimensional surface. Such a measure should be invariant under unitary transformations (the degree of concentration should not depend on the orientation of the distribution in state space). The dependence measure $M(p)$ is not such a measure, as it is not invariant under unitary transformations.

We suggest measuring compactness based on the degree to which $p(\mathbf{x})$ differs from a sphered Gaussian with the same total variance $\mathrm{Tr}\Sigma$. The pdf of such an equivalent sphered Gaussian is

$$p_{SG}(\mathbf{x}) = \left(\frac{N}{2\pi \mathrm{Tr}\Sigma} \right)^{N/2} \exp\left(-\frac{N}{2\mathrm{Tr}\Sigma} \mathbf{x}^T \mathbf{x} \right), \tag{20}$$

from which it follows that

$$\begin{aligned} D(p||p_{SG}) &= -H(p) + H(p_{SG}) \\ &= -H(p) + \frac{N}{2} \ln\left(2\pi e \frac{\mathrm{Tr}\Sigma}{N} \right) \end{aligned} \tag{21}$$

(note that the relative entropy can be expressed as a difference between two entropies as a consequence of the special form of Eq. 20). This measure vanishes for a sphered Gaussian and is never negative. In analogy with Eq. (16), we define the compactness of $p(\mathbf{x})$ as

$$\mathcal{C}(p) = 1 - e^{-2D(p||p_{SG})} = 1 - \left(\frac{\exp(H(p))}{\exp(H(p_{SG}))} \right)^2, \tag{22}$$

which is bounded between 0 and 1, vanishes for a sphered Gaussian, and is invariant under unitary transformations. The compactness measure can be factored as

$$\mathcal{C}(p) = 1 - \left(\frac{e^{H(p)}}{(2\pi e)^{N/2}\sqrt{\det \Sigma}} \right)^2 \left(\frac{N(\det \Sigma)^{1/N}}{\mathrm{Tr}\Sigma} \right)^N . \tag{23}$$

The first factor in parentheses is the exponential of the ratio of the entropy of $p(\mathbf{x})$ to that of a Gaussian distribution with the same covariance matrix; by inequality Eq. (9), this ratio is bounded between zero and one. The second factor in parentheses is bounded between zero and one by the inequality Eq. (14) and the fact that the eigenvalues of Σ are all non-negative, such that the ratio achieves its maximum value when all eigenvalues of Σ are equal (i.e. when the distribution is sphered). This factorisation illustrates that our proposed measure of compactness is fundamentally a combined measure of Gaussianity and covariance isotropy.

For a bivariate Gaussian distribution, $\mathcal{C}(p)$ reduces to:

$$\mathcal{C}(p) = 1 - \frac{4\det\Sigma}{(\mathrm{Tr}\Sigma)^2} \tag{24}$$

For such a distribution, the classical measure of compactness is the fraction of variance accounted for by the first principal component. This measure is expressed mathematically as:

$$F(p) = \frac{\lambda_1}{\mathrm{Tr}\Sigma}. \tag{25}$$

The measure

$$S(p) = \frac{1}{2}\left(1 + \sqrt{\mathcal{C}(p)}\right) \tag{26}$$

is a general measure of compactness that in the limit of $p(\mathbf{x})$ Gaussian reduces to $F(p)$. The quadratic equation for λ_1 following from the facts that $\det\Sigma=\lambda_1\lambda_2$ and $\mathrm{Tr}\Sigma=\lambda_1+\lambda_2$ can be solved to yield

$$\begin{aligned} F(p) &= \frac{1}{2}\left(1 + \sqrt{1 - \frac{4\det\Sigma}{(\mathrm{Tr}\Sigma)^2}}\right) \\ &= \frac{1}{2}\left(1 + \sqrt{\mathcal{C}(p)}\right) = S(p). \end{aligned} \tag{27}$$

In the same way that for a bivariate Gaussian $M(p)$ had a straightforward relationship to the fraction of variance explained by an ordinary least-squares regression, for a bivariate Gaussian $\mathcal{C}(p)$ is naturally related to the fraction of variance explained by the first PCA mode.

5 Measures of structure: non-gaussianity

The compactness measure $\mathcal{C}(p)$ combined measures of covariance isotropy and Gaussianity and therefore cannot distinguish between situations in which the measure is large because (a) the pdf is Gaussian (or nearly so), such that the variance of the first principal component is much larger than that of the second, or because (b) the pdf is narrowly distributed around a nonlinear curve. For this, a direct measure of non-Gaussianity is needed; such a measure is the *negentropy* (Lee et al., 2000), defined as the relative entropy between $p(\mathbf{x})$ and the Gaussian pdf with the same covariance matrix:

$$D(p||p_G) = H(p_G) - H(p_{SG}) + H(p_{SG}) - H(p) \tag{28}$$

$$= \frac{1}{2}\ln\left(\frac{\det\Sigma}{1-\mathcal{C}(p)}\left(\frac{N}{\mathrm{Tr}\Sigma}\right)^N\right), \tag{29}$$

where

$$p_G(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right) \tag{30}$$

and Eq. (28) follows because $p_G(\mathbf{x})$ is Gaussian.

Defining

$$\nu(p) = 1 - e^{-2D(p||p_G)} = 1 - \frac{1-\mathcal{C}(p)}{\det\Sigma}\left(\frac{\mathrm{Tr}\Sigma}{N}\right)^N, \tag{31}$$

we obtain a measure taking values between 0 and 1, with $\nu(p)=0$ if and only if $p(\mathbf{x})$ is Gaussian (as by construction both $p(\mathbf{x})$ and $p_G(\mathbf{x})$ have the same covariance matrix) and $\nu(p)$ increasing as $p(\mathbf{x})$ becomes increasingly non-Gaussian. Note that $\nu(p)$ contains contributions from both the compactness of the pdf and the degree of covariance anisotropy; for a sphered distribution $\mathrm{Tr}\Sigma/N=(\det\Sigma)^{1/N}$ and $\nu(p)=\mathcal{C}(p)$. Furthermore, for Σ fixed, $\nu(p)$ increases as $\mathcal{C}(p)$ increases: among all distributions with the same covariance, the more compact distributions are the more non-Gaussian.

6 Measures of structure: geometric interpretation

The quantity

$$V(p) = \exp(H(p)) \tag{32}$$

is an extensive variable which can be interpreted as a measure of the “effective volume” of a pdf. For instance, for a Gaussian distribution $p_G(\mathbf{x})$ with covariance matrix Σ,

$$V(p_G) = e^{H(p)} = e^{N/2}|\Sigma|^{1/2}(2\pi)^{N/2}. \tag{33}$$

The volume enclosed by a surface of constant probability density for the same distribution is

$$\begin{aligned} V_E(\alpha) &= \text{Volume}[\mathbf{x} : \mathbf{x}^T\Sigma\mathbf{x} \le \alpha] \\ &= \alpha^{N/2}|\Sigma|^{1/2}\left(\frac{\pi^{N/2}}{\Gamma(N/2+1)}\right). \end{aligned} \tag{34}$$

Comparing these two expressions shows that, aside from factors that depend only on the dimension of the space, $V(p)$ is related to the geometric volume of the isoprobability ellipsoid. More generally, $V(p)$ is the volume of a “typical set”, as reviewed in Cover and Thomas (1991). Because of inequalities Eqs. (9) and (13), the pdf with maximum volume for a given covariance matrix is Gaussian, and the pdf with maximum volume for given total variance is a sphered Gaussian.

The measures of dependence, compactness, and non-Gaussianity introduced above have natural interpretations in terms of effective volumes (in the sense of Eq. 32):

$$M(p) = 1 - \left(\frac{V(p(\mathbf{x}))}{V(\prod_{i=1}^N p_{x_i}(x_i))}\right)^2 \tag{35}$$

$$\mathcal{C}(p) = 1 - \left(\frac{V(p)}{V(p_{SG})}\right)^2 \tag{36}$$

$$\nu(p) = 1 - \left(\frac{V(p)}{V(p_G)}\right)^2. \tag{37}$$

That is:

- $M(p)$ is one less the square of the ratio of two effective volumes: that of the full pdf, and that of the pdf produced by the product of the marginals. Dependence among the variables x_i implies a concentration of probability around some lower-dimensional surface, with an associated reduction in $V(p)$ and an increase in $M(p)$.

- $\mathcal{C}(p)$ is one less the square of the ratio of two volumes: that of the full pdf, and that of the equivalent sphered Gaussian. Similarly to $M(p)$, $\mathcal{C}(p)$ is a measure of the degree to which the pdf $p(\mathbf{x})$ clusters around a low-dimensional surface; but unlike $M(p)$, $\mathcal{C}(p)$ is rotationally invariant as the effective volume of $p_{SG}(\mathbf{x})$ does not change under a coordinate rotation (in contrast to the effective volume of the product of the marginals).

- $\nu(p)$ is one less the square of the ratio of the effective volume of the full pdf to that of the Gaussian with the same covariance matrix.

In general, the degree of structure in a pdf increases as the effective volume decreases relative to that of the "unstructured" pdf against which it is compared. This result provides a useful geometrical interpretation of the measures of structure.

7 Conclusions

This study has considered three measures of structure for multivariate datasets, all defined in terms of the relative entropy (the information-theoretic distance) between a given pdf $p(\mathbf{x})$ and an appropriate "unstructured" pdf.

- *Dependence* is measured in terms of the relative entropy between $p(\mathbf{x})$ and the pdf $q(\mathbf{x})=\Pi_{i=1}^{N} p_{x_i}(x_i)$ consisting of the product of the marginal distributions along each individual component x_i of $\mathbf{x}$

- *Compactness* is measured in terms of the relative entropy between $p(\mathbf{x})$ and the equivalent sphered Gaussian $p_{SG}(\mathbf{x})$ (the Gaussian with the same total variance but equal variance along each coordinate direction). This is a combined measure of Gaussianity and covariance isotropy, and is invariant under an orthogonal rotation of the variables.

- *Non-Gaussianity* is measured in terms of the relative entropy between $p(\mathbf{x})$ and the equivalent Gaussian (the Gaussian with the same covariance matrix). This measure has a natural connection with the measure of compactness.

All of these measures admit useful geometrical interpretations in terms of the ratio of the "effective volume" of the pdf to that of the associated "unstructured" pdf against which it is compared.

The dependence measure $M(p)$ is not invariant under an orthogonal rotation of the variable vector $\mathbf{x}$, despite the fact that the integral defining it is in fact invariant. This study has demonstrated that this apparent paradox is resolved by the fact that under the rotation the integral no longer retains the identity of the dependence measure (as the rotated product of marginal distributions is not the product of the rotated marginals).

Table 1 presents values of the various measures of dependence, compactness, and non-Gaussianity considered in this study for the distributions in Fig. 1. For the Gaussian distributions, the dependence measure $M(p)$ (Eq. 16) and compactness measure $\mathcal{C}(p)$ (Eq. 22) coincide with the corresponding measures from ordinary and orthogonal least-squares regression, as expected. For the non-Gaussian distributions, the new measures are larger, demonstrating their better characterisation of dependence and compactness relative to that of their Gaussian counterparts. Note that the compactness of (b)–(e) is measured through comparison with (a); visual inspection demonstrates that (b)–(e) are all more tightly concentrated around a lower dimensional curve (and are therefore have smaller "effective volume") than is (a). Non-Gaussianity of (d) or (e) is measured through comparison with (b) or (c), respectively; it is evident from inspection of (d) that the same probability mass is concentrated in smaller volume in (d) than in (b) [and similarly for (e) and (c)], consistent with the geometric interpretation of our measure of non-Gaussianity.

The measures of dependence, compactness, and non-Gaussianity considered in this study are defined by the distance between the given pdf and an appropriate reference pdf, as measured by the relative entropy. Many other distance measures between pdfs have been proposed, such as Bregman's distance, Bhattacharyya distance, the chi-squared statistic, and the Kolmogorov-Smirnov distance (e.g. Pardo, 2006). Despite the availability of a wide class of measures, we feel that the measures that we have proposed are especially attractive because they connect to more traditional measures used in geophysics (e.g. the fraction of variance explained by least-squares regression or PCA).

For bivariate distributions, the information theoretic measures of dependence and compactness considered in this study are generalisations of the corresponding measures of covariability obtained from the classical linear measures provided with ordinary least-squares regression and Principal Component Analysis. A fundamental challenge with the use of these information theoretic measures in exploratory data analysis is their estimation from a finite sample. Estimators of the measures of structure themselves, as well as the associated sampling error, are required for their practical application. Classical hypothesis testing (e.g. determining if one of the proposed measures is significantly different from zero) will require the development of parametric or non-parametric techniques for computing confidence intervals which is beyond the scope of the present study. The estimation problem for information theoretic measures is an active field of research (e.g. Kleeman and Majda, 2005; Haven et al., 2005); we are confident that as robust estimators become available, the measures of dependence, compactness, and non-Gaussianity discussed in this study will demonstrate their utility as practical tools for exploratory data analysis in geophysical data sets.

Appendix A

Change of mutual information under rotation

Suppose that the distribution of $\mathbf{x}$ is bivariate Gaussian with mean zero and covariance matrix Eq. (2). Under the rotation

$$\mathcal{U} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \tag{A1}$$

the transformed variable $\mathbf{x}'=\mathcal{U}\mathbf{x}$ is Gaussian with covariance matrix

$$\Sigma' = \mathcal{U}\Sigma\mathcal{U}^T = \begin{pmatrix} \sigma_{x'}^2 & \rho'\sigma_{x'}\sigma_{y'} \\ \rho'\sigma_{x'}\sigma_{y'} & \sigma_{y'}^2 \end{pmatrix} \tag{A2}$$

where

$$\sigma_{x'}^2 = \cos^2\phi\sigma_x^2 - 2\cos\phi\sin\phi\rho\sigma_x\sigma_y + \sin^2\phi\sigma_y^2 \tag{A3}$$

$$\sigma_{y'}^2 = \sin^2\phi\sigma_x^2 + 2\cos\phi\sin\phi\rho\sigma_x\sigma_y + \cos^2\phi\sigma_y^2 \tag{A4}$$

$$\rho' = \frac{\sin\phi\cos\phi(\sigma_x^2-\sigma_y^2) + (\cos^2\phi - \sin^2\phi)\rho\sigma_x\sigma_y}{\sigma_{x'}\sigma_{y'}}. \tag{A5}$$

The product of the marginal distributions in the untransformed coordinates is

$$q(\mathbf{x}) = p_x(x)p_y(y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\mathbf{x}^T C^{-1}\mathbf{x}\right), \tag{A6}$$

which is Gaussian with covariance matrix

$$C = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}. \tag{A7}$$

Under the coordinate transformation, $q(\mathbf{x})$ remains Gaussian with new covariance matrix

$$C' = \begin{pmatrix} \cos^2\phi\sigma_x^2 + \sin^2\phi\sigma_y^2 & \cos\phi\sin\phi(\sigma_x^2-\sigma_y^2) \\ \cos\phi\sin\phi(\sigma_x^2-\sigma_y^2) & \sin^2\phi\sigma_x^2 + \cos^2\phi\sigma_y^2 \end{pmatrix}. \tag{A8}$$

Clearly, C' is *not* the covariance matrix of the product of the marginals in the transformed coordinate system:

$$\tilde{C} = \begin{pmatrix} \sigma_{x'}^2 & 0 \\ 0 & \sigma_{y'}^2 \end{pmatrix} \tag{A9}$$

with $\sigma_{x'}$ and $\sigma_{y'}$ given by Eqs. (A3) and (A4). That is, the transformed product of the marginal distributions is not equal to the product of the transformed marginal distributions. While the integral defining mutual information is invariant under an orthogonal rotation mixing variables, its identity as the mutual information is lost.

Acknowledgements. AHM acknowledges support from the Natural Sciences and Engineering Research Council of Canada and from the Canadian Institute for Advanced Research Earth System Evolution Program. TD acknowledges supported from the National Science Foundation (ATM0332910), National Aeronautics and Space Administration (NNG04GG46G), and the National Oceanographic and Atmospheric Administration (NA04OAR4310034). The authors would also like to thank three anonymous reviewers whose helpful comments improved this manuscript.

Edited by: H. A. Dijkstra
Reviewed by: three anonymous referees

References

Cover, T. M. and Thomas, J. A.: Elements of Information Theory, John Wiley & Sons, Inc., New York, 1991.

DelSole, T.: Predictability and Information Theory. Part I: Measures of predictability, J. Atmos. Sci., 61, 2425–2440, 2004.

DelSole, T. and Tippett, M. K.: Predictability: Recent Insights from Information Theory, Rev. Geophys., 45, RG4002, doi:10.1029/2006RG000202, 2007.

Haven, K., Majda, A., and Abramov, R.: Quantifying predictability through information theory: Small sample estimation in a non-Gaussian framework, J. Comp. Phys., 206, 334–362, 2005.

Kleeman, R. and Majda, A.: Predictability in a model of geophysical turbulence, J. Atmos. Sci., 62, 2864–2879, 2005.

Lee, T.-W., Girolami, M., Bell, A., and Sejnowski, T.: A unifying information-theoretic framework for independent component analysis, Comp. Math. App., 39, 1–21, 2000.

Majda, A., Kleeman, R., and Cai, D.: A Mathematical Framework for Quantifying Predictability Through Relative Entropy, Meth. Appl. Anal., 9, 425–444, 2002.

Majda, A. J., Abramov, R. V., and Grote, M. J.: Information Theory and Stochastics for Multiscale Nonlinear Systems, American Mathematical Society, Providence, RI, USA, 2005.

Monahan, A. H., Fyfe, J. C., and Pandolfo, L.: The vertical structure of wintertime climate regimes of the Northern Hemisphere extratropical atmosphere, J. Climate, 16, 2005–2021, 2003.

Pardo, L.: Statistical Inference Based on Divergence Measures, Chapman and Hall, 2006.

Peña, D. and van der Linde, A.: Dimensionless measures of variability and dependence for multivariate continuous distributions, Comm. Stat. Theory Meth., 36, 1845–1854 doi:10.1080/03610920601126449, 2007.

Schneidman, E., Still, S., Berry, M. J., and Bialek, W.: Network information and connected correlations, Phys. Rev. Lett., 91, 238701-1–238701-4, 2003.