ISSN 1099-4300
www.mdpi.com/journal/entropy
Article

# Minimum Mutual Information and Non-Gaussianity Through the Maximum Entropy Method: Theory and Properties 

Carlos A. L. Pires * and Rui A. P. Perdigão

Instituto Dom Luiz, Faculdade de Ciências, University of Lisbon, DEGGE, Ed. C8, Campo-Grande, 1749-016 Lisbon, Portugal; E-Mail: raperdigao@fc.ul.pt

* Author to whom correspondence should be addressed; E-Mail: capires@fc.ul.pt; Tel.: +351-217500886; Fax: +351-217500807.

Received: 20 May 2012; in revised form: 8 June 2012 / Accepted: 15 June 2012 /
Published: 19 June 2012


#### Abstract

The application of the Maximum Entropy (ME) principle leads to a minimum of the Mutual Information (MI), $I(X, Y)$, between random variables $X$, $Y$, which is compatible with prescribed joint expectations and given ME marginal distributions. A sequence of sets of joint constraints leads to a hierarchy of lower MI bounds increasingly approaching the true MI. In particular, using standard bivariate Gaussian marginal distributions, it allows for the MI decomposition into two positive terms: the Gaussian MI $\left(I_{g}\right)$, depending upon the Gaussian correlation or the correlation between 'Gaussianized variables', and a non-Gaussian MI $\left(I_{n g}\right)$, coinciding with joint negentropy and depending upon nonlinear correlations. Joint moments of a prescribed total order $p$ are bounded within a compact set defined by Schwarz-like inequalities, where $I_{n g}$ grows from zero at the 'Gaussian manifold' where moments are those of Gaussian distributions, towards infinity at the set's boundary where a deterministic relationship holds. Sources of joint non-Gaussianity have been systematized by estimating $I_{n g}$ between the input and output from a nonlinear synthetic channel contaminated by multiplicative and non-Gaussian additive noises for a full range of signal-to-noise ratio (snr) variances. We have studied the effect of varying snr on $I_{g}$ and $I_{n g}$ under several signal/noise scenarios.


Keywords: mutual information; non-Gaussianity; maximum entropy distributions; non-Gaussian noise

PACS Codes: 92.60.Wc

## 1. Introduction

One of the most commonly used information theoretic measures is the mutual information (MI) [1], measuring the total amount of probabilistic dependence among random variables (RVs)—see [2] for a unifying perspective and axiomatic review. MI is a positive quantity vanishing iff RVs are independent.

MI is an effective tool for multiple purposes, namely and among others: (a) Blind signal separation [3] and Independent Component Analysis (ICA) [4], both of which look for transformed and/or lagged stochastic time series [5] which minimize MI; (b) Predictability studies, Predictable Component Analysis [6] and Forecast Utility [7], all of which are focused on the analysis and decomposition of the MI between probabilistic forecasts and observed states.

Analytical expressions of MI are known for a few number of parametric joint distributions [8,9]. Alternatively, it can be numerically estimated by different methods, such as Maximum Likelihood estimators, Edgeworth expansion, Bayesian methods, equiprobable and equidistant histograms, kernel-based probability distribution functions (PDFs), K-nearest neighbors technique-see [10,11] and references therein for a survey of estimation methods and scoring comparison studies.

In the bivariate case, treated here, the MI $I(X, Y)$, between RVs $X, Y$ is the Kullback-Leibler (KL) divergence: $I(X, Y) \equiv D\left(p_{X Y} \| p_{X} p_{Y}\right)=E_{p_{X Y}}\left(\log \left(p_{X Y} /\left(p_{X} p_{Y}\right)\right)\right) \geq 0$, between the joint probability function $p_{X Y}$ and the product of marginal probability distributions $p_{X} p_{Y}$ where $E_{p_{X Y}}$ is the expectation operator over the measure $p_{X Y}$. The MI is invariant for smooth invertible transformations of $X, Y$.

The goal is the determination of theoretical lower MI bounds under certain conditions or, in other words, the minimum mutual information (MinMI) [12] between two RVs $X, Y$, consistent, both with imposed marginal distributions and cross-expectations assessing their linear and nonlinear covariability. Those lower bounds can be obtained due to the application of the Maximum Entropy (ME) method to distributions [13] and to the inequality: $D\left(p_{M E} \| q\right) \leq D(p \| q)$. Here, $p_{M E}$ is a Maximum Entropy probability distribution (MEPD) with respect to $q$, say $p_{M E}=\arg \min _{p \in \Omega}(D(p \| q))$, where $\Omega$ is a PDF class verifying a given set of constraints [1]. Therefore, by using the joint MEPD $p=p_{X Y-M E}$ and $q=p_{X} p_{Y}$, the lower MI bound is obtained: $D\left(p_{X Y-M E} \| p_{X} p_{Y}\right) \leq D\left(p_{X Y} \in \Omega \| p_{X} p_{Y}\right) \leq I(X, Y)$. Finding of the bivariate $p_{X Y-M E}$ is straightforward if the marginal distributions are themselves well defined ME distributions. We solve that by transforming the single variables $X, Y$ into others with imposed ME probability mass distributions through the so called ME-anamorphoses [14].

The joint ME probability distribution $p_{X Y-M E}$ is derived from the minimum of a functional in terms of a certain number of Lagrange parameters. The properties of multivariate ME distributions have been studied for various ME constraints, namely: (a) imposed marginals and covariance matrix [15]; (b) generic joint moments [16]. Abramov [17-19] has developed efficient and stable numerical iterative algorithms for computing ME distributions forced by sets of polynomial expectations. Here we use a bivariate version of the algorithm of [20], already tested in [21].

By taking a sequence of encapsulated sets of joint ME constraints we obtain an increasing hierarchy of lower MI bounds converging towards the total MI.

We particularize this methodology to the case where $X, Y$ are standard Gaussians, issued from single homeomorphisms of an original pair of variables $\hat{X}, \hat{Y}$ by Gaussian anamorphosis [14]. Then
we get the MI $I(X, Y)=I(\hat{X}, \hat{Y})$, which is decomposed into two generic positive quantities, a Gaussian MI $I_{g}$ and a non-Gaussian MI $I_{n g}$ [21], vanishing under bivariate Gaussianty. The Gaussian MI is given by $I_{g}\left(c_{g}\right) \equiv-1 / 2 \log \left(1-c_{g}^{2}\right)$, where $c_{g}(\hat{X}, \hat{Y}) \equiv \operatorname{cor}(X, Y)$ is the Pearson linear correlation between the Gaussianized variables. As for the non-Gaussian MI term $I_{n g}$ it relies upon imposed nonlinear correlations between the single Gaussian variables. The MI reduces to $I_{g}\left(c_{g}\right)$ when only moments of order one and two are imposed as ME constraints. We will note that, for certain extreme non-Gaussian marginals, $I_{g}(c=\operatorname{cor}(\hat{X}, \hat{Y}))>I(\hat{X}, \hat{Y})$, thus showing that $I_{g}(c)$ [22] is not a proper MI lower bound in general.

The correlation $c_{g}$, hereafter called Gaussian correlation [21] is a nonlinear concordance measure like the Spearman rank correlation [23] and the Kendall $\tau$, which, by definition are invariant for monotonically growing smooth marginal transformations. These measures have the good property of being expressed as functionals of the copula density functions [24], uniquely dependent on the cross dependency between variables.

The non-Gaussian MI $I_{n g}$ holds some interesting characteristics. It coincides with the joint negentropy (deficit of entropy with respect to that of the Gaussian PDF with the same mean, variance and covariance) in the space of 'Gaussianized' variables, which is invariant for any orthogonal or oblique rotation of them. In particular for uncorrelated rotated variables, it coincides with the 'compactness', which measures the concentration of the joint distribution around a lower-dimensional manifold [25] which is given by $D\left(p_{X Y} \| p_{S G}\right)$, i.e., the KL divergence with respect to the spherical Gaussian $p_{S G}$ (Gaussian with an isotropic covariance matrix with the same trace as that of $p_{X Y}$, say the total variance).

We also show that $I_{n g}$ comprises a series of positive terms associated to a $p$-sequence of imposed monomial expectations of total even order $p(2,4,6,8 \ldots)$. The higher the number of independent constraints, the higher the order of terms that are retained in that series and the more information is extracted from the joint PDF.

We have shown that the possible values of the cross-expectations lie within a bounded set obtained by Schwarz-like inequalities. We illustrate the range of $I_{n g}$ values within those sets as function of third and fourth-order cross moments. Near the set's boundary, $I_{n g}$ tends to infinity where a deterministic relationship holds and the ME problem functional is ill-conditioned.

In order to better understand the possible sources of joint non-Gaussianity and non-Gaussian MI, we have used the preceding method for computing $I_{g}$ and $I_{n g}$ between the input and the output of a nonlinear channel contaminated by multiplicative and non-Gaussian noise for a full range of the signal-to-noise (snr) variance ratio. We put in evidence that sources of non-Gaussian MI arise from the nonlinearity of the transfer function, multiplicative noise and additive non-Gaussian noise [26].

Many of the results of the paper are straightforwardly generalized to the multivariate case with three or more random variables.

The paper is then organized as follows: Section 2 formalizes the Minimum Mutual Information (MinMI) principle from maximum entropy distributions, while Section 3 particularizes that principle to the MI decomposition into Gaussian and non-Gaussian MI parts. Section 4 addresses the non-Gaussianity in a nonlinear non-Gaussian channel. The paper ends with conclusions and appendices with theoretical proofs and the numerical algorithm for solving the ME problem. This
paper is followed by a companion one [27] on the estimation of non-Gaussian MI from finite samples with practical applications.

## 2. MI Estimation from Maximum Entropy PDFs

In this section we present the basis of the MI estimation between bivariate RVs, through the use of joint PDFs inferred by the maximum entropy (ME) method (ME-PDFs for short) in the space of transformed (anamorphed) marginals into specified ME distributions. We start with preliminary general concepts and definitions.

### 2.1. General Properties of Bivariate Mutual Information

Let $(X, Y)$ be continuous RV s with support given by the Cartesian product $S=S_{X} \otimes S_{Y} \subseteq \mathbb{R}^{2}$, and let the joint PDF $\rho_{X, Y}$ be taken absolutely continuous with respect to the product of the marginal PDFs $\rho_{X}$ and $\rho_{Y}$. The mutual information (in nat) between $X$ and $Y$ is a non-negative real functional of $\rho_{X, Y}$ expressed in different equivalent forms as:

$$
\begin{equation*}
I(X, Y) \equiv \int_{S} \rho_{X, Y}(x, y) \log \left(\frac{\rho_{X, Y}(x, y)}{\rho_{X}(x) \rho_{Y}(y)}\right) d x d y=D\left(\rho_{X Y} \| \rho_{X} \rho_{Y}\right)=H_{\rho_{X}}+H_{\rho_{Y}}-H_{\rho_{X Y}} \geq 0, \tag{1}
\end{equation*}
$$

in terms both of a KL divergence and Shannon entropies. The MI equals zero iff the RVs $X$ and $Y$ are statistically independent, i.e., iff $\rho_{X, Y}(x, y)=\rho_{X}(x) \rho_{Y}(y), \forall x \in S_{X}, y \in S_{Y}$, except possibly in a zero-measure set in $S$. Under quite general regularity conditions of PDFs, statistical independence of $X$ and $Y$ is equivalent to the vanishing of the Pearson correlation between any pair of linear and/or nonlinear mapping functions $\tilde{X}(X)$ and $\tilde{Y}(Y)$. Then, linearly uncorrelated non-independent variables have at least a non-zero nonlinear correlation. The MI between transformed variables $\tilde{X}(X)$ and $\tilde{Y}(Y)$, differentiable almost everywhere, satisfies the 'data processing inequality' $I(\tilde{X}(X), \tilde{Y}(Y)) \leq I(X, Y) \quad$ [1] with equality occurring if both $\tilde{X}$ and $\tilde{Y}$ are smooth homeomorphisms of $X$ and $Y$ respectively.

### 2.2. Congruency between Information Moment Sets

Definition 1: Following the notation of [28], we define the moment class $\Omega_{\mathrm{T}, \boldsymbol{\theta}}$ of bivariate PDFs $\rho_{X Y}$ of $(X, Y)$, as:

$$
\begin{equation*}
\Omega_{\mathbf{T}, \boldsymbol{\theta}}=\left\{\rho_{X Y}: E_{\rho}[\mathbf{T}]=\boldsymbol{\theta}\right\} \tag{2}
\end{equation*}
$$

where $E_{\rho}$ is the $\rho$-expectation operator, $\mathbf{T}=\left(T_{1}, \ldots, T_{J}\right)^{T}$ is a $J$-dimensional vector composed of $J$ absolutely ( $X, Y$ ) integrable functions with respect to $\rho_{X Y}$, and $\boldsymbol{\theta}=\left(\theta_{1}, \ldots, \theta_{J}\right)^{T}$ is the vector of function expectations of $\mathbf{T}$. Here and henceforth, $(\mathbf{T}, \boldsymbol{\theta})$ is denoted as an information moment set.

The PDF $\rho \in \Omega_{\mathrm{T}, \theta}$ that maximizes the Shannon entropy or the ME-PDF verifying the constraints associated to $(\mathbf{T}, \boldsymbol{\theta})$, hereby represented by $\rho_{\mathbf{T}, \boldsymbol{\theta}}^{*}$, must exist since $H_{\rho}$ is a concave function in the non-empty convex set $\Omega_{\mathrm{T}, \theta}$. The form and estimation of the ME-PDF is described in Appendix 1. The ME-PDF satisfies the following Lemma [1]:

Lemma 1: Given two information encapsulated information moment sets: $(\mathbf{T}, \boldsymbol{\theta}) \subseteq\left(\mathbf{T}_{1}, \boldsymbol{\theta}_{1}\right)$ i.e., with $\mathbf{T}_{1}$ including more constraining functions than $\mathbf{T}$, the respective ME-PDFs $\rho_{\mathrm{T}, 9}^{*}, \rho_{\mathrm{T}_{1}, \boldsymbol{\theta}_{1}}^{*}$, if they exist, satisfy the following conditions:

$$
\begin{align*}
& E_{\rho}\left[-\log \rho_{\mathbf{T}, \theta}^{*}\right]=E_{\rho_{\mathrm{T}, 91}^{*}}\left[-\log \rho_{\mathbf{T}, \theta}^{*}\right]=E_{\rho_{\mathrm{T}, 9}^{*}}\left[-\log \rho_{\mathbf{T}, \theta}^{*}\right]=H_{\rho_{\mathrm{T}, \theta}^{*}}  \tag{3a}\\
& E_{\rho}\left[-\log \rho_{\mathbf{T} 1, \theta 1}^{*}\right]=E_{\rho_{\mathrm{T}, 0,91}^{*}}\left[-\log \rho_{\mathbf{T} 1, \theta 1}^{*}\right]=H_{\rho_{\mathrm{T}, 01}^{*}} \tag{3b}
\end{align*}
$$

This means that any ME-PDF log-mean is unchanged when it is computed with respect to a more constraining ME-PDF. As a corollary we have $D\left(\rho_{\mathrm{T}_{1}, \theta_{1}}^{*} \| \rho_{\mathrm{T}, \mathrm{e}}^{*}\right)=H_{\rho_{\mathrm{T}, 9}^{*}}-H_{\rho_{\mathrm{T}, \mathrm{T}_{1}}^{*}} \geq 0$, i.e., the ME decreases with the increasing number of constraints. This can be understood bearing in mind that the entropy maximization is performed in a more reduced PDF class, since $\Omega_{\mathrm{T}_{1}, \theta_{1}} \subseteq \Omega_{\mathrm{T}, \boldsymbol{\theta}}$.

Definition 2: If two information moment sets $\left(\mathbf{T}_{1}, \boldsymbol{\theta}_{1}\right),\left(\mathbf{T}_{2}, \boldsymbol{\theta}_{2}\right)$ are related by linear affine relationships, then the sets are referred to as "congruent", a property hereby denoted as $\left(\mathbf{T}_{1}, \boldsymbol{\theta}_{1}\right) \underset{P D F}{\cong}\left(\mathbf{T}_{2}, \boldsymbol{\theta}_{2}\right)$, and consequently both PDF sets are equal, i.e., $\Omega_{T_{1}, \boldsymbol{\theta}_{1}}=\Omega_{\mathrm{T}_{2}, \boldsymbol{\theta}_{2}}$ [15]. A stronger condition than 'congruency' is the ME-congruency, denoted as $\left(\mathbf{T}_{1}, \boldsymbol{\theta}_{1}\right) \cong\left(\mathbf{T}_{2}, \boldsymbol{\theta}_{2}\right)$, holding when both the associated ME-PDFs are equal. For example, both the univariate constraint sets ( $\mathbf{T}_{1}=X^{2}, \boldsymbol{\theta}_{1}=1$ ) and $\left(\mathbf{T}_{2}=\left(X^{2}, X^{4}\right)^{T}, \boldsymbol{\theta}_{2}=(1,3)^{T}\right)$ for $X \in \mathbb{R}$ lead to the same ME-PDF, the standard Gaussian $\mathrm{N}(0,1)$. Consequently both information moment sets are ME-congruent but not congruent since $\Omega_{\mathbf{T}_{2}, \boldsymbol{\theta}_{2}} \subset \Omega_{\mathrm{T}_{1}, \boldsymbol{\theta}_{1}}$. This is because the Lagrange multiplier of the ME functional (see Appendix 1) corresponding to the fourth moment is set to zero without any constraining effect. The congruency implies ME-congruency but not the converse.

### 2.3. MI Estimation from Maximum Entropy Anamorphoses

We are looking for a method of obtaining lower bound MI estimates from ME-PDFs. For that purpose we will decompose the information moment set as: $(\mathbf{T}, \boldsymbol{\theta})=\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right) \cup\left(\mathbf{T}_{c r}, \boldsymbol{\theta}_{c r}\right)$, say into a marginal or independent part $\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right)=\left(\mathbf{T}_{X}, \boldsymbol{\theta}_{X}\right) \cup\left(\mathbf{T}_{Y}, \boldsymbol{\theta}_{Y}\right)$ of single $X$ and $Y$ independent moments on $S_{X}, S_{Y}$, and a cross part $\left(\mathbf{T}_{c r}, \boldsymbol{\theta}_{c r}\right)$ on $S_{X} \otimes S_{Y}$, made by moments of joint $(X, Y)$ functions not expressible as sums of single moments. For example, by taking $\left(\mathbf{T}_{X}, \boldsymbol{\theta}_{X}\right)=\left(\left(X, X^{2}\right),(0,1)\right)$, $\left(\mathbf{T}_{Y}, \boldsymbol{\theta}_{Y}\right)=\left(\left(Y, Y^{2}\right),(0,1)\right)$ and $\left(\mathbf{T}_{c r}, \boldsymbol{\theta}_{c r}\right)=(X Y, c)$ for $(X, Y) \in \mathbb{R}^{2}$ leads to a ME-PDF which is the bivariate Gaussian with correlation $c$ and standard Gaussians as marginal distributions.

The joint ME-PDF associated to the independent part $\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right)$ is the product of two independent ME-PDFs related to $\left(\mathbf{T}_{X}, \boldsymbol{\theta}_{X}\right)$ and $\left(\mathbf{T}_{Y}, \boldsymbol{\theta}_{Y}\right)$, with the joint entropy being the sum of marginal maximum entropies, as in the case of independent random variables [15]. The KL divergence between the ME-PDFs associated to ( $\mathbf{T}, \boldsymbol{\theta})$ and those associated to $\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right)$ is a proper MI lower bound or, expressed in other terms, a constrained mutual information. We denote it as:

Its difference with respect to $I(X, Y)$ is given by:

$$
\begin{equation*}
I(X, Y)-I\left(X, Y:(\mathbf{T}, \boldsymbol{\theta}),\left(\mathbf{T}_{i n d}, \boldsymbol{\theta}_{i n d}\right)\right)=D\left(\rho_{X Y} \| \rho_{\mathbf{T}, \boldsymbol{\theta}}^{*}\right)-\left[D\left(\rho_{X} \| \rho_{\mathrm{T} X, 0 \mathrm{Q}}^{*}\right)+D\left(\rho_{Y} \| \rho_{\mathrm{T},, \hat{}}^{*}\right)\right] \tag{5}
\end{equation*}
$$

which can be negative. However, the positiveness is ensured when marginal distributions are set to the ME-PDFs constrained by $\rho_{X}(x)=\rho_{\mathrm{r} x, Q X}^{*}(x), \forall x$ and $\rho_{Y}(y)=\rho_{\mathrm{T} \gamma, \theta Y}^{*}(y), \forall y$, which results in KL divergences vanishing with respect to the marginal PDFs in Equation (5). This procedure is quite general because $(X, Y)$ can appropriately be obtained through an injective smooth maps $(X=X(\hat{X}), Y=Y(\hat{Y}))$ from an original pair of RVs $(\hat{X}, \hat{Y})$ preserving the MI, i.e., $I(\hat{X}, \hat{Y})=I(X, Y)$. Those maps are monotonically growing homeomorphisms, the hereby called ME-anamorphoses, which are obtained by equaling mass probability functions of the original variable $\rho_{\hat{\chi}}$ to those of the transformed variable $\rho_{X}$ (equally for $\rho_{\hat{Y}}$ and $\rho_{Y}$ ) as:

$$
\begin{equation*}
\int_{-\infty}^{X(\hat{X})} \rho_{\tau X, 0 x}^{*}(u) d u=\int_{-\infty}^{\hat{X}} \rho_{\hat{X}}(u) d u \quad ; \int_{-\infty}^{Y(\hat{Y})} \rho_{\mathrm{T} \gamma, 9 r}^{*}(u) d u=\int_{-\infty}^{\hat{Y}} \rho_{\hat{Y}}(u) d u \tag{6}
\end{equation*}
$$

The moments in $\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right)$ are invariant under the transformation, $(\hat{X}, \hat{Y}) \rightarrow(X, Y)$, i.e., they are the same for both original and transformed variables: $E_{\rho_{\hat{X}}}\left[\mathbf{T}_{X}\right]=E_{\rho_{X}}\left[\mathbf{T}_{X}\right]=\boldsymbol{\theta}_{X}$ (idem for $Y$ ). Therefore, in the space of ME-anamorphed variables the ME-based MI (4) gives the minimum MI (MinMI), compatible with the cross moments ( $\mathbf{T}_{c r}, \boldsymbol{\theta}_{c r}$ ).

Thanks to Lemma 1, a hierarchy of ME-based MI bounds is obtainable by considering successive supersets of the ME constraints on the ME-anamorphed variables, which is justified by the theorem below.

Theorem 1: Let $(X, Y)$ be a pair of single random variables (RVs), distributed as the ME-PDF associated to the independent constraints $\left(\mathbf{T}_{\text {ind }}=\left(\mathbf{T}_{X}, \mathbf{T}_{Y}\right), \boldsymbol{\theta}_{\text {ind }}=\left(\boldsymbol{\theta}_{X}, \boldsymbol{\theta}_{Y}\right)\right)$. Both variables can be obtained from previous ME-anamophosis. Let $\left(\mathbf{T}_{1}, \boldsymbol{\theta}_{1}\right)=\left(\mathbf{T}_{c r 1} \cup \mathbf{T}_{\text {ind } 1}, \boldsymbol{\theta}_{\text {cr1 }} \cup \boldsymbol{\theta}_{\text {ind1 }}\right)$ be a subset of $\left(\mathbf{T}_{2}, \boldsymbol{\theta}_{2}\right)=\left(\mathbf{T}_{c r 2} \cup \mathbf{T}_{\text {ind } 2}, \boldsymbol{\theta}_{c r 2} \cup \boldsymbol{\theta}_{\text {ind } 2}\right)$, i.e., $\mathbf{T}_{c r 1} \subseteq \mathbf{T}_{c r 2}$ and $\mathbf{T}_{i n d} \subseteq \mathbf{T}_{\text {ind } 1} \subseteq \mathbf{T}_{\text {ind } 2}$, such that all independent moment sets are ME-congruent (see Definition 2), i.e., $\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right) \xlongequal[\overline{M E}]{\cong}\left(\mathbf{T}_{\text {ind } 1}, \boldsymbol{\theta}_{\text {ind1 }}\right) \xlongequal[\overline{M E}]{\cong}\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind2 }}\right)$, i.e., such that the independent extra moments in $\left(\mathbf{T}_{2}, \boldsymbol{\theta}_{2}\right)$ are not further constraining the ME-PDF. Each marginal moment set is decomposed as $\left(\mathbf{T}_{\text {ind } j}=\left(\mathbf{T}_{X j}, \mathbf{T}_{Y j}\right), \boldsymbol{\theta}_{\text {ind } j}=\left(\boldsymbol{\theta}_{X j}, \boldsymbol{\theta}_{Y j}\right)\right),(j=1,2)$. For simplicity of notation, let us denote $I(X, Y)_{j} \equiv I\left(X ; Y:\left(\mathbf{T}_{j}, \boldsymbol{\theta}_{j}\right),\left(\mathbf{T}_{\text {ind } j}, \boldsymbol{\theta}_{\text {ind } j}\right)\right)(j=1,2)$. Then, the following inequalities between constrained mutual informations hold:

$$
\begin{align*}
& I(X, Y)_{j}=I(X, Y)-D\left(\rho_{X Y} \| \rho_{\mathbf{T}_{j}, \boldsymbol{\theta}_{j}}^{*}\right) \leq I(X, Y)(j=1,2)  \tag{7a}\\
& I(X, Y)_{2}-I(X, Y)_{1}=D\left(\rho_{\mathbf{T}_{2}, \mathbf{\theta}_{2}}^{*} \| \rho_{\mathbf{T}_{1}, \boldsymbol{\theta}_{1}}^{*}\right)=H_{\rho_{\mathrm{T}_{1}, 0_{1}}^{*}}^{*}-H_{\rho_{\mathrm{T}, 2, \theta_{2}}^{*}} \geq 0 \tag{7b}
\end{align*}
$$

The proof is given in Appendix 2. The Theorem 1 justifies the possibility of building a monotonically growing sequence of lower bounds of $I(X, Y)$ from encapsulated sequences of crossconstraints $\mathbf{T}_{c r 1} \subseteq \ldots \mathbf{T}_{c r j} \subseteq \mathbf{T}_{c r j+1} \subseteq \ldots$ and independent constraints $T_{\text {ind }} \subseteq T_{\text {ind } 1} \subseteq \ldots \subseteq T_{\text {ind } j} \subseteq T_{\text {ind } j+1} \subseteq \ldots$. In the sequence, the entropy associated to independent constraint sets is always constant due to the ME-congruency, while the entropy of the joint ME-PDF decreases, thus allowing the MinMIs to grow.

Therefore, $I(X, Y)_{j}$ is the part of MI due to cross moments in $\mathbf{T}_{c r j}$ and the positive difference $I(X, Y)_{j+1}-I(X, Y)_{j}$ is the increment of MI due to the additional cross moments in $\mathbf{T}_{\text {ind } j+1} / \mathbf{T}_{\text {ind } j}$, while marginals are kept as preset ME-PDFs (e.g., Gaussian, Gamma, Weibull).

## 3. MI Decomposition under Gaussian Marginals

### 3.1. Gaussian Anamorphosis and Gaussian Correlation

In this section we explain how to implement the sequence of MI estimators detailed in Section 2.3 for the particular case where $X$ and $Y$ are standard Gaussian RVs. Our aim is to estimate the MI between two original variables $(\hat{X}, \hat{Y})$ of null mean and unit variance with real support. Those variables are then transformed through a homeomorphism, the Gaussian anamorphosis [14], into standard Gaussian RVs, respectively $X \sim N(0,1)$ and $Y \sim N(0,1)$, given by:

$$
\begin{equation*}
X(\hat{x})=\Phi^{-1}\left(\int_{-\infty}^{\hat{x}} \rho_{\hat{X}}(u) d u\right)=G_{\hat{X}}(\hat{x}) ; Y(\hat{y})=\Phi^{-1}\left(\int_{-\infty}^{\hat{y}} \rho_{\hat{Y}}(u) d u\right)=G_{\hat{Y}}(\hat{y}) \tag{8}
\end{equation*}
$$

where $\Phi$ is the mass distribution function for the standard Gaussian. If ( $\hat{X}, \hat{Y}$ ) are marginally non-Gaussian, then the Gaussian anamorphoses are nonlinear transformations. In practice, Gaussian anamorphoses can be approximated empirically from finite data sets by equaling cumulated histograms. However, for certain cases, it is analytically possible to construct bivariate distributions with specific marginal distributions and the knowledge of the joint cumulative distribution function [29].

In the case of Gaussian anamorphosis, the information moment set $\left(\mathbf{T}_{\text {ind }}, \boldsymbol{\theta}_{\text {ind }}\right)$ of Theorem 1 includes the first and second independent moments of each variable: $E[X]=E[Y]=0 ; E\left[X^{2}\right]=E\left[Y^{2}\right]=1$. Then, following the proposed procedure of Section 2.3 , we will consider a sequence of cross-constraint sets for determining the hierarchy of lower MI bounds.

The most obvious cross moment to be considered is the $X Y$ expectation, equal to the Gaussian correlation $c_{g} \equiv \operatorname{cor}\left(G_{\hat{X}}, G_{\hat{Y}}\right)=\operatorname{cor}(X, Y)$ between the 'Gaussianized' variables $(X, Y)$. The difference between $c_{g}$ and the linear correlation $c=\operatorname{cor}(\hat{X}, \hat{Y})$ is easily expressed as:

$$
\begin{equation*}
c_{g}-c=E\left[\left(G_{\hat{X}}-\hat{X}\right)\left(G_{\hat{Y}}-\hat{Y}\right)\right] \tag{9}
\end{equation*}
$$

The signal of the factor $G_{\hat{X}}-\hat{X}$ in Equation (9) roughly depends on the skewness $\operatorname{sk}(\hat{X})=E\left[\hat{X}^{3}\right]$ and excess of kurtosis $\operatorname{kur}(\hat{X})=E\left[\hat{X}^{4}\right]-3$ through the rule of thumb, stating that $\operatorname{sgn}\left(G_{\hat{X}}-\hat{X}\right)$ is approximated by $-\operatorname{sgn}(\operatorname{sk}(\hat{X}))$ and $-\operatorname{sgn}(k u r(\hat{X})) \operatorname{sgn}(\hat{X})$, respectively for a skewed $\hat{X}$ PDF and a symmetric $\hat{X}$ PDF (idem for $\hat{Y}$ ). Therefore, $c_{g}$ can result in an enhancement of correlation $c$ or in the opposite effect, as shown in [21] for the RV pair of meteorological variables ( $\hat{X}=$ North Atlantic Oscillation Index, $\hat{Y}=$ monthly precipitation). The Gaussian correlation is a concordance measure like the rank correlation and Kendall $\tau$, being thus invariant for a monotonically growing smooth homeomorphism of both $\hat{X}$ and $\hat{Y}$. Those measures are expressed as functionals of the bivariate copula-function $c\left[u=\int_{-\infty}^{X_{1}} \rho_{X_{1}}(x) d x, v=\int_{-\infty}^{X_{2}} \rho_{X_{2}}(y) d y\right]=\rho_{X_{1} X_{2}}\left(X_{1}, X_{2}\right) /\left(\rho_{X_{1}}\left(X_{1}\right) \rho_{X_{2}}\left(X_{2}\right)\right)$, which is uniquely dependent on the cumulated marginal probabilities and equal to the density ratio,
independently from the specific forms of marginal PDFs [24]. In particular, the Gaussian correlation is given by:

$$
\begin{equation*}
c_{g}=\int_{0}^{1} \int_{0}^{1} c[u, v] \Phi^{-1}(u) \Phi^{-1}(v) d u d v \tag{10}
\end{equation*}
$$

### 3.2. Gaussian and Non-Gaussian MI

The purpose of this sub-section is to express which part of MI comes from joint non-Gaussianity. If the 'Gaussianized' variables $(X, Y)$ are jointly non-Gaussian, then the original standardized variables $(\hat{X}, \hat{Y})$, obtained from $(X, Y)$ by invertible smooth monotonic transformations, are jointly non-Gaussian as well. However, the converse is not true. In fact, any nonlinear transformation ( $\hat{X}, \hat{Y}$ ) of jointly Gaussian RVs $(X, Y)$ leads to non-Gaussian $(\hat{X}, \hat{Y})$ with the joint non-Gaussianity arising in a trivial way. Therefore, the 'genuine' joint non-Gaussianity can only be diagnosed in the space of Gaussianized variables $(X, Y)$. When the marginal PDFs of $(\hat{X}, \hat{Y})$ are non-Gaussian then $c_{g} \neq c$ in general. In particular, if the correlation $c$ is null as it occurs when $(\hat{X}, \hat{Y})$ are principal components (PCs), then $c_{g}$ can be non-null, leading to statistically dependent PCs since they are non-linearly correlated.

The MI between Gaussianized variables $(X, Y)$ or $(\hat{X}, \hat{Y})$ is expressed as $I(\hat{X}, \hat{Y})=I(X, Y)=2 H_{g}-H_{\rho_{X Y}}$, where $H_{g} \equiv 1 / 2 \log (2 \pi e)$ is the Shannon entropy of the standard Gaussians $X, Y$ and $H_{\rho_{X Y}}$ is the $(X, Y)$ joint entropy, associated to the joint PDF $\rho_{X Y}$. Given the above equality, the MI is decomposed into two positive terms: $I(\hat{X}, \hat{Y})=I(\hat{X}, \hat{Y})_{g}+I(\hat{X}, \hat{Y})_{n g}$. The first term is the Gaussian MI [21] given by $I(\hat{X}, \hat{Y})_{g}=I(X, Y)_{g}=-1 / 2 \log \left(1-c_{g}^{2}\right) \equiv I_{g}\left(c_{g}\right) \geq 0$, as a function of the Gaussian correlation $c_{g}$ (see its graphic in Figure 1), and the second one is the non-Gaussian MI $I(\hat{X}, \hat{Y})_{n g}=I(X, Y)_{n g}$, which is due to joint non-Gaussianity and nonlinear statistical relationships among variables.

The MI $I(\hat{X}, \hat{Y})$ is related to the negentropy $J(\hat{X}, \hat{Y})$, i.e., to the KL divergence between the PDF and the Gaussian PDF with the same moments of order one and two. That is shown by:

Theorem 2: Given $\left(\hat{X}_{r}, \hat{Y}_{r}\right)^{T}=A(\hat{X}, \hat{Y})^{T}$, a pair of rotated standardized variables ( $A$ being an invertible $2 \times 2$ matrix), one has the following result with proof in Appendix 2:

$$
\begin{align*}
& J(\hat{X}, \hat{Y})=J(\hat{X})+J(\hat{Y})+I(\hat{X}, \hat{Y})-I_{g}(\operatorname{cor}(\hat{X}, \hat{Y}))= \\
& =J\left(\hat{X}_{r}, \hat{Y}_{r}\right)=J\left(\hat{X}_{r}\right)+J\left(\hat{Y}_{r}\right)+I\left(\hat{X}_{r}, \hat{Y}_{r}\right)-I_{g}\left(\operatorname{cor}\left(\hat{X}_{r}, \hat{Y}_{r}\right)\right) \tag{11}
\end{align*}
$$

A simple consequence is that in the space of uncorrelated variables $\left(\right.$ i.e., $\left.I_{g}(\operatorname{cor}(\hat{X}, \hat{Y}))=0\right)$, the joint negentropy is the sum of marginal negentropies with the MI, thus showing that there are intrinsic and joint sources of non-Gaussianity. One interesting corollary is derived from that.

Corollary 1: For standard Gaussian variables $(X, Y)$ and standardized rotated ones $\left(X_{r}, Y_{r}\right)^{T}=A(X, Y)^{T}$, we have

$$
\begin{equation*}
I_{n g}(X, Y)=J(X, Y)=J\left(X_{r}, Y_{r}\right)=J\left(X_{r}\right)+J\left(Y_{r}\right)+I\left(X_{r}, Y_{r}\right)-I_{g}\left(\operatorname{cor}\left(X_{r}, Y_{r}\right)\right) \tag{12}
\end{equation*}
$$

For the proof it suffices to consider Gaussian variables $(\hat{X}, \hat{Y})^{T}=(X, Y)^{T}$ in (11). Their self-negentropy vanishes by definition and the correlation term is the Gaussian MI.

Negentropy has the property of being invariant for any orthogonal or oblique rotations of the Gaussianized variables $(X, Y)$. However, this invariance does not extend to $I(X, Y)_{n g}$. From (12), in particular when $\left(X_{r}, Y_{r}\right)$ are uncorrelated (e.g., standardized principal components of $(X, Y)$ or uncorrelated standardized linear regression residuals), the negentropy equals the KL divergence between the joint PDF and that of an isotropic Gaussian with the same total variance. That KL divergence is the compactness (level of concentration around a lower-dimensional manifold), as defined in [25]. This measure is invariant under orthogonal rotations. The last term of (12) vanishes due the fact that the null correlation allows one to decompose of $I(X, Y)_{n g}$ into positive contributions: the self-negentropies of uncorrelated variables $\left(X_{r}, Y_{r}\right)$ and their MI $I\left(X_{r}, Y_{r}\right)=I\left(X_{r}, Y_{r}\right)_{g}+I\left(X_{r}, Y_{r}\right)_{n g}$. These variables can be 'Gaussianized' and rotated, leading to further decomposition of $I\left(X_{r}, Y_{r}\right)_{n g}$ until the possible "emptying"/depletion of the initial joint non-Gaussianity into Gaussian MIs and univariate negentropies. The PDF of the new rotated variables will be closer to an isotropic spherical Gaussian PDF. Since it is algorithmically easier to compute univariate rather than multivariate entropies, the above method can be used for an efficient estimation of MIs.

The search for rotated variables maximizing the sum of individual negentropies $J\left(X_{r}\right)+J\left(Y_{r}\right)$ in (12) with minimization of $I\left(X_{r}, Y_{r}\right)$ or their statistical dependency is the goal of Independent Component Analysis (ICA) [4].

A natural generalization of the MI decomposition is possible when $(X, Y)$ is obtained from a generic ME-anamorphosis by decomposing the MI into a term associated to correlation, under the constraint that marginals are set to given ME-PDFs (the equivalent to $I_{g}$ ), and to a term not explained by correlation (the equivalent to $I_{n g}$ ). There is however no guarantee that this decomposition is unique as in the case of non-Gaussians, since there is no natural bivariate extension of univariate prescribed PDFs with a given correlation [30].

By looking again at Equation (12), we notice that when original variables are correlated $\left(c=\operatorname{cor}\left(X_{r}, Y_{r}\right) \neq 0\right)$, the sum of marginal negentropies is not necessarily a lower bound of the joint negentropy $J(\hat{X}, \hat{Y})$, because in some cases $I(\hat{X}, \hat{Y})-I_{g}(c)$ can be negative. This means that $I_{g}(c)$ is not generally a proper lower bound of the MI. An example of that is given by the following discrete distribution with support on four points: $(\hat{X}, \hat{Y})=(1,1),(1,-1),(-1,1),(-1,-1)$, with mass probabilities $P_{\hat{X} \hat{Y}}$, respectively of $(1+c) / 4,(1+c) / 4,(1-c) / 4$ and $(1-c) / 4$. This 4-point distribution has $\hat{X}$ and $\hat{Y}$ zero means, unit variances and Pearson correlation $c$. The PDF is made of four Dirac-Deltas. In this case, the mutual information $I(\hat{X}, \hat{Y})$ is easily computed from the 4-point discrete mean of $\log \left(P_{\hat{X} \hat{Y}} /\left(P_{\hat{X}} P_{\hat{Y}}\right)\right)$ while the marginal mass probabilities are $P_{\hat{X}}(1)=P_{\hat{Y}}(1)=P_{\hat{X}}(-1)=P_{\hat{Y}}(-1)=1 / 2$. After some lines of algebra the MI becomes:

$$
\begin{equation*}
I(X, Y)=\log \left[2\left(\frac{1+|c|}{2}\right)^{\left(\frac{1+c \mid c}{2}\right)}\left(\frac{1-|c|}{2}\right)^{\left(\frac{1-|c|}{2}\right)}\right] \equiv I_{d 4}(c) \tag{13}
\end{equation*}
$$

The MI is bounded for any value of $c$ and $I_{d 4}(c) \rightarrow \log (2)<\infty$ as $(|c| \rightarrow 1)$. The graphic of $I_{d 4}(c)$ is included in Figure 1, showing that $I_{g}(c)>I_{d 4}(c)$ for about $|c|>0.45$. This behavior is also reproduced by finite discontinuous PDFs (e.g., replacing the Dirac-Deltas by cylinders of probability) as well as continuous PDFs. We test it by approximating the discrete PDF by the weighted superposition of 4 spherical bivariate Gaussian PDFs, all of which have a sharp isotropic standard deviation $\sigma=0.001$ and are centered at the 4 referred centroids. Figure 1 depicts successively growing MI lower bounds for each value of $c$, using the maximum entropy method (labels, $I_{g}, I_{g}+I_{n g, 4}$, $I_{g}+I_{n g, 6}, I_{g}+I_{n g, 8}$, see Section 3.3 for details), showing the convergence to the asymptotic value $\log (2)$.

Figure 1. Semi-logarithmic graphs of $I_{g}(c)$ (black thick line), $I_{4}(c)$, (grey thick line) and of the successive growing estimates of $I_{4}(c): I_{g}, I_{g}+I_{n g, 4}, I_{g}+I_{n g, 6}$ and $I_{g}+I_{n g, 8}$ (grey thin lines). See text for details.


### 3.3. The Sequence of Non-Gaussian MI Lower Bounds from Cross-Constraints

In order to build a monotonically increasing sequence of lower bounds for $I(X, Y)$ (cf. Section 2.3), we have considered a sequence of encapsulated sets of functions whose moments will constrain the ME-PDFs. Those functions consist of single (univariate) and combined (bivariate) monomials of the standard Gaussians $X$ and $Y$. The numerical implementation of joint ME-PDFs constrained by polynomials in dimensions $d=2,3$ and 4 was studied by Abramov [17-19], with particular emphasis on the efficiency and convergence of iterative algorithms. Here, we use the algorithm proposed in [21] and explained in the Appendix 1. Let us define the information moment set $\mathbf{T}_{p}$ as the set of bivariate monomials with total order up to $p$ :

$$
\begin{equation*}
\mathbf{T}_{p} \equiv\left(X^{i} Y^{j}: 1 \leq i+j \leq p,(i, j) \in N_{0}^{2}\right), p \in N \tag{14}
\end{equation*}
$$

This set is decomposed into marginal (independent) and cross monomials as $\mathbf{T}_{p} \equiv\left(\mathbf{T}_{p \text { ind }}, \mathbf{T}_{p c r}\right)$ with the corresponding vector of expectations $\boldsymbol{\theta}_{p} \equiv\left(\boldsymbol{\theta}_{p \text { ind }}, \boldsymbol{\theta}_{p c r}\right)=\left(E\left[\mathbf{T}_{p \text { ind }}\right], E\left[\mathbf{T}_{p c r}\right]\right)$ referring respectively
to $2 p$ and $p(p-1) / 2$ independent and cross moments. The components of $\boldsymbol{\theta}_{p}$ are of the form $m_{i, j} \equiv E\left[X^{i} Y^{j}\right], 1 \leq i+j \leq p$, where the independent part is set to the moment values of the standard Gaussian: $E_{g}\left[X^{p}\right]=0$ for an odd $p$ and $E_{g}\left[X^{p}\right]=(p-1)!!=(p-1)(p-3) \ldots$ for an even $p$ (idem for $Y$ ), where $E_{g}$ denotes expectation with respect to the standard Gaussian. The cross moments are bounded by a simple application of the Schwartz inequality as $\left(m_{i, j}\right)^{2} \leq E_{g}\left[X^{2 i}\right] E_{g}\left[Y^{2 j}\right]$. The monomials of $\mathbf{T}_{p}$ are linearly independent and space-generating by linear combinations, thus forming a basis in the sense of vector spaces.

For an even integer $p$ and constraint set $\mathbf{T}_{p}$ there exists an integrable ME-PDF with support in $\mathbb{R}$, of the form $\rho^{*}(X, Y)=C \exp \left(P_{p}(X, Y)\right)$, where $P_{p}(X, Y)$ is a $p^{\text {th }}$ order polynomial given by a linear combination of monomials in $\mathbf{T}_{p}$ with weights given by Lagrange multipliers and $C$ being a normalizing constant. That happens because $\left|\left(X^{i} Y^{j}\right)\right| \leq O\left(X^{p}+Y^{p}\right), \forall X, Y \in \mathbb{R}, 1 \leq i+j \leq p$, thus allowing for $P_{p}(X, Y) \rightarrow-\infty$ as $|X|,|Y| \rightarrow \infty$, leading to convergence of ME-integrals [18]. For odd $p$, those integrals diverge because the dominant monomials $X^{p}, Y^{p}$ change sign from positive to negative real and therefore the ME-PDF is not well defined.

In order to build a sequence of MI lower bounds and use the procedure of Section 2.3, we have considered the sequence of information moment sets of even order $\left(\mathbf{T}_{2}, \boldsymbol{\theta}_{2}\right),\left(\mathbf{T}_{4}, \boldsymbol{\theta}_{4}\right),\left(\mathbf{T}_{6}, \boldsymbol{\theta}_{6}\right), \ldots$ with any pair of consecutive sets satisfying the premises of Theorem 1, i.e., all independent moment sets are ME-congruent. This will lead to the corresponding monotonically growing sequence of lower bounds of MI, denoted as $I(X, Y)_{g, 2} \leq I(X, Y)_{g, 4} \leq I(X, Y)_{g, 6} \leq \ldots$, where the subscript $g$ means that variables are marginally standard Gaussian. The first term of the sequence is the Gaussian MI $I_{g, 2}(X, Y) \equiv I(X, Y)_{g}$, dependent upon the Gaussian correlation. The difference between the subsequent terms and the first one leads to the non-Gaussian MI of order $p$ defined as $I(X, Y)_{n g, p} \equiv I(X, Y)_{g, p}-I(X, Y)_{g, 2} \leq I(X, Y)_{n g}$, which increases with $p$ and converges to $I(X, Y)_{n g}$ as $p \rightarrow \infty$, under quite general conditions [15].

In the same manner as stated in (12), the lower bound $I(X, Y)_{n g, p}$ of $I(X, Y)_{n g}$ is also a lower bound for the joint negentropy, which is invariant for any affine transformations $\left(X_{r}, Y_{r}\right)^{T}=A(X, Y)^{T}$, i.e.,

$$
\begin{equation*}
I(X, Y)_{n g, p}=H(X, Y)_{2}-H(X, Y)_{p}=H\left(X_{r}, Y_{r}\right)_{2}-H\left(X_{r}, Y_{r}\right)_{p} \tag{15}
\end{equation*}
$$

where $H(X, Y)_{p}$ is the bivariate ME associated to $\left(\mathbf{T}_{p}, \boldsymbol{\theta}_{p}\right)$ and $H(X, Y)_{2}=2 H_{g}-I(X, Y)_{g}$. The successive differences $I(X, Y)_{n g, p+2}-I(X, Y)_{n g, p}$ are non-negative, with the extra MI being explained by moments of orders $p+1$ and $p+2$, while not explained by lower-order moments.

There is no analytical closed formula for the dependence of non-Gaussian MI on cross moments. However, under the scenario of low joint non-Gaussianity (small KL divergence to the joint Gaussian), the ME-PDF can be approximated by the Edgeworth expansion [31], based on orthogonal Hermite polynomials and $I_{n g}$ approximated as a polynomial of joint bivariate cumulants: $k^{[i, j]} \equiv E\left[X_{r}^{i} Y_{r}^{j}\right]-E_{g}\left[X_{r}^{i}\right] E_{g}\left[Y_{r}^{j}\right], i+j>2$, of any pair of uncorrelated standardized variables $\left(X_{r}, Y_{r}\right)^{T}=A(X, Y)^{T}$. Cross-cumulants are nonlinear correlations measuring joint non-Gaussianity [32],
vanishing when $(X, Y)$ are jointly Gaussian. For example $I(X, Y)_{n g, p=4}$ is approximated as in [13] by the sum of squares:

$$
\begin{align*}
& I(X, Y)_{n g(E d, p=4)} \equiv \frac{1}{12}\left[3\left(k^{[2,1]}\right)^{2}+3\left(k^{[1,2]}\right)^{2}+\left(k^{[3,0]}\right)^{2}+\left(k^{[0,3]}\right)^{2}\right] \\
& +\frac{1}{48}\left[\left(k^{[4,0]}\right)^{2}+4\left(k^{[3,1]}\right)^{2}+6\left(k^{[2,2]}\right)^{2}+4\left(k^{[1,3]}\right)^{2}+\left(k^{[0,4]}\right)^{2}\right]+  \tag{16}\\
& \frac{1}{72}\left[5\left(k^{[3,0]}\right)^{4}+30\left(k^{[3,0]} k^{[2,1]}\right)^{2}+\frac{2}{3}\left(9 k^{[3,0]} k^{[1,2]}+6 k^{[2,1]} k^{[2,1]}\right)^{2}+\frac{2}{3}\left(9 k^{[0,3]} k^{[2,1]}+6 k^{[1,2]} k^{[1,2]}\right)^{2}+\right. \\
& \left.\frac{1}{2}\left(2 k^{[3,0]} k^{[0,3]}+18 k^{[2,1]} k^{[1,2]}\right)^{2}+30\left(k^{[1,2]} k^{[0,3]}\right)^{2}+5\left(k^{[0,3]}\right)^{4}\right]=I(X, Y)_{n g}+O\left(n_{e q}^{-3 / 2}\right)
\end{align*}
$$

where ( $X, Y$ ) is assumed to be the arithmetic average of an equivalent number $n_{e q}$ of independent and identically distributed (iid) bivariate RVs. Therefore, from the multidimensional Central Limit Theorem [33], the larger $n_{e q}$ is, the closer the distribution is to joint Gaussianity, and the smaller the absolute value of cumulants become.

### 3.4. Non-Gaussian MI across the Polytope of Cross Moments

The $(X, Y)$ cross moments in the expectation vector $\theta_{p}$ ( $p$ even) are not completely free. Rather, they satisfy to Schwarz-like inequalities defining a compact set $\mathbf{D}_{p}$ within which cross-moments lie. That set portrays all the possible non-Gaussian ME-PDFs with $p$-order independent moments equal to those of the standard Gaussian. Under these conditions $I(X, Y)_{n g}=I(X, Y)_{n g, p}$. In order to have a better feel on how $I_{n g}$ behaves, we have numerically evaluated $I_{n g, p=4}$ along the allowed set of cross-moments.

In order to determine that set, let us begin by invoking some generalities about polynomials. Any bivariate polynomial $\mathscr{P}_{p}(x, y): \mathbb{R}^{2} \rightarrow \mathbb{R}$ of total order $p$ is expressed as a linear combination of linearly independent monomials from the basis $\mathbf{T}_{p}^{1} \equiv \mathbf{T}_{p} \cup\{1\}$, obtained from $\mathbf{T}_{p}$ including unity. Then:

$$
\begin{equation*}
\mathscr{P}_{p}(x, y)=\sum_{i} a_{i} T_{i, p}(x, y), T_{i, p} \in \mathbf{T}_{p}^{1} \tag{17}
\end{equation*}
$$

If the condition $\mathscr{P}_{p}(x, y) \geq 0, \forall(x, y) \in \mathbb{R}^{2}$ holds i.e., $\mathscr{P}_{p}(x, y)$ is positive semi-definite (PSD), then its expectation is non-negative: $E\left[\mathscr{P}_{p}(x, y)\right]=\sum_{i} a_{i} \theta_{i, p} \geq 0$, where $\theta_{i, p} \equiv E\left[T_{i, p}\right]$, thus imposing a constraint on the components of $\theta_{p}$. A sufficient condition for the positiveness is that $\mathscr{P}_{p}(x, y)$ is a sum of squares (SOS) or, without loss of generality, the square of a certain polynomial $\mathbb{Q}_{p / 2}(x, y)=\mathbf{b}^{T} \mathbf{T}_{p / 2}^{1}$ of total order $p / 2$, where $\mathbf{b}$ is a column vector of coefficients multiplying monomials of $\mathbf{T}_{p / 2}^{1}$. Then, $\mathscr{P}_{p}(x, y)$ is written as a quadratic form $\mathscr{P}_{p}(x, y)=\left(\mathscr{Q}_{p / 2}(x, y)\right)^{2}=\mathbf{b}^{T}\left(\mathbf{T}_{p / 2}^{1} \mathbf{T}_{p / 2}^{1}\right) \mathbf{b} \geq 0$. By taking the expectation operator we have $\mathbf{b}^{T} E\left[\left(\mathbf{T}_{p / 2}^{1} \mathbf{T}_{p / 2}^{1}{ }^{T}\right)\right] \mathbf{b} \geq 0, \forall \mathbf{b}$, which implies the positiveness of the matrix of moments $E\left[\left(\mathbf{T}_{p / 2}^{1} \mathbf{T}_{p / 2}^{1}{ }^{T}\right)\right] \equiv \mathbf{C}_{p}$, which is given in terms of components of $\boldsymbol{\theta}_{p}$.

When $p=4$ and $d=2$, the case of bivariate quartics, any PSD polynomial is a SOS [34] and vice versa. However, for $p \geq 6$ there are PSD-non-SOS polynomials (e.g., those coming from the
inequality between arithmetic and geometric means [35]). Therefore, a necessary and sufficient condition among fourth-order moments is that $E\left[\left(\mathbf{T}_{2}^{1} \mathbf{T}_{2}^{1 T}\right)\right] \equiv \mathbf{C}_{4}$ be a PSD matrix. Let us study the conditions for that.

By ordering the set $\mathbf{T}_{2}^{1} \equiv\left(1, x, y, x^{2}, y^{2}, x y\right)^{T}$, one has the $6 \times 6$ matrix $\mathbf{C}_{4}$, written in the simplified form in terms of moments:

$$
\mathbf{C}_{4}=\left[\begin{array}{l}
\left(1,0,0,1,1, c_{g}\right)  \tag{18}\\
\left(0,1, c_{g}, 0, m_{1,2}, m_{2,1}\right) \\
\left(0, c_{g} 1, m_{2,1}, 0, m_{1,2}\right) \\
\left(1,0, m_{2,1}, m_{2,2}, m_{3,1}\right) \\
\left(1, m_{1,2} 0, m_{2,2} 3, m_{1,3}\right) \\
\left(c_{g}, m_{2,1}, m_{1,2}, m_{3,1}, m_{1,3}, m_{2,2}\right)
\end{array}\right]
$$

A necessary and sufficient condition for the positiveness of $\mathbf{C}_{4}$ is given by the application of the Sylvester criterion, stating that the determinants $d_{1}, d_{2}, d_{3}, d_{4}, d_{5}$ and $d_{6}$ of the 6 upper sub-matrices of $\mathbf{C}_{4}$ are positive. From these, only those of orders 4, 5 and 6 lead to nontrivial relationships, given with help of Mathematica ${ }^{\circledR}$ [36] as:

$$
\begin{gather*}
d_{4} \equiv 2-2 c_{g}^{2}-m_{2,1}^{2} \geq 0 ; d_{4, \text { dual }} \equiv 2-2 c_{g}^{2}-m_{1,2}^{2} \geq 0  \tag{19}\\
d_{5} \equiv-2\left(m_{2,1}^{2}+m_{2,1}^{2}\right)+m_{1,2}^{2} m_{2,1}^{2}-2 m_{1,2} m_{2,1}\left(m_{2,2}-1\right) c_{g}+\left(m_{2,2}-3\right)\left(m_{2,2}+1\right)\left(-1+c_{g}^{2}\right) \geq 0  \tag{20}\\
d_{6} \equiv\left(9-m_{2,2}^{2}\right) c_{g}^{4}+2\left[\left(m_{1,3}+m_{3,1}\right)\left(m_{2,2}-3\right)+m_{1,2} m_{2,1} m_{2,2}\right] c_{g}^{3}+ \\
{\left[\begin{array}{l}
\left(m_{2,1}^{2}+m_{1,2}^{2}\right)\left(9-2 m_{2,2}\right)-m_{1,2} m_{2,1}\left(2\left(m_{1,3}+m_{3,1}\right)+m_{1,2} m_{2,1}\right)+ \\
2\left(m_{3,1}^{2}+m_{1,3}^{2}+m_{1,3} m_{3,1}\left(1-m_{2,2}\right)\right)+\left(m_{2,2}-3\right)\left(3+m_{2,2}\left(2+m_{2,2}\right)\right)
\end{array}\right] c_{g}^{2}+} \\
2\left[\begin{array}{l}
\left(3-2 m_{2,1}^{2}-2 m_{1,2}^{2}-m_{2,2}\right)\left(m_{1,3}+m_{3,1}\right)+m_{2,2}\left(m_{1,2}^{2} m_{3,1}+m_{2,1}^{2} m_{1,3}\right)+ \\
m_{1,2} m_{2,1}\left(m_{1,3} m_{3,1}+m_{2,1}^{2}+m_{1,2}^{2}-3+m_{2,2}\left(5-2 m_{2,2}\right)\right)
\end{array}\right] c_{g}+  \tag{21}\\
{\left[\begin{array}{l}
-m_{2,2}^{3}+\left(m_{2,1}^{2}+m_{1,2}^{2}+2\right) m_{2,2}^{2}+\left(3-4\left(m_{2,1}^{2}+m_{1,2}^{2}\right)+m_{1,2} m_{2,1}\left(3 m_{1,2} m_{2,1}-2\left(m_{1,3}+m_{3,1}\right)\right)\right) m_{2,2}+ \\
m_{1,2} m_{2,1}\left(-2\left(m_{1,2}^{2} m_{3,1}+m_{2,1}^{2} m_{1,3}+m_{1,2} m_{2,1}\right)+6\left(m_{1,3}+m_{3,1}\right)\right)+2\left(m_{2,1}^{4}+m_{1,2}^{4}\right)+ \\
-3\left(m_{2,1}^{2}+m_{1,2}^{2}\right)-2\left(m_{1,3}^{2}+m_{3,1}^{2}\right)-2 m_{1,3} m_{3,1}+\left(m_{1,3} m_{2,1}\right)^{2}+\left(m_{3,1} m_{1,2}\right)^{2}
\end{array}\right] \geq 0}
\end{gather*}
$$

In Equation (22), the inequality for $d_{4}$ has a dual relationship ( $d_{4, \text { dual }}$ ), its sign being reversed by swapping the two indices in $m_{i, j}$, whereas $d_{5}$ and $d_{6}$ are symmetric with respect to indicial swap. The term $d_{6}$ of Equation (21) is a fourth-order polynomial of $c_{g}$. The inequalities for $d_{4}, d_{4 d u a l}, d_{5}$ and $d_{6}$ hold inside a compact domain denoted $\mathbf{D}_{4}$, with the shape of what resembles a rounded polytope in the space of cross moments ( $m_{1,2}, m_{2,1}, m_{1,3}, m_{3,1}, m_{2,2}$ ), for each value of the Gaussian correlation $c_{g}$. The case of Gaussianity lies within the interior of $\mathbf{D}_{4}$, corresponding to $m_{1,2}=m_{2,1}=0 ; m_{1,3}=m_{3,1}=3 c_{g} ; m_{2,2}=2 c_{g}^{2}+1$, thus defining the hereafter called one-dimensional ‘Gaussian manifold'.

In order to illustrate how non-Gaussianity depends on moments of third and fourth order, we have computed the non-Gaussian MI of order $4\left(I_{n g, p=4}\right)$, along a set of 2-dimensional cross-sections of $\mathbf{D}_{4}$ crossing the Gaussian manifold and extending up the boundary of $\mathbf{D}_{4}$. For Gaussian moments, $I_{n g, 4}$ vanishes, being approximated by the Edgeworth expansion (16) near the Gaussian manifold.

In order to get a picture of $I_{n g, p=4}$, we have chosen six particular cross-sections of $\mathbf{D}_{4}$ by varying two moments and setting the remaining to their 'Gaussian values'. The six pairs of varying parameters are: $\mathrm{A}\left(c_{g}, m_{2,1}\right), \mathrm{B}\left(c_{g}, m_{3,1}\right), \mathrm{C}\left(c_{g}, m_{2,2}\right), \mathrm{D}\left(m_{2,1}, m_{3,1}\right)$ at $c_{g}=0, \mathrm{E}\left(m_{2,1}, m_{2,2}\right)$ at $c_{g}=0$ and $\mathrm{F}\left(m_{3,1}, m_{2,2}\right)$ at $c_{g}=0$, with the contours of the corresponding $I_{n g}$ field shown in Figure 2a-f. The fields are retrieved from a discrete mesh of $100 \times 100$ in moment space. The Gaussian state lies at: $\left(c_{g}, 0\right),\left(c_{g}, 3 c_{g}\right)$, $\left(c_{g}, 2 c_{g}{ }^{2}+1\right),(0,0),(0,1)$ and $(0,1)$, respectively for cases A up to $F$. The moment domains are obtained by solving inequalities for $d_{4}, d_{5}$ and $d_{6}$ and applying the restrictions imposed by the crossing of the Gaussian manifold (e.g., $m_{1,2}=0, m_{l, 3}=m_{3,1}=3 c_{g}, m_{2,2}=2 c_{g}{ }^{2}+1$ for case A). We obtain the following restrictions for cases $\mathrm{A}-\mathrm{F}$ :

$$
\begin{align*}
& \text { Case A: } m_{2,1}^{2} \leq \frac{1}{2}\left[3+3 c_{g}^{2}\left(1-2 c_{g}^{2}\right)-\left(1-c_{g}^{2}\right)\left(1+52 c_{g}^{2}+28 c_{g}^{4}\right)^{1 / 2}\right]  \tag{22}\\
& \text { Case B: } c_{g}+2 c_{g}^{3}-\sqrt{2\left(1-c_{g}^{2}-c_{g}^{4}+c_{g}^{6}\right)} \leq m_{3,1} \leq c_{g}+2 c_{g}^{3}+\sqrt{2\left(1-c_{g}^{2}-c_{g}^{4}+c_{g}^{6}\right)}  \tag{23}\\
& \text { Case C: } \frac{1}{2}\left[-1+c_{g}^{2}+\left(1+34 c_{g}^{2}+c_{g}^{4}\right)^{1 / 2}\right] \leq m_{2,2} \leq 3  \tag{24}\\
& \text { Case D: } m_{3,1}^{2} \leq 2-3 m_{2,1}^{2}+4 m_{2,1}^{4} ;\left|m_{21}\right| \leq 1 ;\left|m_{21}\right| \leq \sqrt{2}  \tag{25}\\
& \text { Case E: } m_{2,1}^{2} \leq m_{2,2} \leq 1+\left(4-2 m_{2,1}^{2}\right)^{1 / 2} ; m_{2,1}^{2} \leq \sqrt{3}  \tag{26}\\
& \text { Case F: } m_{2,2}^{3}-2 m_{2,2}^{2}-3 m_{2,2}+2 m_{3,1}^{2} \leq 0 \tag{27}
\end{align*}
$$

The analytical boundary of allowed domains is emphasized with thick solid lines in all Figure 2a-f. There are some common aspects among the figures. As expected, $I_{n g}$ vanishes at the Gaussian states, the Gaussian manifold, marked with G in Figures 2. $I_{n g}$ grows monotonically towards the boundary of the moment domains $\mathbf{D}_{4}$. There, $\operatorname{det}\left(\mathbf{C}_{4}\right)=0$, meaning that $\mathbf{C}_{4}$ is singular and there is a vector $(\mathbf{b} \neq 0) \in \operatorname{Ker}\left(\mathbf{C}_{4}\right)$. This holds if one gets the deterministic relationship $\mathscr{Q}_{p / 2}(x, y)=\mathbf{b}^{T} \mathbf{T}_{2}^{1}=0$, leading to a Dirac-Delta-like ME-PDF along a one-dimensional curve. This in turn leads to $I_{n g}=\infty$, except possibly in a set of singular points of $\mathbf{D}_{4}$ on which $I_{n g}$ is not well defined. In practice, infinity is not reached due to stopping criteria for convergence of the iterative method used for obtaining the ME-PDF.

Figure 2. Field of the non-Gaussian MI $I_{n g, 4}$ along 6 bivariate cross-sections of the set of allowed moments. Two varying moments are featured in each cross-section: A ( $\left.c_{g}, m_{2,1}\right)$, $\mathrm{B}\left(c_{g}, \mathrm{~m} 3_{, 1}\right), \mathrm{C}\left(c_{g}, m_{2,2}\right), \mathrm{D}\left(m_{2,1}, m_{3,1}\right)$ at $c_{g}=0, \mathrm{E}\left(m_{2,1}, m_{2,2}\right)$ at $c_{g}=0$ and $\mathrm{F}\left(m_{3,1}, m_{2,2}\right)$ at $c_{g}=0$. The letter G indicates points or curves where Gaussianity holds.


At states where $\left|c_{g}\right|=1, I_{g}=\infty$ and $I_{n g}$ has a second-kind singularity discontinuity where the contours merge together without a well-defined limit for $I_{n g}$. In the neighborhood of the Gaussian state with $c_{g}=0$ in Figure $2 \mathrm{~d}-\mathrm{f}, I_{n g}$ is approximated by the quadratic form (16) as is confirmed by the elliptic shape of $I_{n g}$ contours. The value of $I_{n g}$ can surpass $I_{g}$, thus emphasizing the fact that in some cases much of the MI may come from nonlinear ( $X, Y$ ) correlations.

The joint entropy is invariant for a mirror symmetry in one or both variables: $X \rightarrow-X$ or $Y \rightarrow-Y$, because the absolute value of the determinant of that transformation equals 1 . As a consequence, the dependence of the Gaussian and non-Gaussian MI on moments also reflects these intrinsic mirror symmetries. For instance, in Figure 2d, the symmetry $X \rightarrow-X$ leads to the dependency relations $I_{n g}\left(m_{2,1}, m_{3,1}\right)=I_{n g}\left(m_{2,1},-m_{3,1}\right)$, where arguments are the varying moments, while symmetry $Y \rightarrow-Y$ leads to $I_{n g}\left(m_{2,1}, m_{3,1}\right)=I_{n g}\left(-m_{2,1},-m_{3,1}\right)$. The symmetries corresponding to the remaining figures are: $I_{n g}\left(c_{g}, m_{2,1}\right)=I_{n g}\left(-c_{g}, m_{2,1}\right)=I_{n g}\left(c_{g},-m_{2,1}\right) ; I_{n g}\left(c_{g}, m_{3,1}\right)=I_{n g}\left(-c_{g},-m_{3,1}\right) ; I_{n g}\left(c_{g}, m_{2,2}\right)=I_{n g}\left(-c_{g}, m_{22}\right)$; $I_{n g}\left(m_{2,1}, m_{2,2}\right)=I_{n g}\left(-m_{2,1}, m_{2,2}\right) ; I_{n g}\left(m_{3,1}, m_{2,2}\right)=I_{n g}\left(-m_{3,1}, m_{2,2}\right)$, respectively for Figure 2a-c, e and f .

Near the boundary of $\mathbf{D}_{4}$, the ME problem is very ill-conditioned. The Hessian matrix of the corresponding ME functional is the inverse $\mathbf{M}_{4}^{-1}$, where $\mathbf{M}_{4} \equiv E\left[\mathbf{T}_{4} \mathbf{T}_{4}^{T}\right]-E\left[\mathbf{T}_{4}\right] E\left[\mathbf{T}_{4}^{T}\right]$ is the covariance matrix of $\mathbf{T}_{4}$, made by 8 independent plus 6 cross moments. The condition number CN (ratio between the largest and smallest eigenvalue) of $\mathbf{M}_{4}$ (and of $\mathbf{M}_{4}^{-1}$ ) tends to $\infty$ at the boundary of $\mathbf{D}_{4}$ where $\mathbf{M}_{4}$ is singular, due to the above deterministic relationship. Therefore, the closer that boundary is, the closer the ME-PDF is to a deterministic relationship, the more ill conditioned the ME problem is and the slower the numerical convergence of the optimization algorithm becomes.

## 4. The Effect of Noise and Nonlinearity on Non-Gaussian MI

The aim of this section is an exploratory analysis of the possible sources of non-Gaussianity in a bivariate statistical relationship. Towards that aim, we explore the qualitative behavior of $I_{g}$ and $I_{n g}$ between a standardized signal $\hat{X}$ (with null mean and unit variance) and an $\hat{X}$-dependent standardized response variable $\hat{Y}$ contaminated by noise. For this purpose, a full range of signal-to-noise variance ratios (snr) shall be considered, from pure signal to pure noise. The statistics are evaluated from one-million-long synthetic $(\hat{X}, \hat{Y})$ iid independent realizations produced by a numeric Gaussian random generator. Many interpretations are possible for the output variable: (i) $\hat{Y}$ taken as the observable outcome emerging from a noisy transmission channel fed by $\hat{X}$; (ii) $\hat{Y}$ given by the direct or indirect observation affected by measurement and representativeness errors corresponding to a certain value $\hat{X}$ of the model state vector [37] (iii) the outcome from a stochastic or deterministic dynamical system [38].

In order to estimate $I(\hat{X}, \hat{Y})$, the working variables $(\hat{X}, \hat{Y})$ are transformed by anamorphosis into standard Gaussian variables $(X, Y)$.

We consider, without loss of generality, $X=\hat{X}$. The variable $Y$ is given by Gaussian anamorphosis $Y=G_{\hat{Y}}(\hat{Y}) \sim N(0,1)$ as in Equation (8), with:

$$
\begin{equation*}
\hat{Y}=s F(X)+\left(1-s^{2}\right)^{1 / 2} n(X, \mathbf{W}) ; s \in[0,1] \tag{28}
\end{equation*}
$$

where $F(X)$ is a purely deterministic transfer function and $n(X, \mathbf{W})$ is a scalar noise uncorrelated with $F(X)$, depending in general on $X$ (e.g., multiplicative noise) and from a vector $\mathbf{W}$ of independent Gaussians contaminating the signal. Both $F(X)$ and $n(X, \mathbf{W})$ have unit variance with $n(X, \mathbf{0})=0$. The signal-to-noise variance ratio is $s n r=s^{2} /\left(1-s^{2}\right)$.

Then, the Gaussian MI $I_{g}$ is computed for each value of $s \in[0,1]$ and compared among several scenarios of $F(X)$ and $n(X, \mathbf{W})$. A similar comparison is done for the non-Gaussian MI, approximated here by $I_{n g, p=8}$. Six case studies have been considered (A, B, C, D, E and F); their signal and noise terms are summarized in Table 1, along with the colors with which they are represented in Figure 3 further below.

Table 1. Types of signal and noise in Equation (35) and corresponding colors used in Figure 3.

| Case study | $F(X)$ | $n(X, \mathbf{W})$ | Color |
| :--- | :--- | :--- | :--- |
| A—Gaussian noise (reference) | $X$ | $W$ | Black |
| B-Additive non-Gaussian independent <br> noise | $X$ | $W^{3} / \sqrt{15}$ | Red |
| C—Multiplicative noise | $X$ |  |  |
| D—Smooth nonlinear homeomorphism | $X^{3} / \sqrt{15}$ | $W$ | Blue |
| E—Smooth non-injective transformation | $\left(X^{3}-X\right) / \sqrt{10}$ | $W$ | Magenta |
| F—Combined non-Gaussianity | $\left(X^{3}-X\right) / \sqrt{10}$ | $X W_{1} / \sqrt{2}+W_{2}^{3} / \sqrt{30}$ | Cyan |

In Table $1, W, W_{1}, W_{2}$ are independent standard Gaussian noises. We begin with a reference case, A, which refers to Gaussian noise. Case B refers to a symmetric leptokurtic (i.e., with kurtosis larger than that of the Gaussian) non-Gaussian noise. In case $C$ the multiplicative noise depends linearly on the
signal $X$. In case D , the signal is a nonlinear cubic homeomorphism of the real domain. For case E , the signal is nonlinear and not injective in the interval $[-1,1]$, thus introducing ambiguity in the relationship $(X, Y)$. Finally, in case F all the factors-non-Gaussian noise, multiplicative noise and signal ambiguity-are pooled together.

Figure 3. Graphs depicting the total MI (a), Gaussian MI (b) and non-Gaussian MI (c) of order 8 for 6 cases (A-F) of different signal-noise combinations with the signal weight $s$ in abscissas varying from 0 up to 1 . See text and Table 1 for details about the cases and their color code.


Figure 3a,b show the graphics of the total MI, estimated by $I_{g}+I_{n g, 8}$ and of the Gaussian MI $\left(I_{g}\right)$ for the six cases (A to F). The graphic of non-Gaussian MI as approximated by $I_{n g, 8}$ is depicted in Figure 3c for five cases (B to F). In Figure 4, we show a 'stamp-format' collection of the contouring of ME-PDFs of polynomial order $p=8$ for all cases (A to F) and extreme and intermediate cases of the snr: $s=0.1, s=0.5$ and $s=0.9$. This illustrates how the $s n r$ and the nature of both the transfer function and noises influence the PDFs.

For the Gaussian noise case (A), the non-Gaussian MI is theoretically set to zero since the joint distribution of $(X, Y)$ is Gaussian. In all scenarios, both $I_{g}$ and the total MI $I_{g}+I_{n g}$ grow, as expected, with the snr. This is in accordance to the Bruijn's equality stating the positiveness of the MI derivative with respect to $s n r$ and established in the literature of signal processing for certain types of noise [39,40]. On the contrary, the monotonic behavior as a function of snr is not a universal characteristic of the non-Gaussian MI.

By observing Figure 3a-c, the following qualitative results are worth mentioning. We begin by comparing the total MI in three cases (A, B and C), which share the same linear signal but feature noise of different kinds (Figure 3a). Both the red (B) and blue (C) lines lie above the black line (A) for each given $s$, thus indicating that the total MI is lowest when the noise is Gaussian. This means that the Gaussian noise is the most signal degrading of noises with the same variance [41]. The extra MI found in the B and C cases come, respectively, from the Gaussian MI (see case B in Figure 3c) and from the
non-Gaussian MI (see case C in Figure 3a), as it is also apparent by looking at ME-PDFs for cases B, C ( $s=0.1$ ) (Figure 4).

We consider now the cases A, D, E, all of which have a Gaussian noise. Their differences lie in the signals, with the one in A being linear and the ones in D and E being nonlinear. By comparing these cases it is seen that $I_{g}$ is highest for the linear signal, the black curve lying above the magenta (D) and green (E) curves for each $s$ in Figure 3b. This indicates that the Gaussian MI, measuring the degree of signal linearity, is lower when the signal introduces nonlinearity (cases D and E) than when no nonlinearity is present (case A).

Figure 4. Collection of stamp-format ME-PDFs for cases $\mathbf{A}-\mathbf{F}$ (see text for details) and signal weight $s=0.1$ (a), $s=0.5$ (b) and $s=0.9$ (c) over the $[-3,3]^{2}$ support set.


It is worth noting that, while the signals in A and D are injective, the one in E is not, thus introducing ambiguity. This will imply loss of information in E, which is visible in the total MI depicted for each $s$ in Figure 3a. In fact, there the green curve (case E) lies lower than the black (A) and magenta (D) curves for every $s$. The effect of nonlinearity is quite evident in ME-PDFs, in particular for high $s$ value (Figure 4, cases D, E, $s=0.9$ ).

We focus now on the non-Gaussian MI, depicted in Figure 3c for each $s$. The curve representing the case B , with a linear signal and a state-independent noise, indicates that the non-Gaussian MI is null for both $s=0$ and $s=1$. The first zero of non-Gaussian MI (at $s=0$ ) is justified by the noise being state-independent, whereas the second zero (at $s=1$ ) is due to the signal being linear, which means that all the MI resides in the Gaussian MI. The non-Gaussian MI is thus positive and maximum at intermediate values of $s$.

By looking at case C (multiplicative noise), it is seen that the non-Gaussian MI remains roughly unchanged for every $s<1$. This holds even at $s=0$ (pure noise), since the noise is state-dependent and thus some information is already present. At $s=1$ the non-Gaussian MI is null due to the signal being linear (as in case B).

By observing the cases with Gaussian noise and nonlinear signals (D and E) in Figure 3c, it can be seen that their non-Gaussian MI grows with $s$ (and thus with the relative weight of the signal), due to
their signals being nonlinear. This gradual behavior is also reflected in the ME-PDFs (Figure 4, cases D, E along the $s$ values).

Finally, we consider the case in which the signal is nonlinear and the noise comprises a multiplicative and a non-Gaussian additive component (case F). As compared with E (which differs from it in that the noise is Gaussian), it can been that non-Gaussian MI is always larger in F independently of $s$. This is due to the fact that in F there is information even at $s=0$, due to the state-dependence of its noise. For all values of $s$, the ME-PDF exhibits quite a large deviation from Gaussianity.

## 7. Discussion and Conclusions

We have addressed the problem of finding the minimum mutual information (MinMI), or the least noncommittal MI between $d=2$ random variables, consistent with a set of marginal and joint expectations. The MinMI is a proper MI lower bound when marginals are set to ME-PDFs through appropriate nonlinear single anamorphoses. Moreover, the MinMI increases as long as one increases number of independent cross-constraints of the bivariate ME problem. Considering a sequence of moments, we have obtained a hierarchy of lower MI bounds approximating the total MI value. The method can easily be generalized for $d>2$ variables with the necessary adaptations.

One straightforward application of that principle follows from the MI estimation from 'Gaussianized' variables with real support, where the marginals are rendered standard Gaussian $\mathrm{N}(0,1)$ by Gaussian anamorphosis. This allows for the MI decomposition into two positive contributions: a Gaussian term $I_{g}$, which depends uniquely on the Gaussian correlation $c_{g}$ (Pearson correlation in the space of 'Gaussianized' variables), and a non-Gaussian term $I_{n g}$ depending on nonlinear correlations. This term is equal to the joint negentropy, which is invariant for any oblique or orthogonal rotation of the 'Gaussianized' variables and is related to the 'compactness' measure or the closeness of the PDF towards a low manifold deterministic relationship. The Gaussian MI is also a 'concordance' measure, invariant for any monotonically growing homeomorphisms of marginals and consequently expressed as a functional of the copula-density function, which is exclusively dependent on marginal cumulated probabilities. In certain extreme cases, very far from Gaussianity, the Pearson correlation among non-Gaussian variables is not a proper measure of the mutual information. An example of that situation is given.

Cross moments under marginal standard Gaussians are bounded by Schwarz-like inequalities defining compact sets, the shape of which resemble a rounded polytope where cross moments live. The allowed moment values portray all possible joint PDFs with Gaussian marginals. Inside that set lies the so called one-dimensional Gaussian manifold, parametrized by $c_{g}$, where joint Gaussinity holds. There, $I_{n g}$ vanishes, growing towards infinity as far as the boundary is approached, where variables satisfy a deterministic relationship and the ME problem is ill conditioned. This behavior is illustrated in cross-sections of the polytope of cross moments of total order $p=4$.

In order to systematize the possible sources of Gaussian and non-Gaussian MI, we have computed it in the context of nonlinear noisy channels. The MI has been computed between a Gaussian input and a panoply of (linear and/or nonlinear) outputs contaminated by different kinds of noise for a full range of the signal-to-noise variance ratio. Sources of non-Gaussian MI include: (a) the nonlinearity of the
signal transfer function, (b) multiplicative noise and (c) non-Gaussian additive noise. This paper is followed by a companion one [27] on the estimation of non-Gaussian MI from finite samples with practical applications.

## Acknowledgments

This research was developed at IDL with the financial support of project PEST-OE/CTE/LA0019/2011-FCT. Thanks are due to two anonymous reviewers, Miguel Teixeira and J. Macke for some discussions and our families for the omnipresent support.

## References

1. Cover, T.M.; Thomas, J.A. Elements of Information Theory; John Wiley \& Sons, Inc.: New York, NY, USA, 1991.
2. Ebrahimi, N.; Soofi, E.S.; Soyer R. Information measures in perspective. Int. Stat. Rev. 2010, 78, 383-412.
3. Stögbauer, H.; Kraskov, A.; Astakhov, S.A.; Grassberger, P. Least-dependent-component analysis based on mutual information. Phys. Rev. E 2004, 70, 066123:1-066123:17.
4. Hyvärinen, A.; Oja, E. Independent Component Analysis: Algorithms and application. Neural Network. 2000, 13, 411-430.
5. Fraser, H.M.; Swinney, H.L. Independent coordinates for strange attractors from mutual information. Phys. Rev. A 1986, 33, 1134-1140.
6. DelSole, T. Predictability and information theory. Part I: Measures of predictability. J. Atmos. Sci. 2004, 61, 2425-2440.
7. Majda, A.; Kleeman, R.; Cai, D. A mathematical framework for quantifying predictability through relative entropy. Methods and applications of analysis. Meth. Appl. Anal. 2002, 9, 425-444.
8. Darbellay, G.A.; Vajda, I. Entropy expressions for multivariate continuous distributions. IEEE Trans. Inform. Theor. 2000, 46, 709-712.
9. Nadarajah, S.; Zografos, K. Expressions for Rényi and Shannon entropies for bivariate distributions. Inform. Sci. 2005, 170, 173-189.
10. Khan, S.; Bandyopadhyay, S.; Ganguly, A.R.; Saigal, S.; Erickson, D.J.; Protopopescu, V.; Ostrouchov, G. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. Phys. Rev. E 2007, 76, 026209:1-026209:15.
11. Walters-Williams, J.; Li, Y. Estimation of mutual information: A survey. Lect. Notes Comput. Sci. 2009, 5589, 389-396.
12. Globerson, A.; Tishby, N. The minimum information principle for discriminative learning. In Proceedings of the 20th conference on Uncertainty in artificial intelligence, Banff, Canada, 7-11 July 2004; pp. 193-200.
13. Jaynes, E.T. On the rationale of maximum-entropy methods. Proc. IEEE 1982, 70, 939-952.
14. Wackernagel, H. Multivariate Geostatistics-An Introduction with Applications; Springer-Verlag: Berlin, Germany, 1995.
15. Shams, S.A. Convergent iterative procedure for constructing bivariate distributions. Comm. Stat. Theor. Meth. 2010, 39, 1026-1037.
16. Ebrahimi, N.; Soofi, E.S.; Soyer, R. Multivariate maximum entropy identification, transformation, and dependence. J. Multivariate Anal. 2008, 99, 1217-1231.
17. Abramov, R. An improved algorithm for the multidimensional moment-constrained maximum entropy problem. J. Comput. Phys. 2007, 226, 621-644.
18. Abramov, R. The multidimensional moment-constrained maximum entropy problem: A BFGS algorithm with constraint scaling. J. Comput. Phys. 2009, 228, 96-108.
19. Abramov, R. The multidimensional maximum entropy moment problem: A review on numerical methods. Commun. Math. Sci. 2010, 8, 377-392.
20. Rockinger, M.; Jondeau, E. Entropy densities with an application to autoregressive conditional skewness and kurtosis. J. Econometrics 2002, 106, 119-142.
21. Pires, C.A.; Perdigão, R.A.P. Non-Gaussianity and asymmetry of the winter monthly precipitation estimation from the NAO. Mon. Wea. Rev. 2007, 135, 430-448.
22. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. Phys. Rev. E 2004, 69, 066138:1-066138:16.
23. Myers, J.L.; Well, A.D. Research Design and Statistical Analysis, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2003.
24. Calsaverini, R.S.; Vicente, R. An information-theoretic approach to statistical dependence: Copula information. Europhys. Lett. 2009, 88, 68003; pp. 1-6.
25. Monahan, A.H.; DelSole, T. Information theoretic measures of dependence, compactness, and non-Gaussianity for multivariate probability distributions. Nonlinear Proc. Geoph. 2009, 16, 57-64.
26. Guo, D.; Shamai, S.; Verdú, S. Mutual information and minimum mean-square error in Gaussian channels. IEEE Trans. Inform. Theor. 2005, 51, 1261-1283.
27. Pires, C.A.; Perdigão, R.A.P. Minimum mutual information and non-Gaussianity through the maximum entropy method: Estimation from finite samples. Entropy 2012, submitted for publication.
28. Shore, J.E.; Johnson, R.W. Axiomatic derivation of the principle of maximum entropy and the principle of the minimum cross-entropy. IEEE Trans. Inform. Theor. 1980, 26, 26-37.
29. Koehler, K.J.; Symmanowski, J.T. Constructing multivariate distributions with specific marginal distributions. J. Multivariate Anal. 1995, 55, 261-282.
30. Cruz-Medina, I.R.; Osorio-Sánchez, M.; García-Páez, F. Generation of multivariate random variables with known marginal distribution and a specified correlation matrix. InterStat 2010, 16, 19-29.
31. van Hulle, M.M. Edgeworth approximation of multivariate differential entropy. Neural Computat. 2005, 17, 1903-1910.
32. Comon, P. Independent component analysis, a new concept? Signal Process. 1994, 36, 287-314.
33. van der Vaart, A.W. Asymptotic Statistics; Cambridge University Press: New York, NY, USA, 1998.
34. Hilbert, D. Über die Darstellung definiter Formen als Summe von Formenquadraten. Math. Ann. 1888, 32, 342-350.
35. Ahmadi, A.A.; Parrilo, P.A. A positive definite polynomial hessian that does not factor. In Proceedings of the Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, Shanghai, China, 15-18 December 2009; pp. 16-18.
36. Wolfram, S. The Mathematica Book, 3rd ed.; Cambridge University Press: Cambridge, UK, 1996.
37. Bocquet, M.; Pires, C.; Lin, W. Beyond Gaussian statistical modeling in geophysical data assimilation. Mon. Wea. Rev. 2010, 138, 2997-3023.
38. Sura, P.; Newman, M.; Penland, C.; Sardeshmuck, P. Multiplicative noise and non-Gaussianity: A paradigm for atmospheric regimes? J. Atmos. Sci. 2005, 62, 1391-1406.
39. Guo, D.; Shamai, S.; Verdú, S. Mutual information and minimum mean-square error in Gaussian channels. IEEE Trans. Inform. Theor. 2005, 51, 1261-1283.
40. Rioul, O. A simple proof of the entropy-power inequality via properties of mutual information. arXiv 2007, arXiv:cs/0701050v2 [cs.IT].
41. Guo, D.; Wu, Y.; Shamai, S.; Verdú, S. Estimation in gaussian noise: Properties of the minimum mean-square error. IEEE T. Inform. Theory 2011, 57, 2371-2385.
42. Gilbert, J.; Lemaréchal, C. Some numerical experiments with variable-storage quasi-Newton algorithms. Math. Program. 1989, 45, 407-435.

## Appendix 1

## Form and Numerical Estimation of ME-PDFs

We hereby present a summary of the ME method for distributions along with the numerical algorithm for computing it. Let us consider an information moment set (2) (T, $\boldsymbol{\theta}$ ) comprising $J$ constraints on the RVs ( $X, Y$ ) with the support set $S=S_{X} \otimes S_{Y} \subseteq \mathbb{R}^{2}$. The associated ME-PDF is obtained from the absolute minimum of a control function:

$$
L(\boldsymbol{\eta}, \mathbf{T}, \boldsymbol{\theta})=\log [Z(\boldsymbol{\eta})]-\sum_{i=}^{J} \eta_{i} \theta_{i}
$$

in terms of a $J$-dimensional vector of Lagrange multipliers:

$$
\boldsymbol{\eta}=\left(\eta_{1}, \ldots, \eta_{J}\right)^{T}
$$

where $Z(\boldsymbol{\eta})=\iint_{S} \exp \left[\sum_{i=1}^{J} \eta_{i} T_{i}(x, y)\right] d x d y$ [13]. The minimum of $L$ lies at a value of $\boldsymbol{\eta}$ dependent on $(\mathbf{T}, \boldsymbol{\theta})$, given by: $\boldsymbol{\lambda}(\mathbf{T}, \boldsymbol{\theta})=\arg \min _{\eta} L(\boldsymbol{\eta}, \mathbf{T}, \boldsymbol{\theta})$. The corresponding $L$ minimum is the value of the maximum entropy: $H_{\rho_{\mathrm{T}, ⿹ 勹}^{*}}=L(\boldsymbol{\lambda}, \mathbf{T}, \boldsymbol{\theta})$.

The ME-PDF is of the form: $\rho_{\mathrm{T}, \boldsymbol{\theta}}^{*}(x, y)=Z(\boldsymbol{\lambda})^{-1} \exp \left[\sum_{i=1}^{J} \lambda_{i} T_{i}(x, y)\right]$, where $Z(\boldsymbol{\lambda})$ is the normalization partition function. Except when no analytical relationship $\lambda(\mathbf{T}, \boldsymbol{\theta})$ exists, this function has to be estimated by iterative techniques of minimization of $L(\boldsymbol{\eta}, \mathbf{T}, \boldsymbol{\theta})$. The numerical algorithm consists of a bivariate version of that presented in [20]. In practice, we have solved the ME problem for a finite square support set $S_{r}=[-r, r]^{2}$ with $r$ large enough in order to prevent significant boundary effects on the ME-PDF, thus obtaining a good estimation of the ME-PDF asymptotic limit when
$r \rightarrow \infty$. By using the two-dimensional Leibnitz differentiation rule, it is easy to obtain the derivative of the ME $H_{\rho^{*}}$ with respect to $r$ :

$$
\begin{equation*}
\frac{d H_{\rho^{\circ}}(X, Y)}{d r}=\oint_{(x, y) \in \theta S_{r}} \rho^{*}(x, y)|d l| \tag{A1}
\end{equation*}
$$

where the line integral is always positive and computed along the boundary line $\partial S_{r}$ of $S_{r}$. When $|x|,|y| \rightarrow \infty$, the bound of (A1) leads to the scaling of the logarithm of the ME-PDF as $O\left(-r^{p}\right)$, where $p$ is the maximum total order of the constraining bivariate monomials in the information moment set $(\mathbf{T}, \boldsymbol{\theta})$. Furthermore, in order to get integrands of the order $\exp (O(1))$ during the optimization process, we solve the ME problem for the scaled variables $(X / r, Y / r)$ in the square $[-1,1]^{2}$, by taking the appropriate scaled constraints. Then, we apply the scaling entropy relationship:

$$
\begin{equation*}
H_{\rho^{*}}(X, Y)=H_{\rho^{*}}(X / r, Y / r)+2 \log (r) \tag{A2}
\end{equation*}
$$

The integrals giving the functional $L(\boldsymbol{\eta}, \mathbf{T}, \boldsymbol{\theta})$ and its $\eta$-derivatives are approximated by the bivariate Gauss truncation rule with $N_{f}$ weighting factors each in the interval $[-1,1]$. In order to get full resolution during the minimization, and to avoid "Not-a-Number" (NAN) and infinity (INF) errors in computation, we subtract the polynomials in the arguments of exponentials from the corresponding maximum in $S$. Finally, the functional $L$ is multiplied by a sufficient high factor $F$ in order to emphasize the gradient. After some preliminary experiments, we have set $r=6, N_{f}=80, F=1000$. Convergence is assumed to be reached when one gets an accuracy of $10^{-6}$ for the gradient of $L$. By setting the first guess of Lagrange multipliers (FGLM) to zero, convergence is reached after about $60-400$ iterations. For the optimization we have used the routine M1QN3 from INRIA [42], which uses the Quasi-Newton BFGS algorithm. Convergence is slower under a higher condition number ( CN ) of the Hessian of $L$, with the convergence time growing in general with the proximity of the boundary of the domain of allowed moments as well as the total maximum order $p$ of constraining monomials. The convergence is faster when a closer FGLM is provided to the exact solution. This is possible in sequential ME problems with quite small successive constraint differences. There, one uses a continuation technique by setting FGLM to the optimized Lagrange multipliers from the previous ME problem. This technique has been used in the computation of the graphics in Figure 3.

## Appendix 2

Proof of Theorem 1: Since $X$ follows the ME-PDF generated by $\left(\mathbf{T}_{X}, \boldsymbol{\theta}_{X}\right)$ and ME-congruency holds, we have $D\left(\rho_{X} \| \rho_{\mathrm{T}_{X}, \theta_{X}}^{*}\right)=D\left(\rho_{X} \| \rho_{\mathrm{T}_{X 1}, \theta_{X 1}}^{*}\right)=D\left(\rho_{X} \| \rho_{\mathrm{T}_{X 2}, \theta_{X}}^{*}\right)=0$ with similar identities for $Y$. Therefore, the first inequality of (7a) follow directly from (5). The second one is obtained by difference and application of Lemma 1 to $\mathbf{T}_{1} \subseteq \mathbf{T}_{2}$ (q.e.d.).

Proof of Theorem 2: The first equality of (11) comes from Equation (1) as $I(X, Y)=2 H_{g}-H_{\rho_{X Y}}$ and from the negentropy definition of Gaussianized variables $J(X, Y)=H_{g}(X, Y)-H_{\rho_{X Y}}=2 H_{g}+I_{g}(X, Y)-H_{\rho_{X Y}}$, where $H_{g}(X, Y)$ is the entropy of the Gaussian fit. From the entropy formula of transformed variables we get $H_{\rho_{X, Y_{r}}}-H_{\rho_{X Y}}=H_{g}\left(X_{r}, Y_{r}\right)-H_{g}(X, Y)=\log |\operatorname{det} A|$, leading to the negentropy equality
$J(X, Y)=J\left(X_{r}, Y_{r}\right)$. Finally, the last equation of (11) comes from the equality $H_{\rho_{X_{r} r_{r}}}=H_{\rho_{X_{r}}}+H_{\rho_{Y_{r}}}-I\left(X_{r}, Y_{r}\right)$ and the definition of negentropy, i.e., $J\left(X_{r}\right)=H_{g}-H_{\rho_{X_{r}}}$ (q.e.d.).
© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

