

Article

Minimum Mutual Information and Non-Gaussianity through the Maximum Entropy Method: Estimation from Finite Samples

Carlos A. L. Pires ^{1,*} and Rui A. P. Perdigão ²

¹ Instituto Dom Luiz (IDL), University of Lisbon (UL), Lisbon, P-1749-016, Portugal

² Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, A-1040, Austria; E-Mail: perdigao@hydro.tuwien.ac.at

* Author to whom correspondence should be addressed; E-Mail: clpires@fc.ul.pt;
Tel.: +351-21-750-0886; Fax: +351-21-750-0807.

Received: 8 November 2012; in revised form: 15 February 2013 / Accepted: 19 February 2013 /
Published: 25 February 2013

Abstract: The Minimum Mutual Information (MinMI) Principle provides the least committed, maximum-joint-entropy (ME) inferential law that is compatible with prescribed marginal distributions and empirical cross constraints. Here, we estimate MI bounds (the MinMI values) generated by constraining sets $\mathbf{T}_{cr}$ comprehended by m_{cr} linear and/or nonlinear joint expectations, computed from samples of N iid outcomes. Marginals (and their entropy) are imposed by single morphisms of the original random variables. N -asymptotic formulas are given both for the distribution of cross expectation's estimation errors, the MinMI estimation bias, its variance and distribution. A growing $\mathbf{T}_{cr}$ leads to an increasing MinMI, converging eventually to the total MI. Under N -sized samples, the MinMI increment relative to two encapsulated sets $\mathbf{T}_{cr1} \subset \mathbf{T}_{cr2}$ (with numbers of constraints $m_{cr1} < m_{cr2}$) is the test-difference $\delta H = H_{\max 1, N} - H_{\max 2, N} \geq 0$ between the two respective estimated MEs. Asymptotically, δH follows a Chi-Squared distribution $\frac{1}{2N} \chi^2_{(m_{cr2}-m_{cr1})}$ whose upper quantiles determine if constraints in $\mathbf{T}_{cr2}/\mathbf{T}_{cr1}$ explain significant extra MI. As an example, we have set marginals to being normally distributed (Gaussian) and have built a sequence of MI bounds, associated to successive non-linear correlations due to joint non-Gaussianity. Noting that in real-world situations available sample sizes can be rather low, the relationship between MinMI bias, probability density over-fitting and outliers is put in evidence for under-sampled data.

Keywords: mutual information; non-Gaussianity; maximum entropy distributions; Entropy bias; mutual information distribution; morphism

MSC2000 Codes: 62B10, 94A17

1. Introduction

1.1. The State of the Art

The seminal work of Shannon on Information Theory [1] gave rise to the concept of Mutual Information (MI) [2] as a measure of probabilistic dependence among random variables (RVs), with a broad range of applications, including neuroscience [3], communications and engineering [4], physics, statistics, economics [5], genetics [6], linguistics [7] and geosciences [8]. MI is the positive difference between two Shannon entropies of the RVs: the one assuming statistical independence (H_{ind}) and the other (H_{dep}) considering their true dependence.

This paper addresses the problem of estimating the MI conveyed by the least committed, inferential law (say the conditional probability density function *pdf* $\rho(Y | X)$) between random variables $RVs Y, X$), which is compatible with prescribed marginal distributions and a set T_{cr} of m_{cr} empirical non-redundant cross constraints (e.g., a set of cross expectations between a stimulus X and a response Y , for example in a neural cell, the Earth's climate, an ecosystem). The constrained MI or the Minimum Mutual Information (MinMI) among RVs Y, X is: $I_{min}(X, Y) = H(X) + H(Y) - H_{max}(X, Y) = H(Y) - H_{max}(Y | X)$, obtained after subtraction to the sum of fixed marginal entropies of the maximum joint entropy (ME) H_{max} , compatible with imposed cross constraints. The solution comes from application of the MinMI principle [9,10]. The MinMI is a MI lower bound depending on the marginal *pdfs* (e.g., Gaussians, Uniforms, Gammas), as well as the particular form of the cross expectations in T_{cr} (e.g., linear and non-linear correlations). There are only a few cases of known closed formulas for the MinMI and $m_{cr}=1$: (a) Gaussian marginals and Pearson linear correlation [8,11,12] and (b) Uniform marginals and rank linear correlation [11]. The authors have presented in [12] (PP12 hereafter), a general formalism for computing, though not in an explicit form, the MinMI in terms of multiple ($m_{cr} > 1$) linear and nonlinear cross expectations included in T_{cr} . This set can consist of a natural population constraint (e.g., a specific neural behavior) or it can grow without limit through additional expectations computed within a sample with the MinMI increasing and converging eventually to the total MI. This paper is the natural follow-up of PP12 [12], studying now the statistics (mean or bias, variance and distribution) of the MinMI estimation errors: $\Delta I_{min,N} = -\Delta H_{max,N} \equiv -(H_{max,N} - H_{max})$ where $H_{max,N}$ is the ME estimation issued from N -sized samples of *iid* outcomes. Those errors are roughly similar to those of MI and entropy generic estimator's errors (see [13,14] for a thorough review and performance comparisons between MI estimators). Their mean (bias), variance and higher-order moments are written in terms of N^{-1} powers, thus covering intermediate and asymptotic N ranges [15], with specific applications in neurophysiology [16,17,18]. Entropy estimators range from: (a) the histogram-based plug-in one [19] with a negative bias or the Miller-Madow correction [20] equal to $-(m-1)/(2N)$, where m is the number of univariate histogram bins to much more improved

estimators (e.g., kernel density estimators, adaptive or non-adaptive grids, next nearest neighbors) and others specially designed for small samples [21,22]

1.2. The Rationale of the Paper

The well-posedness of a MinMI $I_{\min}(X, Y)$ compatible with available cross information needs the knowledge of marginal X and Y PDFs, ρ_X and ρ_Y , either imposed or inferred from sufficiently long samples. For that purpose, we can change X and Y into the cumulated probabilities $u(x) = \int^x \rho_X(t) dt$; $v(y) = \int^y \rho_Y(t) dt$, which are uniform RVs on the interval $[0, 1]$ (i.e., copulas [23]), through appropriate smoothly growing (injective) morphisms (or anamorphoses), while leaving the MI invariant [2]. Then, the MI $I(X, Y)$ becomes the negative copula entropy [24,25] given by $I(X, Y) = \int_0^1 \int_0^1 c[u, v] \log(c[u, v]) du dv$, where the copula density is $c[u, v] = \rho_{XY}(x, y) / [\rho_X(x) \rho_Y(y)]$.

The MinMI, subjected to m_{cr} constraints of type $E[T_i(u, v)] = \theta_i$; $i = 1, \dots, m_{cr}$ in the copula-space, is readily obtained by variational analysis (as in the ME method [2]) for $c[u, v] = \exp[-1 + \lambda_u(u) + \lambda_v(v) + \sum_{i=1}^{m_{cr}} \lambda_i T_i(u, v)]$, where the Lagrange multipliers $\lambda_u(u), \lambda_v(v), \lambda_i$ correspond respectively to the preset (not subjected to sampling) continuum of constraints: $\int c[u, v] du = \int c[u, v] dv = 1$ and to the m_{cr} expectations (subjected to sampling error). The general solution is rather tricky since all the values $\lambda_u(u), \lambda_v(v), \lambda_i$ are implicitly related. The constrained joint PDF and the inferential law are recovered from the constrained copula through the product: $\rho_{XY}(x, y) = c[u, v] \rho_X(x) \rho_Y(y)$.

In PP12 [12], we have generalized this problem to a less constrained MinMI version by changing marginal RVs into ME prescribed ones—the ME-morphisms (e.g., standard Gaussians)—and imposing a finite set of marginal constraints instead of the full marginal PDFs. Under these conditions, the number of control Lagrange multipliers is finite, leaving the possibility of using nonlinear minimization algorithms for the MinMI estimation, as already tested in [8]. The MinMI subjected to a set $\mathbf{T}_{cr}$ of m_{cr} cross constraints is thus given by $H_{ind} - H_{ME,cr}$, where $H_{ME,cr}$ is the joint ME and H_{ind} is the sum of single fixed (preset) entropies. The MinMI estimator is written as $H_{ind} - H_{ME,cr,N}$, where $H_{ME,cr,N}$ is the ME constrained by the m_{cr} sampling expectations obtained from N -sized samples. The MinMI estimation error is $H_{ME,cr} - H_{ME,cr,N}$. Therefore, as a generalization of the ME estimator bias [26], one verifies a MinMI positive bias equal to (larger/smaller than) $m_{cr} / (2N)$ when the true population PDF including the tested sample, follows (is more leptokurtic/platykurtic than) the ME-PDF. This result is supported through Monte-Carlo experiments.

Moreover, we introduce here the positive incremental MinMI given by the difference $H_{ME,cr1} - H_{ME,cr2}$ between two MEs, forced by cross constraint sets $\mathbf{T}_{cr1} \subseteq \mathbf{T}_{cr2}$, which is interpreted as the MinMI coming from the difference set $\mathbf{T}_{cr2} / \mathbf{T}_{cr1}$. The corresponding estimator is $H_{ME,cr1,N} - H_{ME,cr2,N}$. Both the MinMI and incremental MinMI estimators depend basically on errors of the expectations estimated from finite N -sized samples.

In particular, under the null hypothesis H_0 that $H_{ME,cr1} = H_{ME,cr2}$ or $\mathbf{T}_{cr1}, \mathbf{T}_{cr2}$ ME-congruent (see definition in PP12, [12]), the difference $H_{ME,cr1,N} - H_{ME,cr2,N}$ works as a significance test of H_0 .

Those tests can be used: (1) for testing statistical significant MI above zero or significant RV dependence or (2) for testing MI due to nonlinear correlations beyond MI due to linear correlations. Another important case (verified here) is the test of MI explained by joint non-Gaussianity beyond the MI explained by joint Gaussianity, in which Gaussian morphism (*i.e.*, bijective, reversible variable transformation into another with a Gaussian pdf without loss of generality) is used for single variables. According to the above result, the bias of $H_{ME,cr1,N} - H_{ME,cr2,N}$, subjected to H_0 is $(m_{cr2} - m_{cr1})/2N$, *i.e.*, the number of cross constraints in the difference set $\mathbf{T}_{cr2}/\mathbf{T}_{cr1}$ divided by $2N$.

We further provide asymptotic analytical N -scaled formulas for the variance and distribution of MinMI estimation errors as functions of statistics of the ME cross constraints estimation errors. This is possible for N high enough where expectation errors are closely governed by a multivariate Gaussian distribution, uniquely determined by their bias and covariance matrix, thanks to the multivariate Central Limit Theorem. Since marginal morphisms are performed, the single variables are set to values from a look-up table of fixed quantiles (not subjected to sampling) and therefore the estimator's squared-bias decreases faster than the estimator's variance as $N \rightarrow \infty$.

The correct modeling of covariances between sampling expectation's errors under morphism is crucial for the correct computation of MinMI error statistics. We have verified an overall reduction of the cross expectation errors when compared to case where they are issued from *iid* realizations (no morphism performed). For instance the variance, noted as $\text{var}(E_N(T))$ of the N -sized sampling mean $E_N(T)$, of a cross function $T(X, Y)$ is given by $N^{-1} \text{var}_N(T^*)$, where T^* is the residual of the best linear fit of T , using the conditional means $E(T|X), E(T|Y)$ as predictors. Asymptotically, $\text{var}_N(T^*) \rightarrow \text{var}(T^*)$ which is the variance of T , conditioned to the knowledge of marginal PDFs, computed at the joint PDF of the population. These conditional variances are exactly those coming from the MinMI solution, allowing for relating MinMI statistics with asymptotic no-replacement finite statistics under fixed marginals. The results are synthesized in the form of two theorems.

Regarding the conversion of expectation errors to ME and MinMI errors, we have used a perturbative approach—a 2nd order Taylor expansion of the ME. This allows for closed analytical formulas to be obtained for MinMI variance and its distribution in a few cases (*e.g.*, Chi-Squared distributions), in what we hereafter call the *analytical approach*. In order to confirm that, expectation errors are generated by surrogates of the governing multivariate Gaussian PDF; then, they are plugged into the Taylor expansion of MinMI and finally statistics (bias, variances, quantiles) are estimated from a large ensemble (*semi-analytical approach*). These statistics are compared with those obtained from a Monte-Carlo experiment where MinMI is computed *ab initio* from the sampling expectations—the *Monte-Carlo approach*. The closeness of results between the Monte-Carlo, the semi-analytical and the analytical approaches is tested using several statistical tests of bivariate non-Gaussianity and RV independency. This exhaustive validation has already been performed for testing analytical formulas of bias, variance, skewness and kurtosis of MI estimation errors [27].

In accordance to the above synthesis, the paper structure starts with this introduction, followed by the formulation of MinMI and their estimators in Section 2. In Section 3 we present the modeling of sample mean errors that will constrain entropy and the effect of morphisms on statistics. Section 4 is devoted to the modeling of errors of MinMI, incremental MinMI and significance tests, followed by a practical case of MI estimation with under-sampled data (Section 5) and the discussion with conclusions in section 6. An appendix with some proofs is also provided.

2. Minimum Mutual Information and Its Estimators

2.1. Imposing Marginal PDFs

Let us formulate the problem of finding the minimum Mutual Information (MinMI) in the simplest framework of bivariate RVs (X, Y) , over the Cartesian product of support sets $S = S_X \otimes S_Y \subseteq \mathbb{R}^2$. The MinMI is constrained by the imposition of marginal PDFs ρ_X, ρ_Y and a set of cross expectations $\{\mathbf{T}_{cr}, \boldsymbol{\theta}_{cr} \equiv E(\mathbf{T}_{cr})\}$, where $\mathbf{T}_{cr}$ is a vector comprising m_{cr} cross X, Y functions and $\boldsymbol{\theta}_{cr}$ is the vector of their expectations. In the space of imposed PDF marginals, the MinMI comes uniquely as a function of $\boldsymbol{\theta}_{cr}$ as $I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) = H_{\rho_X} + H_{\rho_Y} - H_{\rho_{X,Y}}^*(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y)$, where $H_{\rho_X} = E[-\log(\rho_X)]$, $H_{\rho_Y} = E[-\log(\rho_Y)]$ are preset Shannon entropies of X, Y respectively and $H_{\rho_{X,Y}}^*(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y)$ is the ME subjected to joint constraints and marginal PDFs where the ME-PDF is $\rho_{X,Y}^*$. That leads to the equivalence between computations of MinMI and ME [9]. In particular if ρ_X, ρ_Y are copula marginals (uniform PDFs in $[0,1]$), then $H_{\rho_X} = H_{\rho_Y} = 0$ and the MinMI is the copula entropy [24,25]. For instance, for standard Gaussians X, Y and a given correlation $E(T_{cr} \equiv XY) = c_g$, the MinMI is $I(c_g) = -\frac{1}{2} \log(1 - c_g^2)$. Obviously, the more cross constraints are imposed, the larger the MinMI will be.

The general solution is obtained through variational analysis, rather similar to that for the ME [28] but with a continuity of constraints (the marginal PDFs) and a finite set of expectations:

$$I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) = H_{\rho_X} + H_{\rho_Y} - H_{\rho_{X,Y}}^*(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) ; H_{\rho_{X,Y}}^*(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) = L(\boldsymbol{\lambda}_{cr})$$

$$\boldsymbol{\lambda}_{cr} = \arg \min_{\boldsymbol{\eta}_{cr}} \left[L(\boldsymbol{\eta}_{cr}) \equiv 1 + \int_{S_X} \log Z_X(X, \boldsymbol{\eta}_{cr}) \rho_X(X) dx + \int_{S_Y} \log Z_Y(Y, \boldsymbol{\eta}_{cr}) \rho_Y(Y) dy - \boldsymbol{\eta}_{cr}^T \boldsymbol{\theta}_{cr} \right] \quad (1)$$

The MinMI-PDF $\rho_{X,Y}^*(X, Y)$ and the partition functions Z_X, Z_Y are

$$\rho_{X,Y}^*(X, Y) = [Z_X(X, \boldsymbol{\lambda}_{cr}) Z_Y(Y, \boldsymbol{\lambda}_{cr})]^{-1} \exp[-1 + \boldsymbol{\lambda}_{cr}^T \mathbf{T}_{cr}(X, Y)] ;$$

$$Z_X(X, \boldsymbol{\lambda}_{cr}) \equiv \frac{1}{\rho_X(X)} \int_{S_Y} \frac{\exp[-1 + \boldsymbol{\lambda}_{cr}^T \mathbf{T}_{cr}(X, y)]}{Z_Y(y, \boldsymbol{\lambda}_{cr})} dy ; \quad (2)$$

$$Z_Y(Y, \boldsymbol{\lambda}_{cr}) \equiv \frac{1}{\rho_Y(Y)} \int_{S_X} \frac{\exp[-1 + \boldsymbol{\lambda}_{cr}^T \mathbf{T}_{cr}(x, Y)]}{Z_X(x, \boldsymbol{\lambda}_{cr})} dx$$

The superscript T stands for transpose such that $\boldsymbol{\lambda}_{cr}^T \mathbf{T}_{cr}$ is the canonical inner product between vectors $\boldsymbol{\lambda}_{cr}$ and $\mathbf{T}_{cr}$. The proof is given in Appendix 1. Any PDF $\rho_{XY}(X, Y)$ is a MinMI PDF corresponding to the single constraint $\mathbf{T}_{cr}(X, Y) = 1 + \log[\rho_{XY}(X, Y) / [\rho_X(X) \rho_Y(Y)]]$, leading to $\lambda = 1$, $Z_X(X, \boldsymbol{\lambda}) = \rho_X(X)^{-1}$ and $Z_Y(Y, \boldsymbol{\lambda}) = \rho_Y(Y)^{-1}$.

The minimization of $L(\boldsymbol{\eta})$ in (1) calls for the implementation of an iterative strategy as in [11] with successive adjustments of the implicitly linked partition functions.

The present paper deals with small changes of $I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y)$ coming from estimation errors $\Delta \boldsymbol{\theta}_{cr}$ of the cross expectations evaluated from finite samples. For the purpose of inferring the consequent MinMI error statistics (bias, variance, distribution), we will use the second-order Taylor expansion of $I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y)$ in terms of the variation $\Delta \boldsymbol{\theta}_{cr}$:

$$\begin{aligned} \Delta I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) &\equiv I(\boldsymbol{\theta}_{cr} + \Delta \boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) - I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) = -\Delta H_{\rho_{X,Y}^*}(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y) = \\ &= \boldsymbol{\lambda}_{cr}^T \Delta \boldsymbol{\theta}_{cr} + 1/2 (\Delta \boldsymbol{\theta}_{cr}^T) \mathbf{C}_{cr, \rho_X, \rho_Y}^{-1} (\Delta \boldsymbol{\theta}_{cr}) + O(\|\Delta \boldsymbol{\theta}_{cr}\|^3) \end{aligned} \quad (1)$$

where $\mathbf{C}_{cr, \rho_X, \rho_Y}^{-1}$ is the inverse of the covariance matrix of the vector of constraining functions $\mathbf{T}_{cr}$, conditioned to knowledge of marginal PDFs and evaluated at the MinMI-PDF $\rho_{X,Y}^*$ i.e.,

$$\mathbf{C}_{cr, \rho_X, \rho_Y} = E_{\rho_{X,Y}^*}[(\mathbf{T}_{cr}^* \mathbf{T}_{cr}^{*T} | \rho_X, \rho_Y)] = E_{\rho_{X,Y}^*}[(\mathbf{T}_{cr}^* \mathbf{T}_{cr}^{*T} | E(\mathbf{T} | X), E(\mathbf{T} | Y))] \quad (2)$$

where $E_{\rho_{X,Y}^*}$ is the expectation at $\rho_{X,Y}^*$. The perturbation $\mathbf{T}^* = \mathbf{T} - E_{\rho_{X,Y}^*}(\mathbf{T}_{cr} | \rho_X, \rho_Y)$ is the residual with respect to the conditional mean, obtained by methods of variational and functional analysis as the best linear fit

$$E_{\rho_{X,Y}^*}(\mathbf{T}_{cr} | \rho_X, \rho_Y) = \boldsymbol{\theta}_{cr} + \boldsymbol{\alpha}_X [E_{\rho_{X,Y}^*}(\mathbf{T}_{cr} | X) - \boldsymbol{\theta}_{cr}] + \boldsymbol{\alpha}_Y [E_{\rho_{X,Y}^*}(\mathbf{T}_{cr} | Y) - \boldsymbol{\theta}_{cr}] \quad (3)$$

where $\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_Y$ are vectors of coefficients minimizing the mean square deviations to each component of $\mathbf{T}_{cr}$ using the X and Y conditional means of $\mathbf{T}_{cr}$ as predictors. The proof is given in Appendix 1 as part of the proof of Theorem 1 presented in Section 2.2.

2.2. Imposing Marginals through ME Constraints

2.2.1. The Formalism

In PP12 [12], we address the MinMI problem (1,2) by considering that ρ_X, ρ_Y are themselves ME-PDFs forced by a finite set of marginal, independent constraints, $\{\mathbf{T}_{ind} \equiv (\mathbf{T}_X(X), \mathbf{T}_Y(Y)), \boldsymbol{\theta}_{ind} \equiv E(\mathbf{T}_{ind}) \equiv (\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y)\}$. For that purpose we solve the ME problem [29] by imposing the constraints set $\{\mathbf{T}, \boldsymbol{\theta}\} = \{(\mathbf{T}_{ind}, \mathbf{T}_{cr}), (\boldsymbol{\theta}_{ind}, \boldsymbol{\theta}_{cr})\}$, thus leading to a weaker (i.e., smaller) MinMI solution than that obtained with the full imposition of the marginal PDFs. That is given by $I(\boldsymbol{\theta}_{cr}, \boldsymbol{\theta}_{ind}) = H(\boldsymbol{\theta}_{ind}) - H(\boldsymbol{\theta}) \leq I(\boldsymbol{\theta}_{cr}, \rho_X, \rho_Y)$, where $H(\boldsymbol{\theta})$ is the ME issued from the finite set of constraints (marginal and cross) and $H(\boldsymbol{\theta}_{ind}) \equiv H_X + H_Y$ is the ME corresponding uniquely to the marginal constraints [30]. In particular, if the support sets are $S_X = S_Y = [0,1]$ and $\{\mathbf{T}_{ind}, \boldsymbol{\theta}_{ind}\} = \emptyset$ (no constraints on marginals), then the joint PDF of (X, Y) is a copula [24] since their marginal PDFs are uniform in $[0,1]$. The cross part $\mathbf{T}_{cr}$ includes only cross functions, not redundantly expressed as sums of marginal functions in $\mathbf{T}_{ind}$.

In practice one can impose the marginal PDFs from *a priori* RVs $(\hat{X}, \hat{Y})$ (data variables) through ME-morphisms $(X = X(\hat{X}), Y = Y(\hat{Y}))$ (Equation 6 of PP12), (e.g., standard Gaussians), which are monotonically growing smooth homeomorphisms linking data to transformed (X, Y) variables. Then, thanks to the MI invariance $(X = X(\hat{X}), Y = Y(\hat{Y}))$ [2], one can consistently define the MinMI between $(\hat{X}, \hat{Y})$ as that obtained with (X, Y) .

The joint ME-PDF is written in terms of a vector $\boldsymbol{\lambda}$ of Lagrange multipliers [28] as: $\rho_{\mathbf{T}, \boldsymbol{\theta}}^*(X, Y) = Z(\boldsymbol{\lambda}, \mathbf{T})^{-1} \exp[\boldsymbol{\lambda}^T \mathbf{T}(X, Y)]$, where $Z(\boldsymbol{\lambda}, \mathbf{T}) \equiv \iint_S \exp(\boldsymbol{\lambda}^T \mathbf{T}) dx dy$ is the partition function. The ME functional is $H(\boldsymbol{\theta}) = \min_{\boldsymbol{\eta}} (\log Z(\boldsymbol{\eta}, \mathbf{T}) - \boldsymbol{\theta}^T \boldsymbol{\eta}) = \log Z(\boldsymbol{\lambda}, \mathbf{T}) - \boldsymbol{\theta}^T \boldsymbol{\lambda}$, whose input is the vector $\boldsymbol{\theta}$.

The marginal PDFs are supposed to be the ME-PDFs $\rho_{\mathbf{T}_X, \mathbf{0}_X}^*(X); \rho_{\mathbf{T}_Y, \mathbf{0}_Y}^*(Y)$, verifying the marginal X and Y constraints respectively, since variables were built accordingly by ME-morphisms.

As far as more cross constraints are added to $\{\mathbf{T}_{cr}, \mathbf{\theta}_{cr}\}$, the MinMI $I(\mathbf{\theta}_{cr}, \mathbf{\theta}_{ind})$ increases converging to the full MI $I(X, Y)$. Let us formalize that by supposing that the true joint PDF belongs to the ME-family characterized by an information moment superset $\{\mathbf{T}_\infty, \mathbf{\theta}_\infty\} \supseteq \{\mathbf{T}, \mathbf{\theta}\}$.

The true joint PDF is given by $\rho_{\mathbf{T}_\infty, \mathbf{\theta}_\infty}^*$ with Shannon entropy given by the ME $H(\mathbf{\theta}_\infty)$. The encapsulated moment sets obey to $\mathbf{\theta}_{ind} \subseteq \mathbf{\theta} \subseteq \mathbf{\theta}_\infty$. Therefore, thanks to Lemma 1 of PP12, the monotonic property of MEs is obtained: $H(\mathbf{\theta}_{ind}) \geq H(\mathbf{\theta}) \geq H(\mathbf{\theta}_\infty)$. This, according to Theorem 1 of PP12, allows for the decomposition of the MI $I(X, Y)$ into two positive terms, such that:

$$\begin{aligned} I(X, Y) &= H(\mathbf{\theta}_{ind}) - H(\mathbf{\theta}_\infty) = I_{\mathbf{\theta}/\mathbf{\theta}_{ind}}(X, Y) + I_{\mathbf{\theta}_\infty/\mathbf{\theta}}(X, Y) \geq 0 \\ I_{\mathbf{\theta}/\mathbf{\theta}_{ind}} &\equiv H(\mathbf{\theta}_{ind}) - H(\mathbf{\theta}) \geq 0 \quad ; \quad I_{\mathbf{\theta}_\infty/\mathbf{\theta}} \equiv H(\mathbf{\theta}) - H(\mathbf{\theta}_\infty) \geq 0 \end{aligned} \quad (6)$$

The term $I_{\mathbf{\theta}/\mathbf{\theta}_{ind}}$ is the MinMI associated to the finite set of cross moments $\mathbf{\theta}_{cr}$ and the second one is the remaining MI. The decomposition (6) allows us for defining a monotonic sequence of lower MI bounds converging to the total MI. That follows from the sequence of encapsulated moment sets $\{\mathbf{T}_{ind} = \mathbf{T}_0, \mathbf{\theta}_{ind} = \mathbf{\theta}_0\} \subseteq \{\mathbf{T}_j, \mathbf{\theta}_j\} \equiv \{(\mathbf{T}_{ind,j}, \mathbf{T}_{cr,j}), (\mathbf{\theta}_{ind,j}, \mathbf{\theta}_{cr,j})\} \subseteq \{\mathbf{T}_{j+1}, \mathbf{\theta}_{j+1}\} \subseteq \dots \subseteq \{\mathbf{T}_\infty, \mathbf{\theta}_\infty\}$, $j \geq 1$ (e.g. set of monomial bivariate moments of a certain total order j), whose ME-PDF approximates the true ME-PDF in the sense of the Kullback-Leibler divergence (KBD) i.e., $D_{KL}(\rho_{\mathbf{T}_\infty, \mathbf{\theta}_\infty}^* \parallel \rho_{\mathbf{T}_j, \mathbf{\theta}_j}^*) = H(\mathbf{\theta}_j) - H(\mathbf{\theta}_\infty) \xrightarrow{j \rightarrow \infty} 0$ with the MI given by the limit $I(X, Y) = H(\mathbf{\theta}_{ind}) - \lim_{j \rightarrow \infty} [H(\mathbf{\theta}_j)]$. The sets $\{\mathbf{T}_0, \mathbf{\theta}_0\}$ and $\{\mathbf{T}_{ind,j}, \mathbf{\theta}_{ind,j}\}$ are ME-congruent, i.e., their ME-PDF are the same. The j -th set must include enough constraints so as to keep a finite joint ME issued from $\{\mathbf{T}_j, \mathbf{\theta}_j\}$ and guarantee the convergence of the above KBD towards zero. Moreover that also guarantees that marginals of the joint ME-PDF converge to the preset marginal PDFs ρ_X, ρ_Y in the KBD sense. Therefore, the MinMI $I(\mathbf{\theta}_{cr, \infty}, \rho_X, \rho_Y) = I(X, Y) = H(\mathbf{\theta}_{ind}) - \lim_{j \rightarrow \infty} [H(\mathbf{\theta}_j)]$.

The addition of constraints leads to the decrease of ME, raising the useful concept of incremental MinMI next presented. The MI part that is explained by cross terms in the set difference $\mathbf{T}_j / \mathbf{T}_p$ ($j > p \geq 0$), i.e., $\mathbf{T}_p \subseteq \mathbf{T}_j$ is the incremental MinMI:

$$I_{j/p} \equiv H(\mathbf{\theta}_p) - H(\mathbf{\theta}_j) = D_{KL}(\rho_{\mathbf{T}_j, \mathbf{\theta}_j}^* \parallel \rho_{\mathbf{T}_p, \mathbf{\theta}_p}^*) = I_{j/0} - I_{p/0} \geq 0 \quad (7)$$

Estimation errors of $I_{j/p}$ are affected by the vector of moment errors $\Delta \mathbf{\theta}_j$ (from which $\Delta \mathbf{\theta}_p$ is simply a projection). Since we preset marginal PDFs, $\Delta \mathbf{\theta}_j$ is restricted to the cross part i.e., $\Delta \mathbf{\theta}_j = \Delta \mathbf{\theta}_{cr,j} = \mathbf{P}_{cr,j} \Delta \mathbf{\theta}_j$ where $\mathbf{P}_{cr,j}$ is the diagonal projector operator over cross expectations (cr and ind terms are set to 1 and 0 respectively). Looking for error statistics of $I_{j/p}$, we use the second-order Taylor expression of ME:

$$-\Delta H = H(\mathbf{\theta}) - H(\mathbf{\theta} + \Delta \mathbf{\theta}_{cr}) = (\mathbf{P}_{cr} \boldsymbol{\lambda})^T \Delta \mathbf{\theta}_{cr} + (1/2) \Delta \mathbf{\theta}_{cr}^T (\mathbf{P}_{cr} \mathbf{C}_*^{-1} \mathbf{P}_{cr}) \Delta \mathbf{\theta}_{cr} + O(\|\Delta \mathbf{\theta}_{cr}\|^3) \quad (8)$$

where, as usually, λ (with dropped subscripts) is the whole vector of Lagrange multipliers of dimension $\dim(\theta_{cr}) + \dim(\theta_{ind})$ and C_* is the covariance matrix of the function vector T , both valid for the ME-PDF verifying the constraints $E_*(T) = \theta$. We note that $C_* = E_*[T'T^T]$, where the star stands for evaluation over the ME-PDF and prime denotes deviation from the mean θ , i.e., $T' = T - \theta$. Therefore, by using (8), we express the variation of $I_{j/p}$ ($j > p$) due to variations $\Delta\theta_{cr,j}$ as:

$$\begin{aligned} \Delta I_{j/p} &= (v_{j/p})^T (\Delta\theta_{cr,j}) + (1/2) (\Delta\theta_{cr,j})^T A_{j/p} (\Delta\theta_{cr,j}) + O(\|\Delta\theta_{cr,j}\|^3) \\ v_{j/p}^T &\equiv P_{cr,j} \lambda_j - P_{cr,p} \lambda_p \quad ; \quad A_{j/p} \equiv P_{cr,j} (C_{*j}^{-1} - P_{cr,p} (C_{*p})^{-1} P_{cr,p}) P_{cr,j} \end{aligned} \quad (9)$$

where λ_j, C_{*j} and λ_p, C_{*p} are the whole vectors of Lagrange multipliers and the whole covariance matrices, valid for the ME-PDFs of orders j and p respectively. The matrix $A_{j/p}$ is built from the covariance matrices C_{*j} and C_{*p} valid at the ME-PDFs of order j and p respectively.

When the ME-PDFs of order j and p are the same (which is useful for testing if the estimated $I_{j/p}$ from data is significantly different from zero), or $p = 0$ (in which $P_{cr,p} = 0$), then C_{*p} is a sub-matrix of C_{*j} . In that case, $A_{j/p}$ is positive semi-definite (PSD). This comes from the algebraic generic result stating that $A = C^{-1} - PC_p^{-1}P$ is PSD, where C is PSD, P is a diagonal projection matrix, $C_p = PCP$ is the projected C with generalized inverse C_p^{-1} such that $C_p C_p^{-1} = C_p^{-1} C_p = P$. A is singular with $\text{Ker}(A) = \text{Im}(CP)$. However, one can prove that for small deviations among the ME-PDFs of orders j and p , the matrix $A_{j/p}$ is still PSD. For that one can use the same perturbation approach of [26].

2.2.2. A Theorem about the MinMI Covariance Matrix

The matrix $P_{cr} C_{*}^{-1} P_{cr}$ in (8) has inverse in the cross-expectation subspace, i.e. $(P_{cr} C_{*}^{-1} P_{cr})^{-1} (P_{cr} C_{*}^{-1} P_{cr}) = P_{cr}$. Taking the identity as the sum of complementary projector operators $I = P_{cr} + P_{ind}$, both diagonal and self-adjoint, we have

$$\begin{aligned} (P_{cr} C_{*}^{-1} P_{cr})^{-1} &= (P_{cr} C_{*} P_{cr}) - (P_{cr} C_{*} P_{ind})(P_{ind} C_{*} P_{ind})^{-1} (P_{ind} C_{*} P_{cr}) \\ &= E_*[T'_{cr} T'_{cr}{}^T] - E_*[T'_{cr} T'_{ind}{}^T] E_*[T'_{ind} T'_{ind}{}^T]^{-1} E_*[T'_{ind} T'_{cr}{}^T] = E_*[T'_{cr}{}^{ind} T'_{cr}{}^{ind}{}^T] \end{aligned} \quad (10)$$

which is the covariance matrix between the residuals $T'_{cr}{}^{ind}$ of the best linear fit (in the sense of mean squares error) of T'_{cr} using the X and Y functions in T'_{ind} as predictors, i.e., $T'_{cr}{}^{ind} \equiv T'_{cr} - \alpha_{ind,cr}^T T'_{ind}$ where the matrix of coefficients is $\alpha_{ind,cr} = E_*[T'_{ind} T'_{ind}{}^T]^{-1} E_*[T'_{ind} T'_{cr}{}^T]$. The identity (10) is simply an application to the ME covariance matrix of a generic algebraic result on PSD matrices C_* and projection operators $P_{cr}, P_{ind} = I - P_{cr}$.

Therefore, the variances in $(P_{cr} C_{*}^{-1} P_{cr})^{-1}$ are smaller than those in $(P_{cr} C_{*} P_{cr})$. Moreover, the more marginal constraints are imposed (with increasing j), the smaller the variances from $(P_{cr} C_{*}^{-1} P_{cr})^{-1}$ will be, due to the increasing number of predictors and closer will be the full knowledge of the marginal PDFs. Then, asymptotically the residuals $T'_{cr,j}{}^{ind}$ at step j must converge to the residuals

$\mathbf{T}^* = \mathbf{T} - E_{\rho_{X,Y}}^* (\mathbf{T}_{cr} | \rho_X, \rho_Y)$ with respect to the mean (5) entering in the covariance (4) regarding MinMI. Therefore, that leads us to the Theorem:

Theorem 1: Let $\rho_{X,Y}^*$ be the MinMI-PDF issued from $\{\mathbf{T}_{cr}, \boldsymbol{\theta}_{cr}\}, \rho_X, \rho_Y$, being the same as the ME-PDF issued from $\{(\mathbf{T}_{ind}, \mathbf{T}_{cr}), (\boldsymbol{\theta}_{ind}, \boldsymbol{\theta}_{cr})\}$ for some set $\{\mathbf{T}_{ind}, \boldsymbol{\theta}_{ind}\}$. Then we have:

$$\boldsymbol{\lambda}_{cr} = \mathbf{P}_{cr} \boldsymbol{\lambda} \quad ; \quad \mathbf{C}_{cr, \rho_X, \rho_Y} = (\mathbf{P}_{cr} \mathbf{C}_*^{-1} \mathbf{P}_{cr})^{-1} = E_{\rho_{X,Y}}^* [(\mathbf{T}_{cr}^* \mathbf{T}_{cr}^{*T} | E(\mathbf{T} | X), E(\mathbf{T} | Y))] \quad (41)$$

which states that the Lagrange multipliers of the MinMI-PDF are those of the ME-PDF for the cross constraints and the MinMI covariance matrix (4), say that of the residuals of the best fit of the cross constraints using their conditional means as predictors. The proof, as well of (3–5) is added in Appendix 1.

An illustrative example of the Theorem 1 is given for the bivariate Gaussian $\rho_{XY}^*(X, Y) = (2\pi)^{-1} d_g^{1/2} \exp[\frac{-1}{2} d_g (X^2 - 2c_g XY + Y^2)]$ of correlation c_g with $d_g \equiv (1 - c_g^2)^{-1}$. The marginals ρ_X, ρ_Y are standard Gaussians. $\rho_{XY}^*(X, Y)$ is the MinMI-PDF constrained by correlation as well as the ME-PDF constrained by moments of order one and two: $\{\mathbf{T}_{ind} = (X, X^2, Y, Y^2), \boldsymbol{\theta}_{ind} = (0, 1, 0, 1)\}$ and $\{\mathbf{T}_{cr} = (XY), \boldsymbol{\theta}_{cr} = (c_g)\}$. The vector of Lagrange multipliers is $\boldsymbol{\lambda} = [0, \frac{-1}{2} d_g, 0, \frac{-1}{2} d_g, c_g d_g]^T$ while the diagonal covariance matrix and its inverse (lower triangle parts) are:

$$\begin{aligned} \mathbf{C}_* &= [(1, 0, c_g, 0, 0)^T, (*, 2, 0, 2c_g^2, 2c)^T, (**, 1, 0, 0)^T, (***, 2, 2c)^T, (****, c_g^2 + 1)^T] \\ \mathbf{C}_*^{-1} &= [(d_g, 0, -c_g d_g, 0, 0)^T, (*, \frac{1}{2} d_g^2, 0, \frac{1}{2} c_g^2 d_g^2, -c_g d_g^2)^T, (**, d_g, 0, 0)^T, \\ & \quad (***, \frac{1}{2} d_g^2, -c_g d_g^2)^T, (****, (1 + c_g^2) d_g^2)^T] \end{aligned} \quad (15)$$

The redundant upper triangle part is given by stars. The MinMI is $I_g(c_g) = \frac{-1}{2} \log(1 - c_g^2)$ with its derivatives entering in the Taylor development (3) given by $\frac{\partial I_g}{\partial c_g} = c_g d_g = \mathbf{P}_{cr} \boldsymbol{\lambda}$ which is the fifth component of $\boldsymbol{\lambda}$ and $\frac{\partial^2 I_g}{\partial c_g^2} = d_g^2 (1 + c_g^2) = \mathbf{C}_{cr, \rho_X, \rho_Y}^{-1} = (\mathbf{P}_{cr} \mathbf{C}_*^{-1} \mathbf{P}_{cr})$, i.e., the entry at 5th line, 5th column of $\mathbf{C}_*^{-1}$ as guessed from the Theorem 1. By expressing $Y = c_g X + d_g^{-1/2} W_X$ and $X = c_g Y + d_g^{-1/2} W_Y$ with standard Gaussian noises $W_X, W_Y \sim \mathcal{N}(0, 1)$, and $\text{cor}(X, W_X) = \text{cor}(Y, W_Y) = 0$, one easily gets the conditional means $\mathbf{T}_{cr}$ as $E_{\rho_{X,Y}}^*(XY | X) = c_g X^2$; $E_{\rho_{X,Y}}^*(XY | Y) = c_g Y^2$, leading to the best linear fit with mean square error $\mathbf{C}_{cr, \rho_X, \rho_Y} = d_g^{-2} (1 + c_g^2)^{-1}$, confirming the second part of (11).

2.3. Gaussian and Non-Gaussian MI

There is a particular MI decomposition of the type (6,7), already studied in PP12 [12], in which both RVs X and Y are set to standard Gaussians $\mathcal{N}(0, 1)$ over the real support set $S_X = S_Y = \mathbb{R}$ by Gaussian morphism [31]. The isotropic bivariate standard Gaussian is constrained by the moment set $\mathbf{T}_{ind} = \mathbf{T}_0 = (X, X^2, Y, Y^2)^T$ with the expectations vector $\boldsymbol{\theta}_{ind} = \boldsymbol{\theta}_0 = E(\mathbf{T}_0) = (0, 1, 0, 1)^T$. The sequence of MinMIs is obtained by considering the indexed moment set (Equation 14 of PP12 [12], changing the index p there into j here):

$$\mathbf{T}_j \equiv \{X^r Y^s : 1 \leq r + s \leq j, (r, s) \in \mathbb{N}_0^2\}, j \in \mathbb{N} \quad (16)$$

Comprising bivariate polynomials of total order j . Only natural j even numbers provide integrable ME-PDFs over $\mathbb{R}$, thus excluding odd j values from the sequence $\{\mathbf{T}_0, \mathbf{\theta}_0\}, \{\mathbf{T}_2, \mathbf{\theta}_2\}, \{\mathbf{T}_4, \mathbf{\theta}_4\} \dots \{\mathbf{T}_\infty, \mathbf{\theta}_\infty\}$ of set pairs {moments, expectations}. The independent parts of all sets are ME-congruent with $\{\mathbf{T}_0, \mathbf{\theta}_0\}$, i.e., they include high-order univariate moment expectations of the standard Gaussian. The number of independent and cross moments of $\mathbf{T}_j$ (13) is $2j$ and $j(j-1)/2$ respectively (e.g. (4,1), (8,6), (12,15) and (16,28), for $j=2,4,6,8$). Other more efficient basis cross functions could be used as for example orthogonal polynomials. Using the notation of Section 2.2, the maximum entropy limit $H(\mathbf{\theta}_\infty)$ of the sequence limit coincides to the true (X,Y) Shannon entropy. As presented in PP12, we define the positive Gaussian MI I_g , the non-Gaussian MI I_{ng} and the non-Gaussian MI $I_{ng,j}$ of even order j , respectively as:

$$\begin{aligned} I_g &= I_{2/0} = H(\mathbf{\theta}_0) - H(\mathbf{\theta}_2) = -(1/2) \log(1 - c_g^2) \equiv I_g(c_g) ; \\ I_{ng} &= I_{\infty/2} = H(\mathbf{\theta}_2) - H(\mathbf{\theta}_\infty) ; \quad I_{ng,j} = I_{j/p=2} = H(\mathbf{\theta}_2) - H(\mathbf{\theta}_j) \end{aligned} \quad (17)$$

with the MI decomposed as $I(X,Y) = I_g + I_{ng} \geq I_g + I_{ng,j}$. The Gaussian MI depends on the Gaussian correlation c_g , i.e., the Pearson correlation between the Gaussianized variables (X,Y) . The non-Gaussian MI vanishes iff the joint PDF is Gaussian.

2.4. Estimators of the Minimum MI from Data and Their Errors

This section is devoted to the study of estimators (and their errors) of the incremental MI $I_{j/p}$ ($j > p$), (7) between *a priori* RVs $\hat{X}, \hat{Y}$ or, equivalently, between their transformed RVs X, Y .

In practice, the incremental MI $I_{j/p}$, $j > p$ is estimated by a two-step algorithm: first, the computation of expectations; then, the MEs and the partial MIs. The vector of expectations, $\mathbf{\theta}_{N,j}$, is estimated from the N -sized bivariate series $(X_l, Y_l), l=1, \dots, N$, obtained by morphism from the original N iid realizations of the *a-priori* RVs $(\hat{X}_l, \hat{Y}_l), l=1, \dots, N$ (e.g. time-series, spatially distributed data), as the arithmetic average:

$$E_N(\mathbf{T}_j) \equiv \mathbf{\theta}_{N,j} = N^{-1} \sum_{l=1}^N \mathbf{T}_j(X_l, Y_l) = \mathbf{\theta}_j + \Delta \mathbf{\theta}_{N,j} \quad (15)$$

where E_N stands for expectation over the N realizations and the vector of moment estimation errors is $\Delta \mathbf{\theta}_{N,j}$. The first-step error comes from the difference $H(\mathbf{\theta}_{N,j}) - H(\mathbf{\theta}_j)$, due to marginal morphisms and finite bivariate sampling, i.e., the cross combinations of variable realizations. We will see that MI errors depend crucially from moment estimation errors and their statistics.

Secondly, the true ME $H(\mathbf{\theta}_{N,j})$ is estimated as the minimum $\hat{H}(\mathbf{\theta}_{N,j})$ of a functional that is reached by nonlinear minimization techniques (e.g., gradient-descent), taking as inputs $\mathbf{\theta}_{N,j}$ and a set of calibrated parameters. The second-step error comes from the difference $\hat{H} - H \equiv \delta H$. The estimator of $I_{j/p}$ along with its error, decomposed into the first-step ($\Delta I_{N,j/p,\theta}$) and second-step ($\Delta I_{N,j/p,H}$) contributions, is written as

$$\begin{aligned}
I_{N,j/p} &\equiv \hat{H}(\boldsymbol{\theta}_{N,p}) - \hat{H}(\boldsymbol{\theta}_{N,j}) = I_{j/p} + \Delta I_{N,j/p} \quad ; \quad \Delta I_{N,j/p} = \Delta I_{N,j/p,\boldsymbol{\theta}} + \Delta I_{N,j/p,H} \\
\Delta I_{N,j/p,\boldsymbol{\theta}} &\equiv [H(\boldsymbol{\theta}_j) - H(\boldsymbol{\theta}_{N,j})] - [H(\boldsymbol{\theta}_p) - H(\boldsymbol{\theta}_{N,p})] \equiv -\Delta H_{N,j} + \Delta H_{N,p} \\
\Delta I_{N,j/p,H} &\equiv [\hat{H}(\boldsymbol{\theta}_{N,p}) - H(\boldsymbol{\theta}_{N,p})] - [\hat{H}(\boldsymbol{\theta}_{N,j}) - H(\boldsymbol{\theta}_{N,j})] \equiv (\delta H)_{N,p} - (\delta H)_{N,j}
\end{aligned} \tag{16}$$

where $\Delta I_{N,j/p,\boldsymbol{\theta}}$ is the difference between entropy anomalies ΔH due to input errors. The second-step error comes from the numerical implementation and round-off errors of the entropy functional due to: (a) a coarse graining representation of the continuous PDF; (b) the numerical approximation of the ME functional and its gradient; (c) the stopping criteria of the iterative gradient-descent technique. In this article we will neglect the effect of the second-step error, thus approximating the MinMI error by $\Delta I_{N,j/p} \approx \Delta I_{N,j/p,\boldsymbol{\theta}}$ depending uniquely on the sampling error of the cross expectations $\Delta \boldsymbol{\theta}_{cr} = \Delta \boldsymbol{\theta}_{N,cr,j}$.

3. Errors of the Expectation's Estimators

3.1. Generic Properties

The distribution of the MinMI error and its statistics (bias, variance, quantiles) depends on the distribution of the vector of error moments $\Delta \boldsymbol{\theta}_{N,cr,j}$ entering in (9). Here, we present a generic statistical modeling of those errors giving the emphasis in the influence of variable morphisms and bivariate sampling.

Let us assume the reasonable hypothesis that the discrete estimator $\boldsymbol{\theta}_{N,j}$ (15) is a consistent estimator of the mean $\boldsymbol{\theta}_j$, i.e., the error $\Delta \boldsymbol{\theta}_{N,j} \rightarrow \mathbf{0}$, $N \rightarrow \infty$ in probability, with both the bias and covariance matrix converging to zero as data size grows:

$$\mathbf{b}_{\Delta \boldsymbol{\theta}_{N,j}} \equiv E(\Delta \boldsymbol{\theta}_{N,j}) \xrightarrow{N \rightarrow \infty} \mathbf{0} \quad ; \quad \mathbf{M}_{\Delta \boldsymbol{\theta}_{N,j}} \equiv E[(\Delta \boldsymbol{\theta}_{N,j})'(\Delta \boldsymbol{\theta}_{N,j})^T] \xrightarrow{N \rightarrow \infty} \mathbf{0} \quad ; \quad \Delta \boldsymbol{\theta}_{N,j}' = \Delta \boldsymbol{\theta}_{N,j} - \mathbf{b}_{\Delta \boldsymbol{\theta}_{N,j}} \tag{17}$$

where the prime stands for perturbation with respect to the mean. The exact form of the components of $\mathbf{b}_{\Delta \boldsymbol{\theta}_{N,j}}$ and $\mathbf{M}_{\Delta \boldsymbol{\theta}_{N,j}}$ is rather difficult to establish as a consequence of imposing marginal distributions thus reducing the randomness to the covariate sampling. Estimator variances are scaled as $O(1/N)$, though smaller than in the case of N iid outcomes. Moreover, we assume that the convergence rate is higher (faster convergence) for the squared bias than for variances, which is supported in a few examples in next section.

3.2. The Effects of Morphisms and Bivariate Sampling

Let us start with the effect of morphisms transforming original variables $(\hat{X}, \hat{Y})$ into their transformed (X, Y) . That depends on the rank of variables within the available sample. Without loss of generality, let us sort $\hat{X}$ by ascending order in the sample, i.e., the l -th value equaling the ordered l -th value $\hat{X}_l = \hat{X}_{(l)}$, $l=1, \dots, N$. The bivariate l -th realization is $(\hat{X}_l, \hat{Y}_l = \hat{Y}_{(l'(l))})$, where $l'(l): \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ is the random bivariate rank permutation depending upon the particular sample (e.g. the first of $\hat{X}$ coming with the third of $\hat{Y}$, then $l'(l=1)=3$ and so on). In particular $l'(l)=l$ when correlation equals one. The inverse of the function $l'(l)$ is written $l(l')$. The probability p-values

of $\hat{X}_{(l)}, \hat{Y}_{(l')}$ i.e., their marginal cumulated probability functions (CDFs) are respectively $p_{X,l}, p_{Y,l'}$, growing as function of l, l' . Those p-values can only be inferred from the sample or prescribed from *a-priori* hypotheses. The sorted transformed RVs given by ME-morphisms are:

$$X_{(l)} = \Phi_{ME,X}^{-1}(p_{X,l}); Y_{(l')} = \Phi_{ME,Y}^{-1}(p_{Y,l'}); l, l' = 1, \dots, N \quad (18)$$

where $\Phi_{ME,X}, \Phi_{ME,Y}$ are the ME prescribed CDFs (e.g. CDFs of Gaussians) of X and Y respectively. Then the morphisms relies upon invertible transformations $\hat{X}_{(l)} \rightarrow X_{(l)}; \hat{Y}_{(l')} \rightarrow Y_{(l')}$. The bivariate transformed realizations $(X_l, Y_l = Y_{(l'(l))})$, $l = 1, \dots, N$ are then used to compute expectations (Equation 15). Since the exact marginal distributions are not known, their cumulated probabilities must be prescribed, for example with regular steps $\Delta p_{X,l} = \Delta p_{Y,l} = 1/N$ in which $p_{X,l}, p_{Y,l} = l/(N+1)$, $l = 1, \dots, N$.

In order to obtain moments of $\Delta\theta_{N,j}$ we need rewriting it in a convenient form:

$$\begin{aligned} \Delta\theta_{N,j} &\equiv \theta_{N,j} - \theta_j = \\ &= \sum_{l,l'=1}^N \mathbf{T}_j(\Phi_{ME,X}^{-1}(p_{X,l}), \Phi_{ME,Y}^{-1}(p_{Y,l'})) N^{-1} \delta_{l'(l),l'} - \int_0^1 \int_0^1 \mathbf{T}_j(\Phi_{ME,X}^{-1}(u), \Phi_{ME,Y}^{-1}(v)) c[u,v] du dv \\ &\approx \sum_{l,l'=1}^N \mathbf{T}_j(X_{(l)}, Y_{(l')}) \left[\frac{N^{-1} \delta_{l'(l),l'}}{\Delta p_{X,l} \Delta p_{Y,l'}} - c[p_{X,l}, p_{Y,l'}] \right] \Delta p_{X,l} \Delta p_{Y,l'} \end{aligned} \quad (19)$$

where $\delta_{l'(l),l'} = \delta_{l(l'),l}, \forall l, l' \in \{1, \dots, N\}$ is the Kronecker delta, $u = \int_{-\infty}^X \rho_{TX,0X}^*(t) dt$; $v = \int_{-\infty}^Y \rho_{TY,0Y}^*(t) dt$ are the marginal cumulated probabilities, corresponding respectively to probabilities $p_{X,l}$ and $p_{Y,l'}$ in the sum (19) and $c[u,v]$ is the copula function [23] (ratio between the joint PDF and the product of marginal PDFs). By looking at (19), one sees that $N^{-1} \delta_{l'(l),l'} / (\Delta p_{X,l} \Delta p_{Y,l'})$ is an estimator of the copula $c[p_{X,l}, p_{Y,l'}]$. In particular, if X, Y are independent, then l and $l'(l)$ are independent, $c[p_{X,l}, p_{Y,l'}] = 1$ and $E(\delta_{l'(l),l'} | l, l') = N^{-1}$ i.e. there is an average equipartition of the bivariate ranks.

Equation (19) shows that moments of $\Delta\theta_{N,j}$ depend on statistics of the error of the copula estimator, which can be very tricky due to the imposition of marginal PDFs by morphisms, presenting unusual effects with respect to classical results from samples of *iid* realizations [32].

For that, let us denote the random perturbation $\eta_{l,l'} \equiv \delta_{l'(l),l'} - E[\delta_{l'(l),l'}] = \eta_{l',l}, \forall l, l'$, then $E[\eta_{l,l'}] = 0$, also satisfying to the constraints $\sum_{l=1}^N \delta_{l'(l),l'} = \sum_{l'=1}^N \delta_{l(l'),l} = 1$ or $\sum_{l=1}^N \eta_{l,l'} = \sum_{l'=1}^N \eta_{l,l'} = 0$ as a consequence of the fact that $l'(l)$ and $l(l')$ are index permutations of N values. Therefore, taking into account those constraints, $\Delta\theta_{N,j}$ can be written in different forms in terms of perturbations:

$$\begin{aligned} \Delta\theta_{N,j}' &= \sum_{l,l'=1}^N \mathbf{T}_{j,l,l'} N^{-1} \eta_{l,l'} = \sum_{l,l'=1}^N \mathbf{T}_{j,l,l'}' N^{-1} \delta_{l'(l),l'} = \sum_{l,l'=1}^N \mathbf{T}_{j,l,l'}'^X N^{-1} \delta_{l'(l),l'} = \\ &= \sum_{l,l'=1}^N \mathbf{T}_{j,l,l'}'^Y N^{-1} \delta_{l'(l),l'} = \sum_{l=1}^N \mathbf{T}_{j,l,l'(l)}' N^{-1} = \sum_{l=1}^N \mathbf{T}_{j,l,l'(l)}'^X N^{-1} = \sum_{l=1}^N \mathbf{T}_{j,l,l'(l)}'^Y N^{-1} \end{aligned} \quad (20)$$

where $\mathbf{T}_{j,l,l'} \equiv \mathbf{T}_j(X_{(l)}, Y_{(l')})$ and its perturbation with respect to the global mean is $\mathbf{T}_{j,l,l'}' \equiv \mathbf{T}_{j,l,l'} - E(\theta_{N,j})$. The perturbation with respect to X -conditional mean is $\mathbf{T}_{j,l,l'}'^X \equiv \mathbf{T}_{j,l,l'} - E(T_j | X = X_{(l)})$ where $E(T_j | X = X_{(l)}) = \sum_{l'=1}^N \mathbf{T}_j E[\delta_{l'(l),l'}]$. A similar definition is written for the Y -perturbation $\mathbf{T}_{j,l,l'}'^Y \equiv \mathbf{T}_{j,l,l'} - E(T_j | Y = Y_{(l')})$.

The estimator (15) of independent constraints (components of $\mathbf{T}_j$ uniquely dependent on X or Y) have a bias but vanishing variances (null components of $\Delta\mathbf{\theta}_{N,j}$), since perturbations $\mathbf{T}_j^{X}$ or $\mathbf{T}_j^{Y}$ vanish because the local values of $\mathbf{T}_j$ coincide to one of the (X or Y)-conditional means. That bias reduces to a numerical integration error. For example for X -depending functions expectations, the error reduces to bias $\Delta\mathbf{\theta}_{X,N,j} = \sum_{l=1}^N \mathbf{T}_{X,j}(X_{(l)}) N^{-1} - \int_0^1 \mathbf{T}_{X,j}(\Phi_{ME,X}^{-1}(u)) du$, of order $O(N^{-2})$ as given by the trapezoidal integration rule for bounded $\mathbf{T}_{X,j}$ functions. The estimators of cross expectations have bias and non-vanishing variances.

Now, our goal is to get the estimation of the covariance matrix $\mathbf{M}_{\Delta\mathbf{\theta}_{N,j}}$ (17). As a consequence of the non-replacement of quantiles or rankins, the deviations $\mathbf{T}_{j,l_1,l'(l_1)}$ and $\mathbf{T}_{j,l_2,l'(l_2)}$ in (20) are not necessarily independent for $l_1 \neq l_2$, which will not occur if different realizations would be independent, leading to $\text{var}(\mathbf{\theta}_{N,j}) = N^{-1} \text{var}(\mathbf{T}_j)$. The statistics without replacement generally lead to a deflation of estimator variances as compared to those satisfying the hypothesis of independence of realizations [33] or, in other words, $\text{var}(\mathbf{\theta}_{N,j}) \leq N^{-1} \text{var}(\mathbf{T}_j)$. Therefore, in order to get a N^{-1} -scaled expression for $\text{var}(\mathbf{\theta}_{N,j})$, we will consider another type of deviations of $\mathbf{T}_j$ consistent with (20).

We propose new deviations, denoted by $\mathbf{T}_j^{lms}$, that are given by the linear combination both of the global deviation $\mathbf{T}_j$ and of the marginal deviations $\mathbf{T}_j^{X}, \mathbf{T}_j^{Y}$ with the respective coefficients summing 1 and having the least mean square (lms). Those deviations are consistently given by:

$$\begin{aligned} \mathbf{T}_j^{lms} &= (1 - \alpha_X - \alpha_Y) \mathbf{T}_j + \alpha_X \mathbf{T}_j^{X} + \alpha_Y \mathbf{T}_j^{Y} = \\ &\mathbf{T}_j - \alpha_X [E(\mathbf{T}_j | X) - E(\mathbf{\theta}_{N,j})] - \alpha_Y [E(\mathbf{T}_j | Y) - E(\mathbf{\theta}_{N,j})] \end{aligned} \quad (21)$$

which are the residuals of the best linear fit of $\mathbf{T}_j$ using the conditional means $E(\mathbf{T}_j | X)$ and $E(\mathbf{T}_j | Y)$ as predictors and where the coefficients are those of the linear regression:

$$\begin{bmatrix} \alpha_X \\ \alpha_Y \end{bmatrix} = \begin{bmatrix} \text{var}[E(\mathbf{T}_j | X)] & \text{cov}[E(\mathbf{T}_j | X), E(\mathbf{T}_j | Y)] \\ \text{cov}[E(\mathbf{T}_j | X), E(\mathbf{T}_j | Y)] & \text{var}[E(\mathbf{T}_j | Y)] \end{bmatrix}^{-1} \begin{bmatrix} \text{cov}[E(\mathbf{T}_j | X), \mathbf{T}_j] \\ \text{cov}[E(\mathbf{T}_j | Y), \mathbf{T}_j] \end{bmatrix} \quad (22)$$

Those deviations take into account the maximum implicit knowledge of marginal PDFs through their conditional means. Now we will use them for expressing the error moments.

The expression of the error covariances in $\mathbf{M}_{\Delta\mathbf{\theta}_{N,j}}$ relies upon the expansion (20) with perturbations written as function of mean values of products of deltas $\delta_{l'(l),l'}$. These means depend on the true copula and are written as:

$$E(\delta_{l'(l_1),l_1'} \delta_{l'(l_2),l_2'}) = \begin{cases} 0, & \text{if } [l_1 = l_2, l_1' \neq l_2' \text{ or } l_1' = l_2', l_1 \neq l_2] \\ E(\delta_{l'(l_1),l_1'}), N^{-1} (*) & \text{if } [l_1 = l_2, l_1' = l_2'] \\ N^{-1}(N-1)^{-1} (*) & \text{if } [l_1 \neq l_2, l_1' \neq l_2'] \end{cases} \quad (23)$$

where we have considered the fact that $l'(l)$ and its inverse $l(l')$ are permutations of ranks (no duplication allowed). The values indicated with asterisk in (23) correspond to X, Y independent ($l'(l)$ independent of l). Those moments are difficult to obtain in practice unless variables are independent or

the bivariate PDF is known *a priori*. From these moments, a large ensemble of N -sized surrogate samples is generated from which empirical estimator covariances are computed.

Then, by plugging (23) into the generic (α -th row, β -th column) of $\mathbf{M}_{\Delta\theta_{N,j}}$, and denoting the α -th and β -th components of $\mathbf{T}_j$ by $\mathbf{T}_{j,\alpha}$ and $\mathbf{T}_{j,\beta}$ with estimation errors $\Delta\theta_{N,j,\alpha}, \Delta\theta_{N,j,\beta}$, we get

$$(\mathbf{M}_{\Delta\theta_{N,j}})_{\alpha,\beta} = E(\Delta\theta_{N,j,\alpha} \Delta\theta_{N,j,\beta}) = \sum_{l_1, l_1', l_2, l_2'} \left[\mathbf{T}_{j,\alpha}'(X_{(l_1)}, Y_{(l_1)}) \mathbf{T}_{j,\beta}'(X_{(l_2)}, Y_{(l_2)}) \right] N^{-2} E(\delta_{l'(l_1), l_1'} \delta_{l'(l_2), l_2'}) =$$

$$= N^{-1} E(E_N(\mathbf{T}_{j,\alpha}' \mathbf{T}_{j,\beta}')) + N^{-2} \sum_{l_1 \neq l_2} E \left[\mathbf{T}_{j,\alpha}'(X_{(l_1)}, Y_{(l_1)}) \mathbf{T}_{j,\beta}'(X_{(l_2)}, Y_{(l_2)}) \right] \quad (24)$$

The first term of the *rhs* of (24) is given by $N^{-1} E[\text{cov}_N(\mathbf{T}_{j,\alpha}, \mathbf{T}_{j,\beta})]$ i.e. $1/N$ times the expectation of the covariance among N realizations. That term converges asymptotically to $N^{-1} \text{cov}(\mathbf{T}_{j,\alpha}, \mathbf{T}_{j,\beta})$, i.e., the estimator's covariance in the hypothesis of N iid realizations. However, when marginals are imposed or the morphism of variables is performed, that hypothesis no longer holds because the covariance estimator is a statistic without replacement [33], since quantiles of X and Y are not repeated in the sample. Therefore, the additional term of (24) reduces the estimator's variances with respect to the case of iid trials.

Looking for a correct representation of the cross estimator's variances when marginals are imposed, we represent the $\mathbf{T}_j$ perturbations by $\mathbf{T}_j^{lms}$ (21) (residuals of the best linear regression). There, we will benefit from a generic property of *lse* (least squares error) regression residuals which is the fact that they are uncorrelated with the predictors (here the conditional means of $E(\mathbf{T}_j | X), E(\mathbf{T}_j | Y)$). This means that $\mathbf{T}_j^{lms}$ is represented in terms of noises which are uncorrelated, both with X and Y . Consequently, different realizations of $\mathbf{T}_j^{lms}$ are uncorrelated, which will simplify the expression of the covariance matrix. Therefore, using those *lms* perturbations, the generic matrix entry $(\mathbf{M}_{\Delta\theta_{N,j}})_{\alpha,\beta}$ (24) is rewritten as

$$(\mathbf{M}_{\Delta\theta_{N,j}})_{\alpha,\beta} = N^{-2} \sum_{l_1} \left[E[\mathbf{T}_{j,\alpha}^{lms}(X_{(l_1)}, Y_{(l_1)}) \mathbf{T}_{j,\beta}^{lms}(X_{(l_1)}, Y_{(l_1)})] \right] +$$

$$N^{-2} \sum_{l_1, l_2 \neq l_1} \left[E[\mathbf{T}_{j,\alpha}^{lms}(X_{(l_1)}, Y_{(l_1)}) \mathbf{T}_{j,\beta}^{lms}(X_{(l_2)}, Y_{(l_2)})] \right] = N^{-1} E(E_N(\mathbf{T}_{j,\alpha}^{lms} \mathbf{T}_{j,\beta}^{lms})) + O(N^{-2}) \quad (25)$$

The N^{-1} -scaled term of (25) converges asymptotically (as $N \rightarrow \infty$) to $N^{-1} E(\mathbf{T}_{j,\alpha}^{lms} \mathbf{T}_{j,\beta}^{lms})$, i.e., $1/N$ times the covariances between residuals of the linear regression relying upon conditional variances. This let us to formulate the Theorem:

Theorem 2: Let us suppose imposed X and Y marginal PDFs by variable morphisms. Then, the covariance between the N -sized based estimators $\theta_{N,\alpha}$ and $\theta_{N,\beta}$ of the means of cross functions of $T_\alpha(X, Y)$ and $T_\beta(X, Y)$ is given by

$$\text{cov}(\theta_{N,\alpha}, \theta_{N,\beta}) = N^{-1} E(E_N(\mathbf{T}_\alpha^{lms} \mathbf{T}_\beta^{lms})) \xrightarrow{N \rightarrow \infty} N^{-1} E(\mathbf{T}_\alpha^{lms} \mathbf{T}_\beta^{lms}) \quad (26)$$

where $\mathbf{T}_\alpha^{lms} = \mathbf{T}_\alpha - \alpha_X [E(\mathbf{T}_\alpha | X) - \theta_\alpha] - \alpha_Y [E(\mathbf{T}_\alpha | Y) - \theta_\alpha]$ is the residual of the best linear fit taking conditional means as predictors, and α_X, α_Y are the corresponding coefficients (idem for $\mathbf{T}_\beta^{lms}$). The expectation is computed with the true PDF of the population. The proof was given before in the text.

An immediate corollary of this Theorem applies in the case data are governed by a certain MinMI-PDF issued from $\{\mathbf{T}_{cr}, \boldsymbol{\theta}_{cr}\}, \rho_X, \rho_Y$. In that conditions T_α and T_β are themselves cross functions from the constraining set $\mathbf{T}_{cr}$ and $\text{cov}(\theta_{N,\alpha}, \theta_{N,\beta})$ are entries of $\mathbf{M}_{\Delta\theta_N}$ (17). Then, if the true joint PDF is the MinMI-PDF issued from $\{\mathbf{T}_{cr}, \boldsymbol{\theta}_{cr}\}, \rho_X, \rho_Y$, we get:

$$\mathbf{P}_{cr} \mathbf{M}_{\Delta\theta_N} \mathbf{P}_{cr} = N^{-1} \mathbf{C}_{cr, \rho_X, \rho_Y} \quad (27)$$

where we use the covariance matrix introduced in (4). Under those conditions one has the identity for the matricial product $(\mathbf{P}_{cr} \mathbf{M}_{\Delta\theta_N} \mathbf{P}_{cr}) \mathbf{C}_{cr, \rho_X, \rho_Y}^{-1} = N^{-1} \mathbf{P}_{cr}$, which will be crucial for the evaluation of asymptotic MinMI estimation bias.

3.3. Errors of the Estimators of Polynomial Moments under Gaussian Distributions

In this section we assess the bias, the covariance of estimators and its expression (25) when constraints are bivariate monomials (13) and Gaussian morphisms are performed as described in Section 2.3. For the purpose of discussing statistical tests of non-Gaussianity presented in a next section, we will restrict our study by considering the case of N -sized samples of *iid* realizations of independent variables $\hat{X}, \hat{Y}$ (taken without loss of generality standard Gaussians). There, an empiric Monte-Carlo strategy is used by taking the standard Gaussian morphisms X, Y of the N outcomes, from which one estimates the expectation of a vector of generic functions $\mathbf{T}(X, Y) = X^r Y^s, r, s \in \mathbb{N}_0$ (13). The bias is $\mathbf{b} = E(E_N(\mathbf{T})) - E(\mathbf{T}) = \mu_{N,r} \mu_{N,s} - \mu_r \mu_s$, which is determined by the fixed Gaussian centered moments $\mu_r \equiv E(X^r)$ and $\mu_{N,r} \equiv E_N(X^r), r \in \mathbb{N}_0$. The sample is centered and standardized such that $\mu_{N,1} = 0; \mu_{N,2} = 1$. The variance $\text{var}(E_N(\mathbf{T}))$ of $E_N(\mathbf{T})$ can be rigorously computed from the quadruple sum (25) using the N quantiles from the standard Gaussian and the delta expectations (23) for the case of X, Y independent from each other. However, the computation of that sum is very time-consuming for high N values. For that reason, we approximate it by a Monte-Carlo mean obtained with $N_{rea} = 5000$ independent realizations of the N -sized samples. The finite and asymptotic values of $N^{-1} E(\text{var}_N(\mathbf{T}))$, valid for the case of N *iid* trials, are given by:

$$N^{-1} E(\text{var}_N(\mathbf{T})) = N^{-1} \left(\mu_{N,2r} \mu_{N,2s} - (\mu_{N,r} \mu_{N,s})^2 \right) \xrightarrow{N \rightarrow \infty} N^{-1} \text{var}(\mathbf{T}) = N^{-1} \left(\mu_{2r} \mu_{2s} - (\mu_r \mu_s)^2 \right) \quad (28)$$

whereas those (smaller than those of (28)) obtained from least mean squares (25) are:

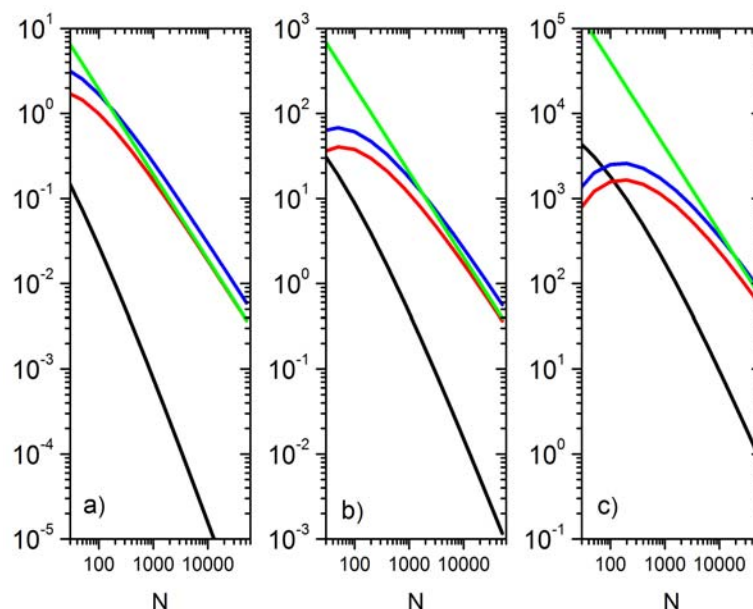
$$\begin{aligned} \text{var}(E_N(\mathbf{T})) &\approx N^{-1} E(\text{var}_N(\mathbf{T} | lms)) = N^{-1} \text{var}_N(\mathbf{T} | lms) = \\ &= N^{-1} \left(\mu_{N,2r} \mu_{N,2s} - \mu_{N,2r} (\mu_{N,s})^2 - \mu_{N,2s} (\mu_{N,r})^2 + (\mu_{N,s} \mu_{N,r})^2 \right) \\ &\xrightarrow{N \rightarrow \infty} N^{-1} \text{var}(\mathbf{T} | lms) = N^{-1} \left(\mu_{2r} \mu_{2s} - \mu_{2r} (\mu_s)^2 - \mu_{2s} (\mu_r)^2 + (\mu_s \mu_r)^2 \right) \end{aligned} \quad (29)$$

Figure 1 compares the variance $\text{var}(E_N(\mathbf{T}))$ with the squared bias $\|\mathbf{b}\|^2$ of the estimator, both relevant in the bias of the MinMI estimation. In the same figure, one compares the empirical variance $\text{var}(E_N(\mathbf{T}))$, with its approximation $N^{-1} \text{var}(\mathbf{T} | lms)$ and with the variance for the case of *iid* trials: $N^{-1} \text{var}(\mathbf{T})$. We use $\mathbf{T} = X^4 Y^2, X^6 Y^2, X^8 Y^2$, respectively in panels a), b), c), sorted by growing total

variance $\text{var}(\mathbf{T})$, specially concentrated at the distribution queues. In all figures, $N=25 \cdot 2^k, k=0, \dots, 11$. We have verified that the empirical variance $\text{var}(E_N(\mathbf{T}))$ agrees very well to the theoretical value $N^{-1} \text{var}_N(\mathbf{T} | lms)$ for all N s. (not shown).

At this point, some generic conclusions can be drawn. The estimator's variance $\text{var}(E_N(\mathbf{T}))$ grows with $\text{var}(\mathbf{T})$ dominating over the squared bias, except for small N values and higher values of $\text{var}(\mathbf{T})$. This will lead us to neglect the bias of covariance estimator's in the MinMI asymptotic statistics.

Figure 1. Squared empirical bias: $\|\mathbf{b}\|^2$ (black lines) of N -based $\mathbf{T}$ - expectations as function of N , empirical variances: $\text{var}(E_N(\mathbf{T}))$ (red lines), approximated variances: $N^{-1} \text{var}(\mathbf{T} | lms)$ (blue lines) and variance for the case of N iid trials: $N^{-1} \text{var}(\mathbf{T})$ (green lines). $\mathbf{T}$ stands for different bivariate monomials: X^4Y^2 (a), X^6Y^2 (b) and X^8Y^2 (c).

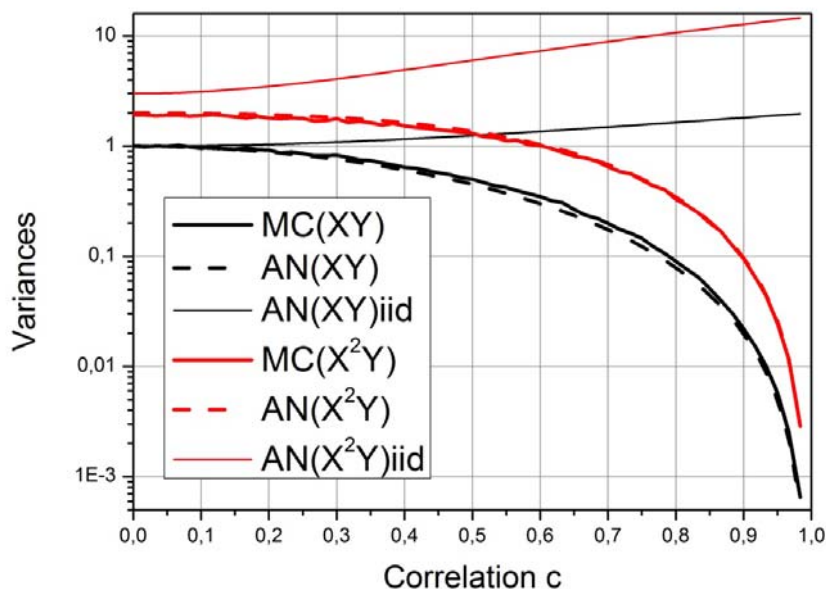


From Figure 1, we also note that the variance reduction coming from morphisms of variables, tends to decrease for higher N values, where the effect of sampling prevails with a N^{-1} scaling on the estimator variance where it is closely approximated by the asymptotic lms variance $N^{-1} \text{var}(\mathbf{T} | lms)$. That can lead to a slight increase of $\text{var}(E_N(\mathbf{T}))$ for small N s, followed by a decrease (e.g., X^6Y^2), due to the effect that $\text{var}_N(\mathbf{T} | lms)$ is small for lower values of N .

Moreover, thanks to the Central Limit Theorem (CLT), the distribution of estimator errors tends towards Gaussianity with increasing N , with a slower convergence rate for higher $\mathbf{T}$ variances. However, the Gaussian PDF limit has an infinite support which must be truncated since the estimated moments $E_N(\mathbf{T})$ must be within a kind of polytope with edges determined by Schwartz-like inequalities as shown by PP12 [12] (e.g., $|E_N(XY)| \leq 1$ and $|E(X^2Y)| / [2(1 - c_g^2)]^{1/2} \leq 1$), working as bounds for nonlinear correlations. Since estimators have bounds, the estimation errors do so as well. This can be solved by using the Fisher Z-transform $\text{arctanh}(c)$ of a generic linear or nonlinear correlation c and projecting it over the real support (not done here).

Now we illustrate in Figure 2, the Theorem 2 under different values of correlation $c_g \in [0,1]$. We consider the variables X, Y with a joint Gaussian PDF of correlation $c_g \in [0,1]$ with marginal standard Gaussians. In Figure 2 we compare the empirical Monte-Carlo value of $N \text{var}(E_N(T))$ (MC in the Figure), within an ensemble of 5000 N -sized samples with the theoretical one $\text{var}(T|lms)$ (case where morphism is performed, AN in the Figure) and $\text{var}(T)$ (case of *iid* realizations, ANiid in the Figure). We have used a sample of $N=200$, which is supposed to be near the beginning of the asymptotic regime and two cross functions: $T(X,Y)=XY$ and $T(X,Y)=X^2Y$. The aforementioned variances are $\text{var}(XY|lms) = (1-c_g^2)/(1+c_g^2)$; $\text{var}(XY) = c_g^2 + 1$ while $\text{var}(X^2Y) = 12c_g^2 + 3$ and $\text{var}(X^2Y|lms)$ is the mean squared residual of the best linear fit using the predictors $E(X^2Y|X) = c_g X^3$ and $E(X^2Y|Y) = c_g^2 Y^3 + (1-c_g^2)Y$. For both functions, a very good agreement is verified between Monte-Carlo values and the theoretical ones within 1–5% relative error. A generic result of Figure 2 is the fact that, under the fixation (presetting) of marginals, the sampling variability of cross estimators falls to zero as far the absolute value of correlation tends to one.

Figure 2. N times Monte-Carlo variances: $N \text{var}(E_N(\mathbf{T}))$ (thick solid lines) and its theoretical analytical value $\text{var}(\mathbf{T} | lms)$ (thick dashed lines), both under imposed marginals (morphisms) and analytical value of $N \text{var}(E_N(\mathbf{T})) = \text{var}(\mathbf{T})$ for *iid* data (thin solid lines). $\mathbf{T}$ means different bivariate monomials: XY (black curves), X^2Y (red curves). $N = 200$.



3.4. Statistical Modeling of Moment Estimation Errors

The above qualitative results gave empirical support to Theorem 2 about the covariance of estimation errors and the neglecting of estimation biases. Therefore, the part of matrix $\mathbf{M}_{\Delta\theta_{N,j}}$ (17) regarding cross components is modeled as:

$$\mathbf{M}_{\Delta\theta_{N,cr,j}} \approx N^{-1} E(E_N(\mathbf{T}_{cr,j}^{lms} \mathbf{T}_{cr,j}^{lms})) \equiv N^{-1} \mathbf{C}_{N,cr,j|lms} \quad (30)$$

with the approximation being valid within terms $o(N^{-1})$. In practice, the matrix $E(\mathbf{T}_{cr,j}^{ilms} \mathbf{T}_{cr,j}^{ilms})$ requires the estimation of conditional means for each value of X and Y .

Now, we will formulate the distribution of moment's estimation errors in the asymptotic regime of high enough N . Then, thanks to the multivariate Central Limit Theorem [34] one can suppose that the unbiased estimation error vector follows a multivariate Gaussian distribution, which is written as

$$\Delta \boldsymbol{\theta}_{N,cr,j} \approx (\mathbf{M}_{\Delta \boldsymbol{\theta}_{N,cr,j}})^{1/2} \mathbf{U}_j \approx N^{-1/2} (\mathbf{C}_{N,cr,j|lms})^{1/2} \mathbf{U}_j ; \quad \mathbf{U}_j \sim \mathcal{N}(\bar{\mathbf{0}}_{cr,j}, \mathbf{P}_{cr,j}) \quad (31)$$

where $(\mathbf{C}_{N,cr,j|lms})^{1/2}$ is the square root matrix of $\mathbf{C}_{N,cr,j|lms}$ and $\mathbf{U}_j$ is a multivariate standard normal RV of dimension equal to $\dim(\boldsymbol{\theta}_{cr,j})$ with zero mean $\bar{\mathbf{0}}_{cr,j}$ and covariance matrix $\mathbf{P}_{cr,j}$.

4. Modeling of MinMI Estimation Errors, Their Bias, Variance and Distribution

Taking into account the Gaussian approximations (31) for estimation errors, their neglected bias, the N^{-1} scaled covariance (30), and the second-order Taylor development of MinMI (9), one can determine approximated bias, variance and distribution of MinMI estimators (15).

Two problems are then addressed:

- I The estimation of bias, variance, quantiles and distribution of estimators of the incremental MinMI $I_{j/p}$ issued from finite samples of N (*iid*) realizations of bivariate original variables $(\hat{X}, \hat{Y})$ and then transformed into RVs (X, Y)
- II The distribution of estimators of $I_{j/p}$ under the null hypothesis H_0 that (X, Y) follows the ME distribution constrained by a weaker constraint set $(\mathbf{T}_p, \boldsymbol{\theta}_p)$ ($j > p$). These estimators work as a significance test for determining whether there is statistically significant MI beyond that explained by cross moments in $(\mathbf{T}_p, \boldsymbol{\theta}_p)$.

4.1. Bias, Variance, Quantiles and Distribution of MI Estimation Error

Considering the moment error distribution (31) and plugging it into the development (9), the error of the MI estimator $I_{N,j/p}$ is then distributed as:

$$\Delta I_{N,j/p,\mathbf{0}} \approx N^{-1/2} [\mathbf{v}_{j/p}^T (\mathbf{C}_{N,cr,j|lms})^{1/2}] \mathbf{U}_j + 1/2 N^{-1} \mathbf{U}_j^T [(\mathbf{C}_{N,cr,j|lms})^{1/2} \mathbf{A}_{j/p} (\mathbf{C}_{N,cr,j|lms})^{1/2}] \mathbf{U}_j \quad (32)$$

where neglected terms are of order $O(N^{-3/2})$. That is a second-order polynomial form of a multivariate standard Gaussian RV $\mathbf{U}_j \sim \mathcal{N}(\bar{\mathbf{0}}_j, \mathbf{P}_{cr,j})$. There is no general analytical expression for the PDF inferred from (32), except in certain cases where $\Delta I_{N,j/p}$ is governed by a non-central Chi-squared distribution [36]. The quantiles determining the confidence intervals of $I_{N,j/p}$ can easily be obtained by sorting of Monte-Carlo surrogates (proxies) of (32) from a pseudo-random generator of a standard Gaussian. Analytical expressions of the distribution of MI estimates are given from a MI Taylor expansion in terms of the anomalies of the estimated probabilities [27,37]. Here, we adopt a different approach by considering anomalies of the estimated expectations.

The bias of $I_{N,j/p}$ or the expectation of $\Delta I_{N,j/p,\mathbf{0}}$ is derived from the mean of the quadratic form term in (32). Therefore, taking the invariance of the trace for the circular permutation of a matrix product, that bias is approximated by the asymptotic value:

$$E(\Delta I_{N,j/p}) \approx (1/2)N^{-1}\text{Tr}(\mathbf{C}_{N,cr,j|lms} \mathbf{A}_{j/p}) \\ = (1/2)N^{-1} \left[\text{Tr}(\mathbf{C}_{N,cr,j|lms} \mathbf{P}_{cr,j} \mathbf{C}_{*j}^{-1} \mathbf{P}_{cr,j}) - \text{Tr}(\mathbf{C}_{N,cr,p|lms} \mathbf{P}_{cr,p} \mathbf{C}_{*p}^{-1} \mathbf{P}_{cr,p}) \right] \quad (33)$$

This is the difference between maximum entropy N^{-1} -scaled biases of orders j and p , subjected to the imposition of marginal PDFs. We must remember that if $p = 0$, $\mathbf{P}_{cr,p}$ is zero. For this case the MinMI bias is simply minus the negative bias of the ME $H(\boldsymbol{\theta}_{N,j})$, which is treated without the effect of variable morphism by [26]. When data is governed by the MinMI-PDF of order j , the matrices $\mathbf{C}_{N,cr,j|lms}$ and $\mathbf{P}_{cr,j} \mathbf{C}_{*j}^{-1} \mathbf{P}_{cr,j}$ are the inverse of each-other, according to Theorems 1 and 2 (11,27), leading to $E(\Delta I_{N,j/0}) = (1/2)N^{-1}\text{Tr}(\mathbf{C}_{N,cr,j|lms} \mathbf{P}_{cr,j} \mathbf{C}_{*j}^{-1} \mathbf{P}_{cr,j}) = (1/2)N^{-1}\text{Tr}(\mathbf{P}_{cr,j})$, i.e., $1/(2N)$ times the number of cross constraints. However, as argued by [26], when the true data distribution is more leptokurtic than the MinMI-PDF, then the bias can be larger than $(1/2)N^{-1}\text{Tr}(\mathbf{P}_{cr,j})$.

By assuming the limit case of Gaussianity, the variance of $\Delta I_{N,j/p}$ comes as:

$$\text{var}(\Delta I_{N,j/p}) \approx N^{-1}\text{Tr} \left[\mathbf{C}_{N,cr,j|lms} (\mathbf{v}_{j/p} \mathbf{v}_{j/p}^T) \right] + (1/2)N^{-2}\text{Tr} \left[(\mathbf{C}_{N,cr,j|lms} \mathbf{A}_{j/p})^2 \right] \quad (34)$$

The leading variance term is N^{-1} -scaled as generally deduced in [15]. Keeping the leading term of (34), and dealing with the trace, we get a given relative error $r_I = \Delta I_{N,j} / I_j$ of the MinMI $I_{j/0}$ ($p=0$) when $N \geq E((\lambda_{cr,j}^T T_{cr,j}')^2) / (I_{j/0} r_I)^2 \approx O(m_{cr,j}) / (I_{j/0} r_I)^2$. The term $O(m_{cr,j})$ increases with a larger rate than $I_{j/0}$ as far as the bound of the polytope of allowed expectations is closer.

4.2. Significance Tests of MinMI Thresholds

The estimators $I_{N,j/p}$ allow for the elaboration of statistical significance tests in order to verify whether the empirical PDF differs considerably from a threshold ME-PDF or in the contrary if the difference can be justified by sampling errors.

Let us suppose the null hypothesis H_0 considering that the true PDF coincides to the ME-PDF constrained by $(\mathbf{T}_p, \boldsymbol{\theta}_p)$. In particular for $(\mathbf{T}_p, \boldsymbol{\theta}_p) = (\mathbf{T}_{p=0}, \boldsymbol{\theta}_{p=0}) = (\mathbf{T}_{ind}, \boldsymbol{\theta}_{ind})$, the null hypothesis states that (X, Y) are statistically independent. Therefore under H_0 , the moment sets $(\mathbf{T}_p, \boldsymbol{\theta}_p), (\mathbf{T}_j, \boldsymbol{\theta}_j)$ are ME-congruent and the moments of order $j \geq p$ remain well determined by expectations over the less restricted p -th ME-PDF i.e., $\boldsymbol{\theta}_j = E_{\rho_{\mathbf{T}_p, \boldsymbol{\theta}_p}}^*(\mathbf{T}_j) \equiv \boldsymbol{\theta}_{j \leftarrow p}$ where the subscript arrow $j \leftarrow p$ means that j -order statistics are obtained by the p -order ME-PDF. The same holds for the ME covariance matrices, i.e., $\mathbf{C}_{*p} = \mathbf{C}_p$ and $\mathbf{C}_{*j} = \mathbf{C}_{*j \leftarrow p} = \mathbf{C}_j$; $j \geq p$. In these conditions, the matrix $\mathbf{C}_p$ is simply a sub-matrix of $\mathbf{C}_j$. The Lagrange multipliers are restricted to the p -order i.e. $\boldsymbol{\lambda}_j = \boldsymbol{\lambda}_{j \leftarrow p} = (\boldsymbol{\lambda}_p, \vec{0}_{j/p})$; $j \geq p$, where entries of higher order than p are set to zero leading to $\mathbf{v}_{j/p} = \mathbf{0}$ in (9). Therefore, the incremental MinMI vanishes, i.e. $H(\boldsymbol{\theta}_j) - H(\boldsymbol{\theta}_p) = I_{j/p} = 0$, but the estimator of $I_{N,j/p}$ is positive due to artificial MI generation stemming from the presence of sampling errors. Then, under H_0 , and using (9), the MI estimation is provided by the following approximation:

$$H(\boldsymbol{\theta}_{N,p}) - H(\boldsymbol{\theta}_{N,j}) | H_0 \equiv \delta I_{N,j/p} \approx (1/2)N^{-1} \mathbf{U}_j^T [(\mathbf{C}_{N,cr,j|lms})^{1/2} \mathbf{A}_{j \leftarrow p} (\mathbf{C}_{N,cr,j|lms})^{1/2}] \mathbf{U}_j \\ \mathbf{U}_j \sim \mathcal{N}(\vec{0}_j, \mathbf{P}_{cr,j}) \quad ; \quad \mathbf{A}_{j \leftarrow p} = \mathbf{P}_{cr,j} (\mathbf{C}_j)^{-1} \mathbf{P}_{cr,j} - \mathbf{P}_{cr,p} (\mathbf{C}_p)^{-1} \mathbf{P}_{cr,p} \quad (35)$$

where $\mathbf{A}_{j \leftarrow p}$ is a positive semi-definite matrix. That works as a significance test for the non-verification of H_0 ; in other words, if $I_{N,j/p}$ is larger than an upper $1-\alpha$ quantile (e.g., $1-\alpha=95\%$) of $\delta I_{N,j/p}$, then H_0 is rejected with a significance level α . Those quantiles determine the significant MI thresholds and can be computed empirically as for the MinMI error (32) by a Monte-Carlo strategy. Another possibility is the fitting of the $\delta I_{N,j/p}$ distribution to a Gamma PDF with prescribed mean and variance (not done here). The bias and variance of $\delta I_{N,j/p}$ are straightforward, coming as:

$$E[\delta I_{N,j/p}] \approx (1/2)N^{-1} \text{Tr}[\mathbf{C}_{N,cr,j|lms} \mathbf{A}_{j \leftarrow p}] ; \quad \text{var}[\delta I_{N,j/p}] \approx (1/2)N^{-2} \text{Tr}[(\mathbf{C}_{N,cr,j|lms} \mathbf{A}_{j \leftarrow p})^2] \quad (36)$$

The N^{-2} -scale for variance is also present in other MI estimate errors under the hypothesis of variable independency [27]. Under the Theorems 1 [11] and 2 [27], along with the null hypothesis, one gets $\mathbf{C}_{N,cr,j|lms} \mathbf{A}_{j \leftarrow p} = \mathbf{P}_{cr,j} - \mathbf{P}_{cr,p}$, thus leading to a Chi-Squared distribution for $\delta I_{N,j/p}$:

$$\delta I_{N,j/p} \sim (1/2)N^{-1} \chi_{nd}^2 ; \quad nd = \text{Tr}(\mathbf{P}_{cr,j} - \mathbf{P}_{cr,p}) \quad (37)$$

with nd degrees of freedom, *i.e.*, the difference between the number of cross moments of order j and p . From that, the upper quantiles necessary for statistical significance are easily obtained from χ^2 probability lookup tables. The bias and variance are, respectively:

$$E[\delta I_{N,j/p}] \approx (1/2)N^{-1} [\text{Tr}(\mathbf{P}_{cr,j} - \mathbf{P}_{cr,p})] ; \quad \text{var}[\delta I_{N,j/p}] \approx (1/2)N^{-2} [\text{Tr}(\mathbf{P}_{cr,j} - \mathbf{P}_{cr,p})] \quad (38)$$

By analyzing (38), and in order to get a test with a relative error $r_I = (\Delta I_{\min} / I_{\min})$, one must choose $N \geq ((m_{cr2} - m_{cr1}) / 2)^{1/2} / (I_{\min} r_I)$.

4.3. Significance Tests of the Gaussian and Non-Gaussian MI

In this section we particularize the theory presented in Section 4.1 and 4.2 (Equations 35–38) for the case of Gaussian and non-Gaussian MIs as defined in Section 2.3. For this purpose, let us consider the moment sets (13) and the MI components I_g and $I_{ng,j}$ (11). Their finite estimators are:

$$\begin{aligned} I_{N,g} &= H(\boldsymbol{\theta}_0) - H(\boldsymbol{\theta}_{N,2}) = I_g + \Delta I_{N,g} = I_{N,j=2/p=0} ; \quad \Delta I_{N,g} = I_g(c_g + \Delta c_{g,N}) - I_g(c_g) = -\Delta H(\boldsymbol{\theta}_{N,2}) ; \\ I_{N,ng,j} &= H(\boldsymbol{\theta}_{N,2}) - H(\boldsymbol{\theta}_{N,j}) = I_{ng,j} + \Delta I_{N,ng,j} = I_{N,j/p=2} ; \quad \Delta I_{N,ng,j} = \Delta H(\boldsymbol{\theta}_{N,2}) - \Delta H(\boldsymbol{\theta}_{N,j}) \end{aligned} \quad (39)$$

where $\Delta I_{N,g}, \Delta I_{N,ng,j}$ are MinMI errors, $\Delta c_{g,N}$ is the Gaussian correlation estimation error, $H(\boldsymbol{\theta}_0) = 2H_g$ with $H_g \equiv \frac{1}{2} \log(2\pi e)$ being the entropy of the univariate standard Gaussian; $\boldsymbol{\theta}_{N,j} = \boldsymbol{\theta}_j + \Delta \boldsymbol{\theta}_{N,j}$; $j \geq 1$ are the expectations obtained from the N -sized Gaussianized standardized sample.

The numerical implementation of the maximum entropy estimator $\hat{H}$ (16), approximating H is computed over a number N_b bins of an extended enough finite interval $[-L_i, L_i]$. In the corresponding experiments (and as in PP12), we have used the calibrated values $L_i=6$ and $N_b=80$. The used algorithm is explained in detail in the appendix 2 of PP12 [12], following an adapted bivariate version of that of [35]. The error $\delta H = \hat{H} - H$ is of the order of round-off errors, only becoming comparable to the sampling ME errors at very high values of N .

4.3.1. Error and Significance Tests of the Gaussian MI

The Gaussian MI error $\Delta I_{N,g}$ depends on the Gaussian correlation estimation's error $\Delta c_{g,N} \equiv c_{g,N} - c_g$ where $c_{g,N} = E_N(XY)$ is inferred from the sample. Let us write (9) for $\Delta I_{N,g}$. The Gaussian bivariate ME-PDF, constrained by $(\mathbf{T}_2 = (X, X^2, Y, Y^2, XY)^T, \boldsymbol{\theta}_2 = (0, 1, 0, 1, c_g)^T)$ is $\rho_{\mathbf{T}_2, \boldsymbol{\theta}_2}^*(X, Y) = [4\pi^2(1 - c_g^2)]^{-1/2} \exp[-(1/2)(1 - c_g^2)^{-1}(X^2 - 2c_gXY + Y^2)]$, leading to the vector of Lagrange multipliers $\boldsymbol{\lambda}_2 = [0, -(1/2)(1 - c_g^2)^{-1}, 0, -(1/2)(1 - c_g^2)^{-1}, c_g(1 - c_g^2)^{-1}]^T$. The projector operator $P_{cr,2}$ onto cross moments is the 5x5 matrix that extracts the 5th entry (row and column) of $\mathbf{T}_2$, corresponding to the unique cross moment XY . The necessary 5x5 covariance matrix is $\mathbf{C}_{*,2} = E_{\rho_{\mathbf{T}_2, \boldsymbol{\theta}_2}^*}[\mathbf{T}_2 \mathbf{T}_2^T] - \boldsymbol{\theta}_2 \boldsymbol{\theta}_2^T$, where the E operator is the expectation over the bivariate Gaussian $\rho_{\mathbf{T}_2, \boldsymbol{\theta}_2}^*$. Then, we apply (9) for $j=2, p=0$ where $\Delta \boldsymbol{\theta}_{N,j} = (0, 0, 0, 0, \Delta c_{g,N})^T$. The Gaussian MI error is written in different forms as:

$$\begin{aligned} \Delta I_{N,g} &\approx (P_{cr,2} \boldsymbol{\lambda}_2)^T (\Delta c_{g,N}) + \frac{1}{2} (P_{cr,2} \mathbf{C}_{*,2}^{-1} P_{cr,2}) (\Delta c_{g,N})^2 = \frac{c_g}{1 - c_g^2} (\Delta c_{g,N}) + \frac{1 + c_g^2}{2(1 - c_g^2)^2} (\Delta c_{g,N})^2 = \\ &= \frac{\partial I_g}{\partial c_g} \Delta c_{g,N} + \frac{1}{2} \frac{\partial^2 I_g}{\partial c_g^2} (\Delta c_{g,N})^2 \end{aligned} \quad (40)$$

There, the term $P_{cr,2} \boldsymbol{\lambda}_2$ is the fifth component of $\boldsymbol{\lambda}_2$, corresponding to the first derivative of I_g with respect to c_g whereas the term $P_{cr,2} \mathbf{C}_{*,2}^{-1} P_{cr,2}$ is the entry of $\mathbf{C}_{*,2}^{-1}$ at row 5, column 5, corresponding to the second derivative of I_g . The bias and variance of $\Delta I_{N,g}$ depend on the distribution of the Gaussian correlation error $\Delta c_{g,N}$. According to the proposed modeling of moment estimation errors (Theorem 2 of section 3.4), $\Delta c_{g,N}$ is asymptotically Gaussian with a negligible bias $E(\Delta c_{g,N}) \approx 0$ and a variance (under imposed marginals) given by:

$$\text{var}(\Delta c_{g,N}) \approx N^{-1} \text{var}(XY | E(XY | X), E(XY | Y)) = (1 - c_g^2)^2 / (1 + c_g^2) \quad (41)$$

However, in order to keep the simulated $c_g = c_{g,N} - \Delta c_{g,N}$ within the interval $[-1, 1]$, one can use the more precise Fisher Z-transform [38] such that $\Delta c_{g,N} = \tanh(\tanh^{-1}(c_g) + \Delta Z_N) - c_g$, where ΔZ_N has a mean and variance of order $O(N^{-1})$.

In order to test the null hypothesis that the variable pair (X, Y) has a joint bivariate isotropic Gaussian distribution, we must compare the estimated $I_{N,g}$ with upper quantiles of the significance test $\delta I_{N,g}$, given by $\Delta I_{N,g}$ (40) with $c_g = 0$ and $\Delta c_{g,N} \sim \mathcal{N}(0, N^{-1})$. This is a Gaussian correlation significance test that is Chi-squared distributed, with:

$$\begin{aligned} \delta I_{N,g} &= (1/2)(\Delta c_{g,N})^2 = (1/2)N^{-1}U^2 \sim (1/2)N^{-1}\chi_1^2; \quad U \sim \mathcal{N}(0, 1) \\ E(\delta I_{N,g}) &= (1/2)N^{-1}; \quad \text{var}(\delta I_{N,g}) = (1/2)N^{-2} \end{aligned} \quad (42)$$

4.3.2. Error and Significance Tests of the Non-Gaussian MI

The estimation error $\Delta I_{N,ng,j}$ of the non-Gaussian MI as defined in (39) can be written as a particular form of (9) for an even order $j \geq 4$ and $p=2$ as function of the vector $\Delta \boldsymbol{\theta}_{N,j}$ of moment errors of the moment vector $\mathbf{T}_j$ (13) with a certain chosen component indexation. Therefore, the matrix $\mathbf{A}_{j/p} = \mathbf{A}_{j/p=2} \equiv \mathbf{P}_{cr,j}(\mathbf{C}_{*j})^{-1}\mathbf{P}_{cr,j} - \mathbf{P}_{cr,2}(\mathbf{C}_{*2})^{-1}\mathbf{P}_{cr,2}$ of (9) comprises the inverses of covariance matrices $\mathbf{C}_{*j}$ and $\mathbf{C}_{*2}$, respectively of the j -th and 2nd order ME solutions.

Algebraic consistency sets the matrix $\mathbf{P}_2(\mathbf{C}_{*2})^{-1}\mathbf{P}_2$ to the embedding of $(\mathbf{C}_{*2})^{-1}$ onto the j -th moment subspace. Then we will perform a range of experiments for the validation of approximations in Section 4.2. The vector $\mathbf{v}_{j/p=2} \equiv \mathbf{P}_{cr,j}\boldsymbol{\lambda}_j - \mathbf{P}_{cr,2}\boldsymbol{\lambda}_2$ comprises Lagrange multiplier vectors of the ME solutions of orders j and 2.

In order to compute the bias, variance, quantiles and confidence intervals of $I_{N,ng,j}$, from N -sized samples, there are two possible strategies: either pure Monte-Carlo simulations or the analytical and the semi-analytical (analytical with moment's error surrogates) approaches as explained in section 1. In the pure Monte-Carlo approach, either a known bivariate PDF is assumed or surrogates of the joint PDF are generated through multivariate bootstrapping techniques [39] preserving the copula structure. For each generated sample from an extended ensemble of N_{rea} (e.g., 5000) realizations, we compute moments and solve the ME problem gathering statistics afterwards. Alternatively, ME errors can be computed from the Taylor expansion (9) from moment deviations over the ensemble.

In the analytical and semi-analytical approaches, moment errors $\Delta \boldsymbol{\theta}_{N,j}$ are assumed to follow a certain parametric distribution that can be multivariate Gaussian as in (31), based on a given bias-covariance matrix modeling or a more sophisticated approach taking into account the natural bounds of the simulated moments $\boldsymbol{\theta}_{cr,j} = \boldsymbol{\theta}_{N,cr,j} - \Delta \boldsymbol{\theta}_{N,cr,j}$. Then, MinMI statistics are computed from statistics (bias, variance, quantiles) on ensembles of error surrogates.

The non-Gaussian MIs $I_{N,ng,j}$ (even $j \geq 4$) work as tests measuring significant statistical deviations from the null hypotheses of joint Gaussianity. These statistical tests are given by Kullback-Leibler distances (7) and constitute an alternative to the use of algebraic deviations of moments from those given by the bivariate Gaussian (e.g., bivariate cumulants) [40].

The non-Gaussianity test of order j is given by $\delta I_{N,ng,j} \equiv H(\boldsymbol{\theta}_{N,2}) - H(\boldsymbol{\theta}_{N,j}) | H_0$ under the null hypothesis H_0 that the true PDF is bivariate Gaussian and is written as a particular case of (35). However, a simplification of the statistical test formula can be achieved by considering a null Gaussian correlation. This holds thanks the non-Gaussian MI invariance under variable rotations (see PP12), in particular for uncorrelated standardized variables $(X_r, Y_r)^T = \mathbf{A}(X, Y)^T$, where $\mathbf{A}$ is the rotation matrix (e.g. $X_r = X, Y_r = (Y - c_g X)(1 - c_g^2)^{-1/2}$, i.e., the residual of the linear prediction). Under H_0 , the rotated variables are still bivariate Gaussian and therefore the non-Gaussianity significance test $\delta I_{N,ng,j}$ has the same distribution as that for $c_g = 0$. The matrices $\mathbf{C}_{N,cr,j|lms}$ and $\mathbf{A}_{j \leftarrow 2}$ entering in Equation (35) are now evaluated for Gaussian isotropic conditions. For the sake of clarity, we represent them respectively by $\mathbf{C}_{g,N,cr,j|lms}$, $\mathbf{A}_{g,j \leftarrow 2} = \mathbf{P}_j(\mathbf{C}_{g,j})^{-1}\mathbf{P}_j - \mathbf{P}_2(\mathbf{C}_{g,2})^{-1}\mathbf{P}_2$, where the subscript g stands for

evaluation at $(X, Y)^T \sim \mathcal{N}(\vec{0}, \mathbf{I})$. For high N , $\mathbf{C}_{g,N,cr,j|lms} = \mathbf{C}_{g,j}$, i.e., the covariance matrix of cross j -th order moments for the isotropic Gaussian. Then we write:

$$\delta I_{N,ng,j} \approx (1/2)N^{-1}\mathbf{U}_j^T \left[(\mathbf{C}_{g,N,cr,j|lms})^{1/2} \mathbf{A}_{g,j \leftarrow 2} (\mathbf{C}_{g,N,cr,j|lms})^{1/2} \right] \mathbf{U}_j \quad (43)$$

Let us specify generic entries at row α , column β of those matrices, corresponding to monomials $X^{r_\alpha} Y^{s_\alpha}$ and $X^{r_\beta} Y^{s_\beta}$ of $\mathbf{T}_j$, i.e. with $r_\alpha + s_\alpha, r_\beta + s_\beta \leq j$. Then, using the notation introduced in Section 3.3 for Gaussian standard moments $\mu_r \equiv E(X^r)$; $\mu_{N,r} \equiv E_N(X^r)$, $r \in \mathbb{N}_0$, the components of $\mathbf{C}_{g,j}$ become:

$$(\mathbf{C}_{g,j})_{\alpha,\beta} = \mu_{r_\alpha+r_\beta} \mu_{s_\alpha+s_\beta} - \mu_{r_\alpha} \mu_{r_\beta} \mu_{s_\alpha} \mu_{s_\beta} \quad (44)$$

whereas the components of the *lms* covariances are:

$$(\mathbf{C}_{g,N,cr,j|lms})_{\alpha,\beta} = \mu_{N,r_\alpha+r_\beta} \mu_{N,s_\alpha+s_\beta} - \mu_{N,s_\alpha+s_\beta} \mu_{N,r_\alpha} \mu_{N,r_\beta} - \mu_{N,r_\alpha+r_\beta} \mu_{N,s_\alpha} \mu_{N,s_\beta} + \mu_{N,r_\alpha} \mu_{N,r_\beta} \mu_{N,s_\alpha} \mu_{N,s_\beta} \quad (45)$$

The bias of the non-Gaussian MinMI and its asymptotic approximation (36) are given by:

$$E[\delta I_{N,ng,j}] \approx (1/2)N^{-1}[\text{Tr}(\mathbf{C}_{g,N,cr,j|lms} \mathbf{P}_{cr,j} \mathbf{C}_{g,j}^{-1}) - 1] = (1/2)N^{-1}(\text{Tr}(\mathbf{P}_{cr,j}) - 1) \quad (46)$$

Similarly and following (36), the variance becomes:

$$\text{var}[\delta I_{N,ng,j}] \approx (1/2)N^{-2} \text{Tr}[(\mathbf{C}_{g,N,cr,j|lms} \mathbf{A}_{g,j \leftarrow 2})^2] = (1/2)N^{-2}(\text{Tr}(\mathbf{P}_{cr,j}) - 1) \quad (47)$$

and the reasonable distribution approximation following (37):

$$\delta I_{N,ng,j} \sim (1/2)N^{-1} \chi_{nd}^2 \quad ; \quad nd = \text{Tr}(\mathbf{P}_{cr,j}) - 1 = j(j-1)/2 - 1 \quad (48)$$

from which bounds of significance levels of non-Gaussianity can be computed through quantiles of the Chi-squared distribution.

4.4. Validation of Significance Tests by Monte-Carlo Experiments

We have presented the theoretical expressions for the bias, variance and distribution, both for the Gaussian correlation test (42) and for the ME non-Gaussianity test of order j (46–48). Now we validate those expressions by comparing their results with statistics from large Monte-Carlo ensembles of ME computations. For that purpose, we have generated $N_{rea} = 5000$ independent synthetic datasets of N iid uncorrelated (X, Y) from a Gaussian random generator. We have set N from a duplication sequence: $N=25, 2^1*25, \dots, 2^{11}*25 = 51200$. Then, we have computed the 5,000 realizations for the independency test $\delta I_{N,g}$ as well as for the non-Gaussianity tests $\delta I_{N,ng,j}$ for $j = 4, 6, 8$. In order to minimize errors of type $\delta H(8)$, from the ME functional, we have retained only those Monte-Carlo realizations whose ME-PDF moments are within a relative square error of 10^{-5} .

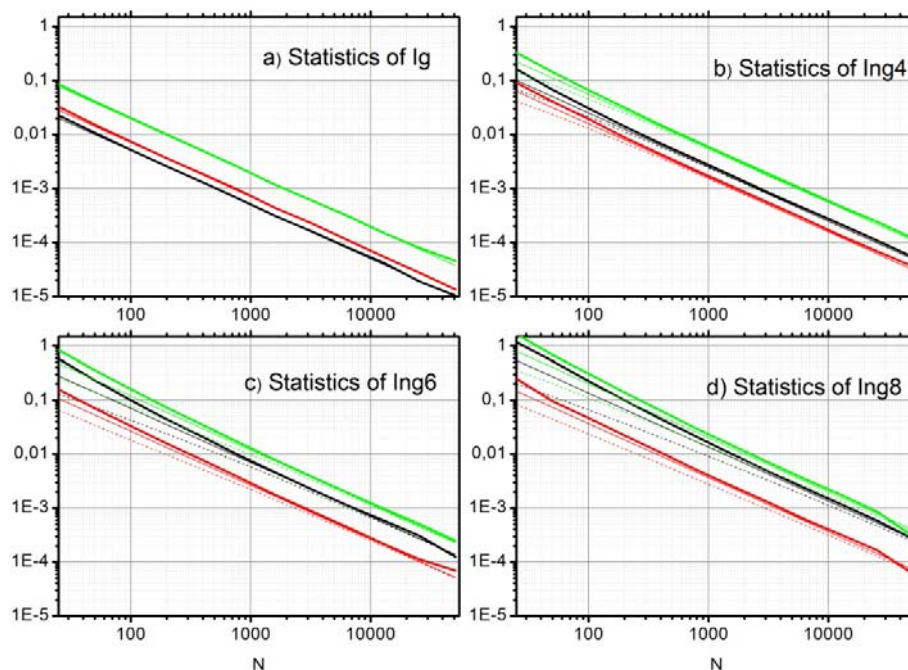
In the sequel, we have collected and compared the estimates of bias, standard deviation and the 95%-quantile, all provided by the three approaches: the Monte-Carlo (extended ensemble of ME computations), the semi-analytical (generation of Gaussian surrogates in the Taylor expansion of ME) and the analytical (analytical formulas based on the Theorems 1 and 2). The Figure 3a, b, c and d depict the above statistics of significance tests, respectively for $\delta I_{N,g}$ and $\delta I_{N,ng,j}$ ($j = 4, 6, 8$). The truth is assumed to be provided by the Monte-Carlo estimate.

As previously expected, significance tests are all scaled by $N^{-1}O(1)$, and consequently their bias, standard deviation and quantiles are $N^{-1}O(1)$ as shown in Figures 3a-d by estimates coming from the different approaches. MinMI biases and significance thresholds (the 95% quantiles) grow for higher number of constraints as in the sequence $I_{N,g}, I_{N,ng,j=4}, I_{N,ng,j=6}, I_{N,ng,j=8}$.

These results mean that those estimators are progressively better (stronger) evaluations of MI (or the MI beyond that explained by Gaussianity), though they call for progressively higher significance thresholds. Therefore, especially in cases of under-sampled data (small N) or very low MI (or Non-Gaussian MI) values (weakly dependent variables or weak joint non-Gaussianity), there must be a tradeoff between N and the number of parameters of the MinMI estimator (here the number of cross constraints).

At this point, we discuss how the analytical and semi-analytical estimates of MinMI error statistics fit the Monte-Carlo (true) statistics. There are three crucial factors in our approximations: (1) The accuracy of the ME Taylor expansion, valid for small enough sampling errors (N large); (2) The convergence rate towards Gaussian statistics (from the CLT) for high N .

Figure 3. Test statistics: bias (black lines), standard deviation (red lines) and 95%-quantiles (green lines), provided by the Monte-Carlo approach (thick full lines), the semi-analytical approach (thin dashed lines) and the analytical approach (tick full lines). The tests are $\delta I_{N,g}$ (a); $\delta I_{N,ng,j=4}$ (b); $\delta I_{N,ng,j=6}$ (c) and $\delta I_{N,ng,j=8}$ (d).

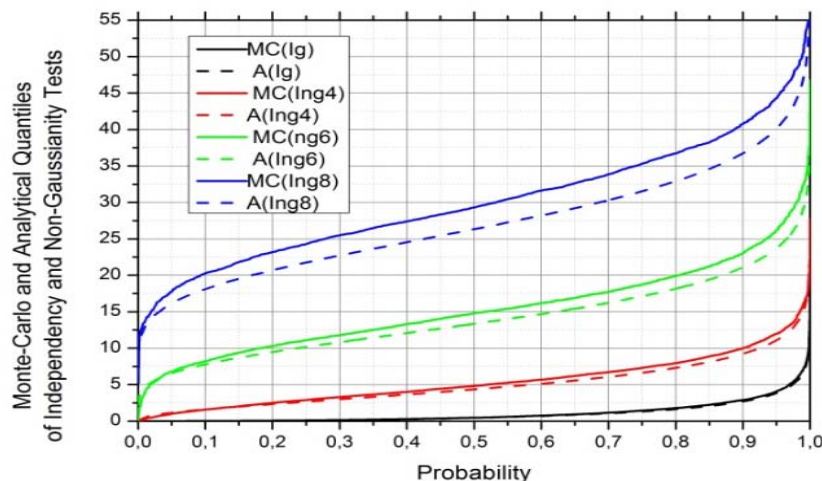


The analytical bias depends on factors 1 and 3, while formulas for variance, distribution and quantiles depend on all above factors, being only valid for N high enough. From Figure 3a-d, we see that the agreement between analytical and Monte-Carlo statistics is quite good for all tests (with a slight analytical underestimation), though only for large enough $N > N_{test}$ values where N_{test} depends on how later (in N) the factors 1-3 hold together. We have $N_{test} \approx 50, 400, 1600, 3200$, respectively for $\delta I_{N,g}, \delta I_{N,ng,j=4}, \delta I_{N,ng,j=6}, \delta I_{N,ng,j=8}$, growing with the number of constraints. The exception is when N

is so large that errors δH of the operational ME (typically, round-off errors) are of the same order of the small value tests δI , starting to influence the Monte-Carlo statistics.

In order to validate the analytical Chi-Squared distributions for the tests, we present in Figure 4, the empirical cumulative histograms, respectively of $2N\delta I_{N,g}$, $2N\delta I_{N,ng,j}$, $2N\delta I_{N,ng,6}$, $2N\delta I_{N,ng,8}$ for $N \approx N_{test}$ and the corresponding theoretical cumulative Chi-Squared PDF fits, respectively χ^2_1 , χ^2_5 , χ^2_{14} and χ^2_{27} . The agreement is shown to be quite good, with a slight deficit in the theoretical number of degrees of freedom, possibly due to uncontrolled aspects (e.g., the numerical implementation of the ME algorithm and bound effects) leading to extra randomness. In fact, the theoretical prediction of MinMI bias results from two matrices, theoretically equal, which are issued from extraordinary complicated outputs (the MinMI covariance matrix and the covariance matrix of estimators under fixed marginals). The theoretical result depends on the matching of a huge number of algorithmic details. The results provide good support to the presented Theorems, the hypotheses on the basis of the analytical and semi-analytical approaches. The slightly higher MinMI bias than the theoretical one is due to a small difference between the data PDF and the ME-PDF.

Figure 4. Monte-Carlo empirical cumulative histogram (solid lines) and theoretical cumulative Chi-Squared fit (dashed lines) normalized by N : $2N\delta I_{N,g}(\chi^2_1)$ for $N = 50$ (black curves); $2N\delta I_{N,ng,j=4}(\chi^2_5)$ for $N = 400$ (red curves); $2N\delta I_{N,ng,6}(\chi^2_{14})$ for $N = 1600$ (green curves) and $2N\delta I_{N,ng,8}(\chi^2_{27})$ for $N = 3200$ (blue curves).



5. MI Estimation from Under-Sampled Data

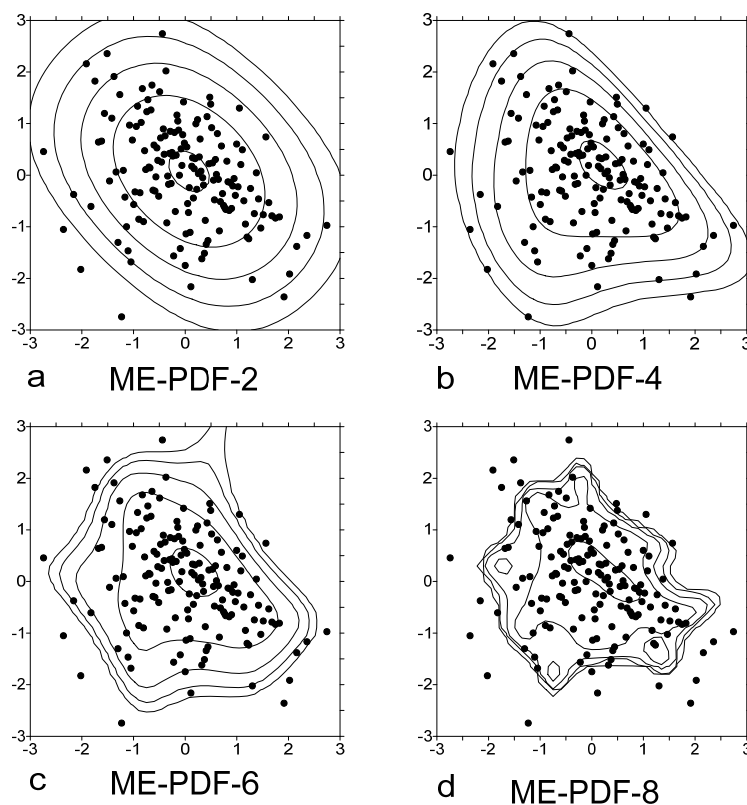
In this section, we present a case of MinMI estimation from under-sampled data (N small), emphasizing the effect of MI bias and its relation to PDF over-fitting. For this purpose, we consider an example from meteorology, already introduced by authors [8] in which X, Y are the standard Gaussian morphism ($X, Y \sim \mathcal{N}(0,1)$) of monthly means in winter (December to February), respectively of the North Atlantic Index (X) (a quite useful planetary-scale atmospheric index [41]), and the amount of rainfall in Greenland (Y). The paper [8] has shown the existence of statistically significant nonlinear correlations between X and Y , i.e., non-Gaussian MI. The data used in the study comes from the

NCEP/NCAR meteorological reanalysis for the period 1951–2003, leading to temporal series with length equal to 159, from which we have estimated the number $N \sim 100$ of *iid* data (temporal degrees of freedom), after discarding the effect of temporal auto-correlation [42].

Figure 5a–d present the scatter-plot of the (X, Y) pairs along with the contours of the ME-PDF fitting constrained by bivariate monomial expectations $\mathbf{T}_j$ (13) of total order $j = 2, 4, 6$ and 8 respectively. There is pictorial evidence of PDF over-fitting for cases of a high number of cross constraints (14 and 27 for $j = 6, 8$ respectively) in Figures 5c and d. In those cases, the dataset bivariate outliers, which lie at very poorly probable regions of the PDF, tend to give a polygonal character to the PDF extreme contours.

The MinMI values in *nats* are $I_{N,g} = \mathbf{0.053}$ (0.048), $I_{N,ng,4} = \mathbf{0.071}$ (0.041), $I_{N,ng,6} = 0.086$ (~ 0) and $I_{N,ng,8} = 0.196$ (~ 0) with unbiased values in parenthesis and figures marked bold where the null hypothesis H_0 is rejected at the 5% significance level (values above the 95% error quantile). That means that variables are significantly correlated with the unbiased Gaussian correlation $c_g = -0.30$ and a statistically significant, though small, non-Gaussian unbiased MI of order $j = 4$ of 0.041 *nats*, which has been shown to be of the same order of the Gaussian MI. None of the remaining incremental MinMIs are significant, which corroborates the fact that the values of $I_{N,ng,6}$ and $I_{N,ng,8}$ are purely artificial.

Figure 5. Scatter-plot of the Gaussianized variables X (in abscissas) Y (in ordinates) (see text for details) along with ME-PDF fitting constrained by monomial bivariate moments up to order $j = 2$ (a), $j = 4$ (b), $j = 6$ (c) and $j = 8$ (d). Contour levels are set to 0.0005, 0.005, 0.05, 0.5, and 5.



6. Discussion and Conclusions

This paper presents theoretical formulas for statistics (bias, variance, distribution) of estimation errors of information theoretical measures. This is quite relevant because finite samples can apparently exhibit artificial statistical structures leading to negatively biased estimations of Entropy or positively biased estimations of Mutual Information. By using Monte-Carlo experiments, we empirically validate certain results about the asymptotic distribution of estimation errors of the minimum Mutual Information (MinMI) between two random variables X, Y .

That MinMI is the least committed MI compatible with prescribed marginal X and Y distributions and a set $\mathbf{T}_{cr}$ of a number m_{cr} of expectations of cross X, Y joint functions $\mathbf{T}_{cr}(X, Y)$, filling up a vector $\boldsymbol{\theta}_{cr} = E(\mathbf{T}_{cr})$ where MinMI is written in terms of Shannon entropies (H) as: $I_{\min}(X, Y) = H(X) + H(Y) - H_{\max}(X, Y)$. There, $H_{\max}$ is the maximum entropy (ME) constrained by marginals and cross mean constraints. The MinMI is a lower MI bound, converging to the total MI when the set $\mathbf{T}_{cr}$ converges to the sufficient joint statistics. Sampling $\boldsymbol{\theta}_{cr}$ errors from N -sized samples, say $\Delta\boldsymbol{\theta}_{N,cr} = \boldsymbol{\theta}_{N,cr} - \boldsymbol{\theta}_{cr}$ lead to MinMI errors. In order to compute MinMI, the marginal PDFs of finite samples must be preset by morphisms, setting the X and Y single values to fixed quantiles. This reduces the sampling randomness to the covariate sampling in the form of random permutations in the bivariate trials (X, Y) . Then, the estimator variance $\text{var}(\Delta\boldsymbol{\theta}_{N,cr})$ is scaled by N^{-1} , being lower than the value $N^{-1} \text{var}(\mathbf{T}_{cr})$, valid in the case of random *iid* marginal trials. In order to get a given MinMI relative error $r_I = (\Delta I_{\min} / I_{\min})$, one must choose $N \geq E((\lambda_{cr}^T \mathbf{T}_{cr})^2) / (I_{\min} r_I)^2 \approx O(m_{cr}) / (I_{\min} r_I)^2$ where one uses the Lagrange multipliers associated to cross moments and also the perturbations $\mathbf{T}_{cr}'$.

The detailed analysis of $\Delta\boldsymbol{\theta}_{N,cr}$ has shown that $\text{var}(\Delta\boldsymbol{\theta}_{N,cr})$ under variable morphisms is given by $N^{-1} \text{var}(\mathbf{T}_{cr} | E(\mathbf{T}_{cr} | X), E(\mathbf{T}_{cr} | Y))$, which is the mean squared residual of the best linear fit of $\mathbf{T}_{cr}$ using the conditional means $E(\mathbf{T}_{cr} | X)$ and $E(\mathbf{T}_{cr} | Y)$ as predictors. This is supported by a few examples using a Monte-Carlo methodology. We have shown that $\text{var}(\Delta\boldsymbol{\theta}_{N,cr})$ is closely related to the Maximum Entropy solution constrained by $\mathbf{T}$ and marginal distributions, *i.e.*, the MinMI solution constrained by the cross constraints $\boldsymbol{\theta}_{cr} = E(\mathbf{T}_{cr})$.

The MinMI errors are readily obtained from MinMI second-order Taylor development in terms of $\Delta\boldsymbol{\theta}_{N,cr}$. Asymptotically, $\Delta\boldsymbol{\theta}_{N,cr}$ is multivariate Gaussian thanks to the Central Limit Theorem. The MinMI bias is positive, given by the mean of a positive quadratic form of Gaussians. When data samples come from the same distribution as the one generated from MinMI, the MinMI bias is simply $1/(2N) m_{cr}$. However, the bias can increase/decrease when data comes from a more leptokurtic/platykurtic distribution. That expression of bias comes from the fact that the Hessian matrix of MinMI in terms of the vector of cross constraints $\boldsymbol{\theta}$ is the inverse of the covariance matrix of the cross functions $\mathbf{T}$, conditioned to the knowledge of marginal PDFs. That matrix is the matrix of mean squared residuals of best linear fit of $\mathbf{T}$ using predictors $E(\mathbf{T}_{cr} | X), E(\mathbf{T}_{cr} | Y)$ evaluated at the MinMI-PDF.

We have further introduced the incremental MinMI given by the difference $H_{\max 1} - H_{\max 2}$ between two MEs, forced by cross constraint sets $\mathbf{T}_{cr1} \subseteq \mathbf{T}_{cr2}$. Under the null hypothesis $H_{\max 1} = H_{\max 2}$, the incremental MinMI stands for a statistical test evaluating the existence of statistically significant MI

explained by cross expectations in the set difference $\mathbf{T}_{cr2}/\mathbf{T}_{cr1}$. This test is distributed as $\frac{1}{2N} \chi^2_{(m_{cr2}-m_{cr1})}$ where m_{cr2}, m_{cr1} are the numbers of cross constraints respectively in $\mathbf{T}_{cr2}, \mathbf{T}_{cr1}$. In order to get a test with a relative error $r_I = (\Delta I_{\min} / I_{\min})$, one must choose $N \geq ((m_{cr2} - m_{cr1}) / 2)^{1/2} / (I_{\min} r_I)$.

By setting X, Y to single standard Gaussians by Gaussian morphisms and the single constraint product $\mathbf{T}_{cr} = XY$, we have evaluated the MI parcel that is explained by joint Gaussianity – the Gaussian MI. By adding further monomial bivariate as constraints, we can define the non-Gaussian MI, attributed to joint non-Gaussianity. Under the null hypothesis of null non-Gaussian MI tests the existence of statistically significant MI explained by nonlinear correlations, beyond the scope of Pearson correlation. This is an Information-Theoretic-based significance test of non-Gaussianity, beyond others based on multivariate cumulants.

Finally, we have evaluated the Gaussian and non-Gaussian MIs for real under-sampled data allowing illustrating the relationship between MI bias, probability density over-fitting and data outliers. Some questions do remain for future work, namely the implementation of fast algorithms for computing non-Gaussian MI and its generalization to more than two random variables.

Acknowledgments

This research was supported by the ERC advanced grant “Flood Change”, project No. 291152 and also the Projects PTDC/GEO-MT/3476/2012 and PEST-OE/CTE/LA0019/2011-FCT, funded by the Portuguese Foundation for Science and Technology (FCT). Thanks are due to three anonymous referees, to J. Macke and Susana Barbosa for some discussions and also our families for the omnipresent support.

References

1. Shannon, C.E. The mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons, Inc.: New York, NY, USA, 1991.
3. Averbeck, B.B.; Latham, P.E.; Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **2006**, *7*, 358 – 366.
4. Goldie, C.M.; Pinch, R.G.E. Communication Theory. In *London Mathematical Society Student Texts (No. 20)*; Cambridge University Press: Cambridge, UK, 1991.
5. Sims, C.A. Rational Inattention: Beyond the Linear-Quadratic Case. *Am. Econ. Rev.* **2006**, *96*, 158–163.
6. Sherwin, W.E. Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography. *Entropy* **2010**, *12*, 1765–1798.
7. Pothos, E.M.; Juola, P. Characterizing linguistic structure with mutual information. *Br. J. Psychol.* **2007**, *98*, 291–304.
8. 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.

9. 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.
10. Globerson, A.; Stark, E.; Vaadia, E.; Tishby, N. The minimum information principle and its application to neural code analysis. *Proc. Natl. Accd. Sci. USA* **2009**, *106*, 3490–3495.
11. Foster, D.V.; Grassberger, P. Lower bounds on mutual information. *Phys. Rev. E* **2011**, *83*, 010101(R):1–010101(R):4.
12. Pires, C.A.; Perdigão, R.A.P. Minimum Mutual Information and Non-Gaussianity Through the Maximum Entropy Method: Theory and Properties. *Entropy* **2012**, *14*, 1103–1126.
13. Walters-Williams, J.; Li, Y. Estimation of mutual information: A survey. *Lect. Notes Comput. Sci.* **2009**, *5589*, 389–396.
14. 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.
15. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1254.
16. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Comp. Neur. Syst.* **1996**, *7*, 87–107.
17. Victor, J.D. Asymptotic Bias in Information Estimates and the Exponential (Bell) Polynomials. *Neur. Comput.* **2000**, *12*, 2797–2804.
18. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Train Information Measures Correcting for the Sampling Bias Problem in Spike Information Measures. *J. Neurophysiol.* **2007**, *98*, 1064–1072.
19. Strong, S.P.; Koberle, R.; de Ruyter van Steveninck, R.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *86*, 197–200.
20. Miller, G. Note on the bias of information estimates. In *Information Theory in Psychology*; Quastler, H., Ed.; *II-B* Free Press: Glencoe, IL, USA, 1955; pp. 95–100.
21. Grassberger, P. Entropy Estimates from Insufficient Samplings. **2008**, *arXiv:physics/0307138v2.pdf*.
22. Bonachela, J.A.; Hinrichsen, H.; Muñoz, M.A. Entropy estimates of small data sets. *J. Phys. A* **2008**, *41*, 202001.
23. Nelsen, R.B. *An Introduction to Copulas*; Springer: New York, NY, USA, 1999; ISBN: 0-387-98623-5.
24. Calsaverini, R.S.; Vicente, R. An information-theoretic approach to statistical dependence: Copula information. *Europhys. Lett.* **2009**, *88*, 68003.
25. Ma, J.; Sun, Z.; Mutual information is copula entropy. **2008**, *arXiv:0808.0845v1*.
26. Macke, J.H.; Murray, I.; Latham, P.E. How biased are maximum entropy models? *Adv. Neur. Inf. Proc. Syst.* **2011**, *24*, 2034–2042.
27. Hutter, M.; Zaffalon, M. Distribution of mutual information from complete and incomplete data. *Comput. Stat. Data An.* **2005**, *48*, 633–657.
28. Jaynes, E.T. On the Rationale of Maximum-entropy methods. *P. IEEE* **1982**, *70*, 939–952.
29. 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.

30. Ebrahimi, N.; Soofi, E.S.; Soyer, R. Information Measures in Perspective. *Int. Stat. Rev.* **2010**, *78*, 383–412.
31. Wackernagel, H. *Multivariate Geostatistics—An Introduction with Applications*; Springer Verlag: Berlin, Germany, 1995.
32. Charpentier, A.; Fermanian, J.D. *Copulas: From Theory to Application in Finance*; Rank, J., Ed.; Risk Publications: London, UK, 2007; Section 2.
33. Tam, S.M. On Covariance in Finite Population Sampling. *J. Roy. Stat. Soc. D-Stat.* **1985**, *34*, 429–433.
34. Van der Vaart, A.W. *Asymptotic statistics*. Cambridge University Press: New York, NY, USA 1998; ISBN 978-0-521-49603-2, LCCN. V22 1998 QA276. V22.
35. Rockinger, M.; Jondeau, E. Entropy densities with an application to autoregressive conditional skewness and kurtosis. *J. Econometrics* **2002**, *106*, 119–142.
36. Bates, D. Quadratic Forms of Random Variables. STAT 849 lectures. Available online: <http://www.stat.wisc.edu/~st849-1/lectures/Ch02.pdf> (accessed on 22 February 2013).
37. Goebel, B.; Dawy, Z.; Hagenauer, J.; Mueller, J.C. An approximation to the distribution of finite sample size mutual information estimates. 2005. In Proceedings of IEEE International Conference on Communications (ICC' 05), Seoul, Korea, 16–20 May 2005; pp. 1102–1106.
38. Fisher, R.A. On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* **1921**, *1*, 3–32.
39. Zientek, L.R.; Thompson, B. Applying the bootstrap to the multivariate case: bootstrap component/factor analysis. *Behav. Res. Methods* **2007**, *39*, 318–325.
40. Mardia, K.V. Algorithm AS 84: Measures of multivariate skewness and kurtosis. *Appl. Stat.* **1975**, *24*, 262–265.
41. Hurrell, J.W.; Kushnir, Y.; Visbeck, M. The North Atlantic Oscillation. *Science* **2001**, *26*, 291.
42. The NCEP/NCAR Reanalysis Project. Available online: <http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml/> (accessed on 22 February 2013).

Appendix 1

Proof of Equations 1 and 2

We are looking for a PDF $\rho_{XY}(X, Y)$ satisfying: (1) the discrete constraints $\iint_S \mathbf{T}_{cr}(X, Y) \rho_{XY}^*(X, Y) dX dY = \boldsymbol{\theta}_{cr}$, corresponding to the vector $\boldsymbol{\eta}_{cr}$ of Lagrange multipliers and (2) the continuum of constraints $\iint_S \delta(X - u) \rho_{XY}(X, Y) dX dY = \rho_X(u)$ and $\iint_S \delta(Y - v) \rho_{XY}(X, Y) dX dY = \rho_Y(v)$, corresponding to the continuum of Lagrange multipliers $\lambda_X(u), \lambda_Y(v)$, $u \in S_X, v \in S_Y$, where the integrals of ρ_X, ρ_Y are both equal to one. The Lagrangian functional of Entropy is therefore

$$\begin{aligned} \mathcal{L}(\boldsymbol{\eta}_{cr}, \lambda_X, \lambda_Y) = & \iint_S \left[-\log \rho_{XY}(X, Y) + \lambda_X(X) + \lambda_Y(Y) + \boldsymbol{\eta}_{cr}^T \mathbf{T}_{cr}(X, Y) \right] \rho_{XY}(X, Y) dX dY \\ & - \int_{S_X} \rho_X(X) \lambda_X(X) dX - \int_{S_Y} \rho_Y(Y) \lambda_Y(Y) dY - \boldsymbol{\eta}_{cr}^T \boldsymbol{\theta}_{cr} \end{aligned} \quad (\text{A1})$$

The maximum Entropy is obtained by taking the differential $\delta\mathcal{L}$ of $\mathcal{L}$ in terms of $\delta\lambda_X(X)$, $\delta\lambda_Y(Y)$, $\delta\mathbf{\eta}_{cr}$ and setting vanishing gradient components, leading to the PDF $\rho_{XY}(X,Y) = \exp[-1 + \mathbf{\eta}_{cr}^T \mathbf{T}_{cr}(X,Y) + \lambda_X(X) + \lambda_Y(Y)]$. Now, considering the partition functions $Z_X(X, \mathbf{\eta}_{cr}) \equiv \exp[-\lambda_X(X)]$ and $Z_Y(Y, \mathbf{\eta}_{cr}) \equiv \exp[-\lambda_Y(Y)]$ and imposing the marginal PDF constraints leads directly to the expressions (2) where the continuum of Lagrange multipliers depend implicitly from the discrete ones $\mathbf{\eta}_{cr}$. Plugging that into $\mathcal{L}$ leads to the definition of the concave function $L(\mathbf{\eta}_{cr})$ in (1) with its global minimum at $\mathbf{\eta}_{cr} = \lambda_{cr}$. The MinMI-PDF (2) is $\rho_{XY}(X,Y) = \rho_{XY}^*(X,Y)$ at that minimum.

Proof of Equations 3, 4, 5 and Theorem 1

At the ME-PDF solution, the $\mathcal{L}$ functional of the MinMI solution is an implicit function of the constraining means $\mathbf{\theta}_{cr}$ and the differential satisfies $\delta\mathcal{L} = \delta H = -\delta I$. By expanding it in terms of $\delta\lambda_X(X)$, $\delta\lambda_Y(Y)$, $\delta\lambda_{cr}$, $\delta\mathbf{\theta}_{cr}$ and using $\int_{S_X} \rho_{XY}^*(X,Y) dY = \rho_X(X)$; $\int_{S_Y} \rho_{XY}^*(X,Y) dX = \rho_Y(Y)$, and $\iint_S \mathbf{T}_{cr}(X,Y) \rho_{XY}^*(X,Y) dX dY = \mathbf{\theta}_{cr}$, one gets $\delta I(\mathbf{\theta}_{cr}) = -\delta\mathcal{L} = \lambda_{cr}^T \delta\mathbf{\theta}_{cr}$, thus showing that the gradient of $I(\mathbf{\theta}_{cr})$ with respect to $\mathbf{\theta}_{cr}$ is λ_{cr} .

Regarding the Hessian of $I(\mathbf{\theta}_{cr})$, we must differentiate $\mathbf{\theta}_{cr}$ using the same technique for the ME problems with a finite number of constraints.

Therefore, as postulated in Section 2.2, let us consider a finite sequence of constraint sets $\{\mathbf{T}_j, \mathbf{\theta}_j\}$ whose ME-PDF converge to MinMI solution as $(j \rightarrow \infty)$. The differentials of expectations $\delta\mathbf{\theta}_j$ and the differential $\delta\lambda_j$ of Lagrange multipliers are related through $\delta\mathbf{\theta}_j = \mathbf{C}_{*j} \delta\lambda_j$, where $\mathbf{C}_{*j}$ is the covariance matrix of the constraining functions $\mathbf{T}_j$ at the ME-PDF solution (denoted with *), i.e., $\mathbf{C}_{*j} = E_*(\mathbf{T}_j' \mathbf{T}_j'^T)$ where the perturbations are $\mathbf{T}_j' = \mathbf{T}_j - \mathbf{\theta}_j$. Inverting that relationship we have $\delta\lambda_j = \mathbf{C}_{*j}^{-1} \delta\mathbf{\theta}_j$. In the case of MinMI, the constraining functions have a discrete part ($\mathbf{T}_{cr}$) and a continuous part (the Dirac deltas), being merged together into a whole vector $\mathbf{T}_{cr,\rho} = (\mathbf{T}_{cr}(X,Y), \delta(X-u), \delta(Y-v))^T$ corresponding to the whole vector of expectations $\mathbf{\theta}_{cr,\rho} = (\mathbf{\theta}_{cr}(X,Y), \rho_X(u), \rho_Y(v))^T$ and to the whole vector of Lagrange multipliers $\lambda_{cr,\rho} = (\lambda_{cr}, \lambda_X(u), \lambda_Y(v))^T$. Therefore, as for the discrete case, the differentials are related by $\delta\mathbf{\theta}_{cr,\rho} = E_*(\mathbf{T}_{cr,\rho}' \mathbf{T}_{cr,\rho}'^T) \delta\lambda_{cr,\rho} = \mathbf{C}_{cr,\rho} \delta\lambda_{cr,\rho}$, where the covariance matrix is now replaced by an operator (continuous matrix) along the u, v , and the discrete index of $\mathbf{\theta}_{cr}$. The multiplication of the continuous matrix by the continuous vector $\delta\lambda_{cr,\rho}$ is the sum of an integral in u , an integral in v and a discrete sum. The inverse relationship comes as $\delta\lambda_{cr,\rho} = [\mathbf{C}_{cr,\rho}]^{-1} \delta\mathbf{\theta}_{cr,\rho}$ where $[\mathbf{C}_{cr,\rho}]^{-1}$ is the inverse operator of $\mathbf{C}_{cr,\rho}$, i.e., the product $[\mathbf{C}_{cr,\rho}]^{-1} \mathbf{C}_{cr,\rho} = \mathbf{C}_{cr,\rho} [\mathbf{C}_{cr,\rho}]^{-1} = (\mathbf{I}_{cr}, \delta(X-u), \delta(Y-v))$ equals the identity operator. Therefore, the fixation of marginal PDFs in the MinMI problem leads to variations on cross expectations alone $\delta\mathbf{\theta}_{cr,\rho} = P_{cr} \delta\mathbf{\theta}_{cr,\rho} = \delta\mathbf{\theta}_{cr}$, where P_{cr} is the projection operator over the discrete part. Therefore, since $\delta I = \delta\mathbf{\theta}_{cr}^T \lambda_{cr}$, the second MI variation is $\delta^2 I = \frac{1}{2} \delta\mathbf{\theta}_{cr}^T \delta\lambda_{cr} = \frac{1}{2} \delta\mathbf{\theta}_{cr}^T [P_{cr}^T [\mathbf{C}_{cr,\rho}]^{-1} P_{cr}] \delta\mathbf{\theta}_{cr}$ and the matrix identity $P_{cr}^T [\mathbf{C}_{cr,\rho}]^{-1} P_{cr} = \mathbf{C}_{cr,\rho_X,\rho_Y}^{-1}$ appearing in (3). The discrete matrix $\mathbf{C}_{cr,\rho_X,\rho_Y}$ is positively

defined, being different from $P_{cr}^T [C_{cr,\rho}] P_{cr}$, which is the single covariance matrix of functions T_{cr} at the MinMI-PDF. Its computation is quite difficult in practice, involving the convolution (continuous product) of operators $[C_{cr,\rho}]^{-1}$ and P_{cr} .

Since the ME-PDF for $\{T_j, \theta_j\}$ converges to the MinMI PDF, the same holds for the covariance matrix conditioned to the marginal PDFs. Therefore, one has the Equation 10 at step j

$$(P_{cr} C_{*j}^{-1} P_{cr})^{-1} = (P_{cr} C_{*j} P_{cr}) - (P_{cr} C_{*j} P_{ind})(P_{ind} C_{*j} P_{ind})^{-1} (P_{ind} C_{*j} P_{cr}) = E_*[T_{cr,j}^{ind} T_{cr,j}^{ind T}] \xrightarrow{j \rightarrow \infty} C_{cr,\rho_X,\rho_Y} \quad (A2)$$

The matrix C_{cr,ρ_X,ρ_Y} can be obtained from the limit of ME covariance matrices where one adds progressively independent moments of the marginal variables X and Y as constraints. In the limit, the perturbations $T_{cr,j}^{ind} = T_{cr,j}' - \alpha_j^T T_{ind,j}'$ must converge to the perturbations $T^* = T_{cr} - E_{\rho_{X,Y}}^*(T_{cr} | \rho_X, \rho_Y)$ appearing in (4). They are residuals of the best fit on marginal functions on X and Y as $T^*(X, Y) = T_{cr}'(X, Y) - [\beta_X(X) + \beta_Y(Y)]$ where $\beta_X(X) + \beta_Y(Y)$ is a sum of marginal functions. The minimum of the total mean squares of residuals $\iint_S \rho_{XY}^*(X, Y) \|T^*\|^2 dX dY = E_*(\|T^*\|^2)$ is obtained through variational analysis by taking small variations $\delta\beta_X(X)$, $\delta\beta_Y(Y)$ and vanishing the gradients. We get the solution

$$T^*(X, Y) = T_{cr}'(X, Y) - [\alpha_X E(T_{cr}' | X) + \alpha_Y E(T_{cr}' | Y)] \quad (A3)$$

where fitting is done on conditional means and α_X, α_Y are the best linear fit coefficients for each function in $T_{cr}'(X, Y)$. This completes the proof of (5) and Theorem 1. The Taylor expansion (3) comes by taking $\Delta I(\theta_{cr}, \rho_X, \rho_Y) = \delta I + \delta^2 I + O(\|\Delta \theta_{cr}\|^3)$.

© 2013 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/>).