Dependence assessment based on generalized relative complexity measures: application to sampling network design

F. J. Alonso[1,*], M. C. Bueso[2] and J. M. Angulo[1]

[1] *Department of Statistics and O. R., University of Granada, Campus Fuente Nueva s/n, 18071 Granada, Spain; falonso@ugr.es, jmangulo@ugr.es*
[2] *Department of Applied Mathematics and Statistics, Technical University of Cartagena, Campus Muralla del Mar s/n, 30202 Cartagena, Murcia, Spain; mcarmen.bueso@upct.es*
[*] *Corresponding author*

Abstract. *The behavior of generalized relative complexity measures is studied for assessment of structural dependence in a random vector. A related optimality criterion to sampling network design, which provides a flexible extension of mutual information based methods previously introduced, is formulated. Aspects related to practical implementation and conceptual issues regarding the meaning and potential use of this new approach are discussed. Numerical examples are used for illustration.*

Keywords. *Generalized information measure; Rényi's entropy; Shannon's entropy; Spatial sampling; Statistical complexity.*

1 Introduction

In the last two decades, a variety of definitions of statistical complexity have been proposed in the literature (see, for example, [10]), with the aim to study structural characteristics of physical systems described by probability distributions. In particular, so-called 'product complexity measures' are formulated in terms of two factors providing, in a certain sense, a balanced quantification of inner information and relative structural richness of a given model. Such measures are usually required to satisfy the property to be zero for the two extreme cases of perfect order and maximum disorder. In this framework, the LMC complexity measure (after López-Ruiz, Mancini and Calbet, [7]) is defined by the product of a measure of information and a measure of disequilibrium; the exponential form, usually adopted for continuous distributions, is written for a density f as

$$C_f^{LMC} = \exp(H_f) \times D_f,$$

where H_f is the Shannon entropy and D_f is the second-order entropic moment given by $D_f = \int f^2(x)dx$, which quantifies departure from equilibrium. Other definitions of complexity measures have been introduced considering generalizations of the Shannon entropy, such as the Rényi entropy, defined as

$$
\begin{aligned}
R_f^{(\alpha)} &:= \frac{1}{1-\alpha}\log\left(\int f^{\alpha}(x)dx\right), \quad \text{for } 0<\alpha<\infty, \quad \alpha\neq 1,\\
R_f^{(1)} &:= \lim_{\alpha\to 1} R_f^{(\alpha)} = H_f.
\end{aligned}
$$

In particular, the second-order disequilibrium factor D_f can be also expressed in terms of the Rényi entropy as $D_f = \exp\left(-R_f^{(2)}\right)$, and hence $C_f^{LMC} = \exp\left(R_f^{(1)} - R_f^{(2)}\right)$. Following the same scheme, for the local comparison of two probability distributions, [9] introduced a family of generalized relative complexity measures depending on two parameters α and β, defined by

$$
C_{f,g}^{(\alpha,\beta)} = \exp\left(R_{f,g}^{(\alpha)} - R_{f,g}^{(\beta)}\right), \quad 0<\alpha,\beta<\infty,
$$

where $R_{f,g}^{(\alpha)}$ is the relative Rényi entropy of order α given by

$$
\begin{aligned}
R_{f,g}^{(\alpha)} &:= \frac{1}{\alpha-1}\log\left(\int \frac{f^{\alpha}(x)}{g^{\alpha-1}(x)}dx\right), \quad \text{for } 0<\alpha<\infty, \quad \alpha\neq 1,\\
R_{f,g}^{(1)} &:= \lim_{\alpha\to 1} R_{f,g}^{(\alpha)} = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx = KL(f,g),
\end{aligned}
$$

with $KL(f,g)$ being the relative entropy or Kullback-Leibler divergence. Among other elementary properties, we note that for $\alpha<\beta$, $C_{f,g}^{(\alpha,\beta)} \leq 1$, and $C_{f,g}^{(\alpha,\beta)} = 1/C_{f,g}^{(\beta,\alpha)}$.

In this work we consider this definition of generalized relative complexity (other alternatives of interest are referred in concluding Section 3). We first study its behaviour in the case of comparison of a multivariate distribution and the product of its marginal distributions, for dependence assessment. This is a natural extension of the concept of mutual information in the complexity context. In Section 2, we propose a complexity related criterion for sampling network design based on this approach.

In order to analyze the behavior of relative complexity for different structures of dependence and selections of the parameters (α,β), we have considered the case of a bivariate random vector (X,Y) with different specifications for marginal distributions and dependence. We show here results for a bivariate Gaussian with standarized marginal distributions, taking $\beta=2$ and $\beta=6$ (see Figure 1). In these two scenarios, α is increased from 0.5 to $\beta-0.05$ and the correlation between the variables is varied within the range $[-0.96,0.96]$. The value of the relative complexity is equal for positive or negative correlation, and it only depends on its magnitude. It can be observed that if α is close to β, the relative complexity does not decrease as strongly as for more distant specifications of the parameters (α,β). For instance, in the case $(\alpha,\beta)=(1.95,2)$ the relative complexity is close to 0.9 for high correlation, and in the case $(\alpha,\beta)=(5.95,6)$ is over 0.98. In this last case, the complexity decreases for correlations near to zero and then it remains almost constant. In the opposite case, when the distance between α and β is significant, the relative complexity decreases more drastically as the correlation increases. However, it can be noted that in the case $(\alpha,\beta)=(0.5,2)$ the relative complexity decreases with a parabola-like shape, while in the case $(\alpha,\beta)=(0.5,6)$ it decreases with a bell-like shape.

2 Application to sampling network design

In previous papers the authors have proposed optimality entropy-based criteria for designing sampling networks based on maximizing the Shannon mutual information in different contexts (e.g. [1], [2], [3],

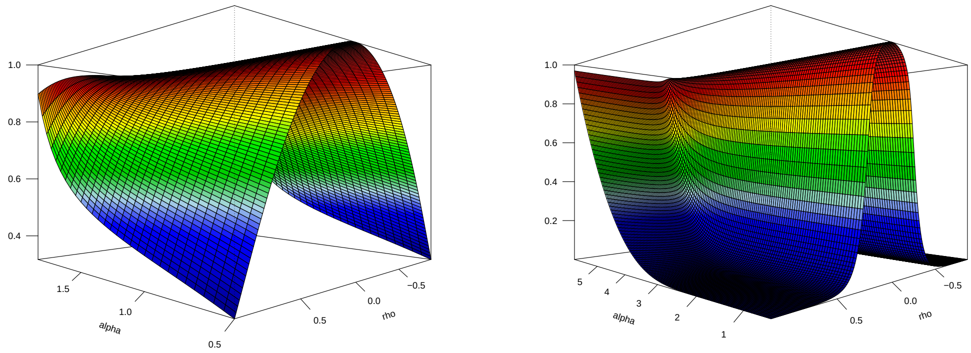

Figure 1: Generalized relative complexity values for $\beta = 2$ (left) and $\beta = 6$ (right), varying α and ρ.

[5]). Here, we extend this approach by formulating an optimality criterion based on the generalized relative complexity. For given values of α and β, with $\alpha < \beta$, the relative complexity of the joint distribution of the variables involved in the sampling design problem, i.e. the variables of interest and the observable variables, and the product of the marginal distributions corresponding to the case of independence between the two types of variables, is evaluated. Denote by X the unobservable spatial process of interest in the set Λ, related to the spatial process Y potentially observable on the set Π. The criterion consists of selecting a configuration $S \subseteq \Pi$ to observe the process Y for estimating X in Λ, that minimizes the relative complexity of the joint distribution of X_Λ and Y_S and the product of the corresponding marginal distributions:

$$S^* = \arg \min_{S \in \mathcal{C}(\Pi)} C_{f,g}^{(\alpha,\beta)},$$

where f is the joint density function of X_Λ and Y_S, g is the product of the corresponding marginal density functions, and $\mathcal{C}(\Pi)$ is the class of admissible candidate configurations considered in the specific case. For illustration, we assume that X is defined on the domain $D = [0,1]^2 \subset \mathbb{R}^2$ and is related to Y by the observation equation $Y(\mathbf{s}) = X(\mathbf{s}) + \varepsilon(\mathbf{s}), \mathbf{s} \in D$, where ε is a spatial white noise process with variance σ_ε^2, independent of X. We also assume a joint Gaussian distribution for all the variables involved and that X has zero mean and an exponential covariance model with range a. All the parameters are assumed to be known and a sequential inclusion of the sites is performed. Specifically, we set $a = 5$ and $\sigma_\varepsilon^2 = 0.1$ and, for simplicity in the computation, $\Lambda = \{(0.5, 0.5)\}$ and $\Pi = \{(0.6883, 0.3614), (0.6987, 0.8740), (0.4771, 0.6699), (0.2457, 0.1345), (0.6587, 0.4670)\}$. Table 1 displays the results from the two first sampling design steps by considering the relative complexity criterion for the cases $(\alpha, \beta) = (1,2)$ and $(3,4)$. For comparison, we also include the selected sites by using the Shannon mutual information criterion. We can observe that the selected sites for the values $(\alpha, \beta) = (1,2)$ coincide with those selected by the Shannon-based criterion, while the ordering changes for the values $(\alpha, \beta) = (3,4)$, which shows that the combined specification of the values of the deformation parameters α and β is significant for the selection of the optimal configuration for a given model.

3 Conclusion

The behavior of Rényi-divergence-based generalized relative complexity is studied in the case of comparison of a multivariate distribution and the product of its marginal distributions, for dependence assessment. A related criterion for sampling network design is formulated. Numerical examples are given for

s	Shannon's information	**s**	$C_{f,g}^{(1,2)}$	**s**	$C_{f,g}^{(3,4)}$
(0.6883,0.3614)	0.5803	(0.6883,0.3614)	0.5744	(0.6883,0.3614)	0.07699
(0.6987,0.8740)	0.3958	(0.6987,0.8740)	0.6767	**(0.6987,0.8740)**	0.07640
(0.4771,0.6699)	0.6736	(0.4771,0.6699)	0.5337	(0.4771,0.6699)	0.07801
(0.2457,0.1345)	0.3805	(0.2457,0.1345)	0.6865	(0.2457,0.1345)	0.07661
(0.6587,0.4670)	0.6899	**(0.6587,0.4670)**	0.5273	(0.6587,0.4670)	0.07819
(0.6883,0.3614)	0.7555	(0.6883,0.3614)	0.5045	(0.6883,0.3614)	0.06535
(0.6987,0.8740)	0.7562	(0.6987,0.8740)	0.5029	(0.4771,0.6699)	0.07286
(0.4771,0.6699)	0.9163	**(0.4771,0.6699)**	0.4565	**(0.2457,0.1345)**	0.05859
(0.2457,0.1345)	0.7867	(0.2457,0.1345)	0.4922	(0.6587,0.4670)	0.06996

Table 1: Sequentially selected sites (marked in boldface) at the first two sampling design steps using the Shannon mutual information, and $C_{f,g}^{(1,2)}$ and $C_{f,g}^{(3,4)}$ relative complexities.

illustration under different scenarios. Further extensions are under consideration in this context, in terms of other generalized relative complexity measures, such as the Rényi-based two-folded relative complexity version introduced in [4], as well as complexity measures derived from alternative formulations of Rényi and Tsallis mutual information (see, for example, [6], [8]).

Acknowledgments. This work has been partially supported by Spanish grant MTM2012-32666 of Ministerio de Economía y Competitividad (co-financed by FEDER).

References

[1] Angulo, J. M., Bueso, M. C., Alonso, F. J. (2000). A study on sampling design for optimal prediction of space-time stochastic processes. *Stochastic Environmental Research and Risk Assessment* **14**, 412–427.

[2] Angulo, J. M., Bueso, M. C. (2001). Random perturbation methods applied to multivariate spatial sampling design. *Environmetrics* **12**, 631–646.

[3] Angulo, J. M., Bueso, M. C., Alonso, F. J. (2013). Space-time adaptive sampling and data transformations. In *Spatial-Temporal Designs. Advances in Efficient Data Acquisition* (J. Mateu and W. G. Müller, eds.), Ch. 10, pp. 231-248. Wiley. Chichester.

[4] Bouvrie, P. A., Angulo, J. C., Antolín, J. (2012). A generalized relative complexity: Application to atomic one-particle densities. *Chemical Physics Letters* **539-540**, 191-196.

[5] Bueso, M. C., Angulo, J. M., Alonso, F. J. (1998). A state-space-model approach to optimal spatial sampling design based on entropy. *Environmental and Ecological Statistics* **5**, 29–44.

[6] Furuichi, S. (2006). Information theoretical properties of Tsallis entropies. *Journal of Mathematical Physics* **47**, 023302.

[7] López-Ruiz, R., Nagy, Á., Romera, E., Sañudo, J. (2009). A generalized statistical complexity measure: applications to quantum systems. *Journal of Mathematical Physics* **50**, 123528.

[8] Rényi, A. (1970). *Probability Theory*. North-Holland Publishing Company. Amsterdam.

[9] Romera, E., Sen, K. D., Nagy, Á. (2011). A generalized relative complexity measure. *Journal of Statistical Mechanics: Theory and Experiment* **09**, P09016.

[10] Sen, K. D. (2011). *Statistical Complexity: Applications in Electronic Structure*. Springer. Dordrecht.