On Fisher Information and the Spatial Dependence Structure of Isotropic Pairwise Gaussian-Markov Random Field Models

Alexandre L. M. Levada

Universidade Federal de São Carlos, SP, Brazil

(Dated: June 12, 2013)

Abstract

Markov Random Field (MRF) models are powerful tools for contextual modelling in the study of complex systems. However, little is known about how the spatial dependence between their elements is encoded in terms of information-theoretic measures. In this paper, we enlight the connection between Fisher information, Shannon entropy and spatial properties of the random field in case of Gaussian random variables (a Gaussian Markov random field, or GMRF), by defining analytical expressions to compute local and global versions of these measures using Besag’s pseudo-likelihood function (conditional independence assumption). The proposed expressions provide analytical tools for the analysis of contextual patterns in data, allowing among other things, the detection of the most informative ones. Besides, we show that from a statistical inference perspective these measures are directly related to the uncertainty in the estimation of the global system behavior by means of the asymptotic variance of the maximum pseudo-likelihood estimator of the spatial dependence parameter. Moreover, the results indicate that the accuracy on the estimation of the global behavior on GMRF’s (inverse temperature) depends essentially on two quite intuitive conditions: concentration of patterns with high local log-likelihood value (minimization of type-I Fisher information), which means patterns “aligned” to the expected global behavior, and also concentration of patterns showing high local log-likelihood curvature (maximization of type-II Fisher information), which means that small perturbations on data cannot cause abrupt changes on the system’s spatial dependence structure (stability). Inspired by these findings we defined $L$-information, a measure calculated by the ratio of the first two derivatives of the log-pseudo-likelihood function regarding the spatial dependence parameter on a MRF. Computational experiments with both simulated and real data show the effectiveness of the proposed equations in practical applications involving spatially dependent random variables.
INTRODUCTION

With the increasing value of information in the modern society and the massive volume of digital data that is available nowadays, there is an urgent need of developing novel methodologies for knowledge discovery from the vast sets of symbols found in several domains of science. In this scenario, where the focus is on data itself, complex systems arise as a powerful framework for a new model of data analysis and interpretation. Patterns that at first may appear to be locally irrelevant may turn out to be extremely informative in a more global perspective. Hence, the concept of what is an informative pattern is a top priority in dealing with a complex system, where often its components (elements, patterns and clusters) relate in a non-linear fashion making the whole greater than the sum of its parts, since there is relevant information encoded in these complex relationships.

Within this context, information theoretic measures play a fundamental role in a huge variety of applications once they represent statistical knowledge in a systematic, elegant and formal framework. Since the first works of Shannon [1], and later with many other generalizations [2–4], the concept of entropy has been adapted and successfully applied to almost every field of science, among which we can cite physics [5], mathematics [6–8], economics [9] and fundamentally, information theory [10–12]. Similarly, the concept of Fisher information has been shown to reveal important properties of statistical procedures, from lower bounds on estimation methods [13–15] to information geometry [16, 17]. Basically, Fisher information can be thought as the likelihood analog of entropy, which is a probability-based measure of uncertainty.

In general, classical statistical inference is focused on capturing information about location and dispersion of unknown parameters of a given family of distribution and studying how this information is related to uncertainty in estimation procedures. In typical situations, exponential family of distributions and independence hypothesis (independent random variables) are often assumed, giving the likelihood function a series of desirable mathematical properties [13–15].

Although mathematically convenient, in many applications such as image processing, contextual data mining, and complex systems analysis, independence assumption may not be reasonable [18, 19], because much of the information is somehow encoded in the relations between the random variables. In this scenario, Markov Random Field (MRF) models
appear as a natural generalization of the classical model by the simple replacement of the independence assumption by the conditional independence assumption. Roughly speaking, in every MRF, knowledge of a finite-support neighborhood around a given variable isolates it from all the remaining variables. A further simplification is to consider a pairwise interaction model, constraining the size of the maximum clique to be two. Moreover, if the MRF model is isotropic, which means the spatial dependence parameter is not dependent on the directions, all information regarding the spatial dependence structure of the system is conveyed by a single parameter, from now on denoted by \( \beta \) (also known as inverse temperature).

In this paper, we assume an isotropic pairwise Gaussian Markov Random Field (GMRF) model \([20, 21]\) (also known as auto-normal model or conditional auto-regressive model \([22, 23]\)). Basically, the question that motivated this work and we are trying to elucidate here is: What kind of information is encoded by the \( \beta \) parameter in such a model? We want to know how this parameter, and consequently the whole spatial dependence structure of a complex system modelled by a Gauss Markov random field, is related to both local and global information theoretic measures, more precisely the observed and expected Fisher information as well as self-information and Shannon entropy.

In searching for answers for our fundamental question, investigations led us to an exact expression for the asymptotic variance of the maximum pseudo-likelihood (MPL) estimator of the spatial dependence parameter on a pairwise GMRF model as a ratio of two forms of Fisher information, indicating that asymptotic efficiency is not granted, once information equality fails. An approximation for the asymptotic variance of the spatial dependence parameter using the observed Fisher information has been proposed in \([24]\). Here, however, we use the expected Fisher information as it appears on the Cramer-Rao lower bound. To the best of our knowledge, closed expressions for the expected Fisher information in the GMRF model have not been derived before. In the context of statistical data analysis, Fisher information plays a central role in providing tools and insights for modeling the interactions between complex systems and their components. The advantage of MRF models over the traditional statistical ones is that MRF’s take into account the dependence between pieces of information as a function of the parameter that controls the temperature of the system, which may be variable through time. Briefly speaking, this research aims to explore ways to measure and quantify distances between complex systems in different thermodynamical conditions by analyzing and comparing the behavior of local patterns that can be observed
in a regular 2D lattice and by determining how informative these patterns are regarding that particular inverse temperature ($\beta$).

The remaining of the paper is organized as follows: Section 2 discusses maximum pseudo-likelihood (MPL) estimation and provides derivations for the expected Fisher information regarding the spatial dependence parameter $\beta$ using both first and second derivatives of the pseudo-likelihood function on a pairwise isotropic GMRF model. Intuitive interpretations for these two measures are also discussed. In Section 3 we show an expression for the global entropy on a GMRF model (a probability-based counterpart to Fisher information), given by the expected value of self-information, a local uncertainty measure based on the observation of a contextual configuration pattern defined by a Markovian neighborhood. The results suggest a connection between maximum pseudo-likelihood and minimum entropy conditions on a MRF. Section 4 presents an exact expression for the asymptotic variance of the MPL estimator of $\beta$ as a ratio of both forms of Fisher information, showing that accuracy on the estimation of $\beta$ depends essentially on the massive presence of contextual patterns satisfying two intuitive conditions: high local likelihood (minimization of one form of Fisher information) and stability, that is, the local log-likelihood is not flat, which means that small variations on data cannot cause abrupt changes on the spatial dependence structure (maximization of another form of Fisher information). Section 6 shows some illustrative computational experiments using Markov Chain Monte Carlo simulations and some results on detecting and extracting relevant information from GMRF outcomes, represented here by noisy images. Finally, Section 7 presents the conclusions, final remarks and possibilities for future works.

**FISHER INFORMATION ON PAIRWISE GMRF’S**

The remarkable Hammersley-Clifford theorem [25] states the equivalence between Gibbs Random Fields (GRF) and Markov Random Fields (MRF), which implies that any MRF can be defined either in terms of a global (joint Gibbs distribution) or a local (set of local conditional density functions) model. For our purposes, we will choose the later representation.

Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ be a set of gaussian random variables defined on a rectangular lattice and $\eta$ a non-causal neighborhood system. A GMRF is completely characterized by a
set of $n$ (number of variables) local conditional density functions (LCDF’s), given by [26]:

$$p(x_i|\eta_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} \left[ x_i - \mu - \sum_{x_j \in \eta_i} \beta (x_j - \mu) \right]^2 \right\}$$

(1)

where $\theta = (\mu, \sigma^2, \beta)$ is the vector of parameters, with $\mu$ and $\sigma^2$ denoting the mean (expected value) and variance, $\beta$ denoting the spatial dependence parameter (inverse temperature) and $\eta_i$ representing the neighborhood around the $i$-th random variable in the field. It is interesting to note that for $\beta = 0$, the expression degenerates to a Gaussian density. From an information geometry perspective [16, 17], it means we are constrained to a sub-manifold within the Riemmanian manifold of probability distributions, where the natural Riemmanian metric (tensor) is given by the Fisher information. It has been shown that the geometric structure of exponential family distributions exhibit constant curvature. However, little is known about information geometry on more general statistical models, such as Gaussian Markov Random Fields.

**Maximum Pseudo-Likelihood Estimation**

Maximum likelihood estimation is intractable in MRF parameter estimation due to the existence of the partition function in the joint Gibbs distribution. An alternative, proposed by Besag [22], is maximum pseudo-likelihood estimation, which is based on the conditional independence principle. The pseudo-likelihood function is defined as the product of the LCDF’s for all the $n$ elements of the complex system, modelled as a random field. So, for a GMRF model, the log pseudo-likelihood function is defined by the following:

$$PL(\theta|X) = -\frac{n}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[ x_i - \mu - \beta \sum_{x_j \in \eta_i} (x_j - \mu) \right]^2$$

(2)

By differentiating equation (2) with respect to each parameter and properly solving the pseudo-likelihood equations one obtains the following MPL estimators:
\[
\hat{\beta}_{MPL} = \frac{\sum_{i=1}^{n} (x_i - \mu) \sum_{j \in \eta_i} (x_j - \mu)}{\sum_{i=1}^{n} \left( \sum_{j \in \eta_i} (x_j - \mu) \right)^2}
\] (3)

\[
\hat{\mu}_{MPL} = \frac{1}{n(1-k\beta)} \sum_{i=1}^{n} \left( x_i - \beta \sum_{j \in \eta_i} x_j \right)
\] (4)

\[
\hat{\sigma}^2_{MPL} = \frac{1}{n} \sum_{i=1}^{n} \left[ x_i - \mu - \beta \sum_{j \in \eta_i} (x_j - \mu) \right]^2
\] (5)

where \( k \) denotes the cardinality of the non-causal neighborhood set \( \eta_i \). Furthermore, assuming a regular neighborhood system such as the 2D lattice (each variable is dependent on a fixed number of neighboring variables), we can rewrite \( \hat{\beta}_{MPL} \) as:

\[
\hat{\beta}_{MPL} = \sum_{j \in \eta_i} \hat{\sigma}_{ij} \sum_{j \in \eta_i} \sum_{k \in \eta_i} \hat{\sigma}_{jk}
\] (6)

where \( \sigma_{ij} \) denotes the covariance between the central variable \( x_i \) and a variable \( x_j \) belonging to a neighborhood system (typical choices are first or second order systems, which correspond to the four and eight nearest neighbors on a rectangular bidimensional lattice) and \( \sigma_{jk} \) denotes the covariance between two variables belonging to the neighborhood of \( x_i \). Note that if \( \beta = 0 \), the MPL estimators of both \( \mu \) and \( \sigma^2 \) become the widely known sample mean and sample variance.

**Fisher information of spatial dependence parameters**

Basically, Fisher information measures the amount of information the observation of a random variable conveys about an unknown parameter. It can be thought as the likelihood analog of entropy, which is a probability-based measure of uncertainty. Often, when we are
dealing with independent and identically distributed (i.i.d) random variables, the computation of the global Fisher Information presented in a random sample $X = \{x_1, x_2, \ldots, x_n\}$ is quite straightforward, since each observation $x_i$, $i = 1, 2, \ldots, n$, brings exactly the same amount of information. However, this is not true for spatial dependence parameters in MRF’s, since different configuration patterns provide distinct contributions to the local observed Fisher information, which can be used to derive a reasonable approximation to the global Fisher information [27].

**Observed Fisher information**

Considering a complex system modelled by a GMRF defined by a set of LCDF’s (eq. [1]), the observed Fisher information can be calculated in terms of the pseudo-likelihood equation as (we will recall this measure type-I observed Fisher information):

$$I_{\text{obs}}^{(1)}(\beta) = \left[ \frac{\partial}{\partial \beta} \log P L(\beta) \right]^2$$

and it can be estimated by a sample average, justified by the Law of Large Numbers:

$$\phi_\beta = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{\partial}{\partial \beta} \log p(x_i | \eta_i, \beta) \right]^2_{\beta = \hat{\beta}}$$

(8)

where $p(x_i | \eta_i, \theta)$ is the LCDF of the Markovian model. Thus, $\phi_\beta$ is an unbiased estimator of the observed Fisher information, that is, $I_{\text{obs}}^{(1)}(\beta) = E[\phi_\beta]$, making $\phi_\beta$ a good approximation to $I_{\text{obs}}^{(1)}(\beta)$. Replacing equation [1] in [3] and after some manipulations, a closed expression for the type-I observed Fisher information, $\phi_\beta$, in the pairwise GMRF model is given by the following:

$$\phi_\beta = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{\sigma^2} \left[ x_i - \mu - \sum_{x_j \in \eta_i} \beta (x_j - \mu) \right] \left[ \sum_{x_j \in \eta_i} (x_j - \mu) \right] \right\}^2$$

(9)

Note that $\phi_\beta$ is simply an average of the local observed Fisher information along the random field. Thus, we can think of $\phi_\beta(x_i)$, for $i = 1, 2, \ldots, n$, as being the information that a particular contextual pattern provides as contribution to the global observed Fisher information. In this sense, the observed Fisher information is explicitly defined in terms of local measures. Note the similarity between $\phi_\beta(x_i)$ and self-information, $-\log p(x_i)$. 

7
Basically, the main difference is that while the former is based on the likelihood, the latter is based on the probability.

Alternatively, one can compute the observed Fisher information by the negative of the second derivative (we will recall this measure type-II global Fisher information):

\[ I^{(2)}_{\text{obs}}(\beta) = -\frac{\partial^2}{\partial \beta^2} \log PL(\beta) \]  

which, after some basic algebra, results in the following approximation:

\[ \psi(\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma^2} \left[ \sum_{x_j \in \eta_i} (x_j - \mu) \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{\sigma^2} \sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} (x_j - \mu) (x_k - \mu) \right] \]  

Note that while \( \phi(\beta) \) is a function of \( \beta \), \( \psi(\beta) \) does not depend explicitly on the spatial dependence parameter. Once again, \( \psi(\beta) \) is the average of another local Fisher information measure, \( \psi(\beta)(x_i) \), along the entire random field.

Therefore, with these two local measures, \( \phi(\beta)(x_i) \) and \( \psi(\beta)(x_i) \), we can assign two information values to every element of a complex system modelled by a MRF. Note that the same measures can be derived for other models, such as the Potts model [28].

A relevant question concerns the interpretation of these local information values. Roughly speaking, \( \phi(\beta)(x_i) \) is the quadratic rate of change of the local likelihood, which means that observations showing low values of \( \phi(\beta)(x_i) \) are very likely to occur throughout the field (they are close to the maximum of the local likelihood). In other words, these patterns or patches are "aligned" to the expected global behavior (once they are not informative). On the other hand, observations showing high values of \( \phi(\beta)(x_i) \) are typically landmarks, because they bring a significant amount of information about the global spatial dependence structure, since they are not likely to occur for that particular value of \( \beta \) (informative patterns or patches). Therefore, type-I local Fisher information minimization in MRF’s produces configuration patterns or patches that are more likely to occur. Basically, \( \phi(\beta)(x_i) \) tell us how informative a given pattern is for a specific global system behavior (given by the value of the inverse temperature parameter \( \beta \)). In other words, \( \Phi(\beta) \) measures the degree of agreement between the central element of the pattern and its neighbors, given a neighborhood system.

Recent works in the signal and image processing literature have been investigating the use of this criterion in Gaussian denoising by considering a tradeoff between local mean square error (MSE) minimization and local observed Fisher information minimization, showing
FIG. 1. Simple geometrical interpretation for local Fisher information minimization: by changing the value of a variable in order to minimize its local Fisher information, a more likely configuration pattern for a complex system in that specific inverse temperature raises.

promissing results [29, 30]. Figure [1] depicts a simple geometrical illustration of the type-I local Fisher information minimization process, showing that patterns exhibiting low values of $\phi_\beta(x_i)$ are close to a maximum of the local likelihood.

Regarding the type-II local Fisher information $\psi_\beta(x_i)$, informally speaking, it can be interpreted as a curvature measure. Thus, points showing low values of this measure have a nearly flat local likelihood (all possible values of $\beta$ have approximately the same likelihood), which means that small perturbations on a significant number of points exhibiting low values of this measure can cause a sharp change on $\beta$ parameter, and consequently in the global spatial dependence structure (in other works, this patterns are more susceptible to changes since they do not show a “stable” configuration). On the other side, if we have many patterns exhibiting large values of $\psi_\beta(x_i)$, change on the global spatial structure is unlikely to happen. In other words, $\Psi_\beta$ measures the degree of agreement among the neighboring elements of the pattern themselves. These rather informal arguments defines the basis for understanding the meaning of the asymptotic variance of maximum pseudo-likelihood estimators, as it will be discussed in the next Sections. Basically, $\psi_\beta(x_i)$ is a measure of how sure or confident we are about the local spatial dependence structure (at a given point $x_i$), since a high average curvature is desired for predicting the system’s global spatial dependence structure (better accuracy of $\beta$ MPL estimator).
**Expected Fisher information**

Unlike the observed Fisher information, the expected Fisher information is strictly a global measure. It is defined by the expected value of the squared score function. In our approach, we replace the likelihood function (the intractable joint Gibbs distribution) by the pseudo-likelihood function in the definition of the score function:

\[
\Phi_\beta = E \left[ \left( \frac{\partial}{\partial \beta} \log PL(\beta) \right)^2 \right]
\]

or equivalently, in classical inference (exponential family i.i.d random variables), by taking the expectation of the negative of the second derivative, which can be interpreted as being an average curvature:

\[
\Psi_\beta = -E \left[ \frac{\partial^2}{\partial \beta^2} \log PL(\beta) \right]
\]

In the following, closed-form expressions for both \( \Phi_\beta \) and \( \Psi_\beta \) on the GMRF model are presented, showing that, in general, information equality \( \Phi_\beta = \Psi_\beta \) fails. Plugging equation (2) in (12) and after some algebra, we obtain the following expression, which is composed by three main terms:

\[
\Phi_\beta = \frac{1}{\sigma^4} \sum_{i=1}^{n} E \left\{ \left[ (x_i - \mu - \beta \sum_{x_j \in \eta_i} (x_j - \mu)) \left( \sum_{x_j \in \eta_i} (x_j - \mu) \right) \right]^2 \right\} =
\]

\[
= \frac{1}{\sigma^4} \sum_{i=1}^{n} E \left\{ \left[ (x_i - \mu) \sum_{x_j \in \eta_i} (x_j - \mu) - \beta \left( \sum_{x_j \in \eta_i} (x_j - \mu) \right)^2 \right]^2 \right\} =
\]

\[
= \frac{1}{\sigma^4} \sum_{i=1}^{n} E \left\{ \left[ \sum_{x_j \in \eta_i} (x_i - \mu) (x_j - \mu) \right]^2 - 2\beta \left[ \sum_{x_j \in \eta_i} (x_i - \mu) (x_j - \mu) \right] \left[ \sum_{x_j \in \eta_i} (x_j - \mu) \right]^2 \right. 
+ \left. \beta^2 \left[ \sum_{x_j \in \eta_i} (x_j - \mu) \right]^4 \right\}
\]
Expanding the first term of the previous expression gives us:

\[
E \left\{ \left[ \sum_{x_j \in \eta_i} (x_i - \mu) (x_j - \mu) \right]^2 \right\} = E \left[ \sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} (x_i - \mu)^2 (x_j - \mu) (x_k - \mu) \right] = \\
= \sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} E [(x_i - \mu) (x_i - \mu) (x_j - \mu) (x_k - \mu)] \tag{15}
\]

According to the Isserlis' theorem \cite{31}, for normally distributed random variables, we can compute higher order moments in terms of the covariance matrix through the following identity:

\[
E [X_1 X_2 X_3 X_4] = E [X_1 X_2] E [X_3 X_4] + E [X_1 X_3] E [X_2 X_4] + E [X_2 X_3] E [X_1 X_4] \tag{16}
\]

which finally lead us to:

\[
\sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} E [(x_i - \mu) (x_i - \mu) (x_j - \mu) (x_k - \mu)] = \tag{17}
\]

\[
\sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} \left\{ E [(x_i - \mu)^2] E [(x_j - \mu) (x_k - \mu)] + 2E [(x_i - \mu) (x_j - \mu)] E [(x_i - \mu) (x_k - \mu)] \right\} = \\
\sum_{j \in \eta_i} \sum_{k \in \eta_i} [\sigma^2 \sigma_{jk} + 2\sigma_{ij} \sigma_{ik}]
\]

We now proceed to the expansion of the second main term of (14). By expanding the square, multiplying the summations and using the identity in (16), we have:
Finally, the third term of (14) is given by:

\[
E \left\{ \left[ \sum_{x_j \in \eta_i} (x_i - \mu) (x_j - \mu) \right]^4 \right\} = E \left\{ \left[ \sum_{x_j \in \eta_i} (x_i - \mu) \right]^2 \left[ \sum_{x_j \in \eta_i} (x_i - \mu) \right]^2 \right\} = (19)
\]

\[
= \sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} \sum_{x_l \in \eta_i} \sum_{x_m \in \eta_i} E \left[ (x_i - \mu) (x_j - \mu) (x_k - \mu) (x_l - \mu) \right] =
\]

\[
= \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sum_{l \in \eta_i} \sum_{m \in \eta_i} \left[ \sigma_{jk} \sigma_{lm} + \sigma_{jl} \sigma_{km} + \sigma_{jm} \sigma_{kl} \right]
\]

Therefore, combining all the parts, the complete expression for \( \Phi_\beta \) (Fisher information for the pairwise GMRF spatial dependence parameter given by the square of the score function) is given by the following equation:

\[
\Phi_\beta = \frac{1}{\sigma^4} \sum_{i=1}^{n} \left\{ \sum_{j \in \eta_i} \sum_{k \in \eta_i} \left[ \sigma^2 \sigma_{jk} + 2 \sigma_{ij} \sigma_{ik} \right] - 2 \beta \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sum_{l \in \eta_i} \sum_{m \in \eta_i} \left[ \sigma_{ij} \sigma_{kl} + \sigma_{ik} \sigma_{jl} + \sigma_{il} \sigma_{jk} \right] \right\}
\] (20)
At this moment some observations can be pointed out. First, the summation involving the variable \( i \) may be interpreted in two ways: in a temporal aspect, when we are observing the outcomes of a random pattern along time (a fixed position of the lattice is assumed to produce several outcomes); or in a spatial aspect, when the patterns are arranged in a 2D lattice, where a moving window is used to extract the observed patterns from the entire lattice at a given time. Thus, the proposed expression allows us to model and analyze different kinds of data: spatial, temporal or even spatiotemporal. In case of static observations, like an image (or possibly a random regular graph) in which each vertex represents a Gaussian random variable, all possible observed patterns extracted with a moving window are treated as outcomes of a single random variable at a fixed time instant (in this case we are interested only in studying spatial correlations). In this case, the summation over \( n \) in the previous equation vanish, since we have \( n = 1 \). From now on, we will consider this specific situation where all the patterns in a given time instant \( t \) are outcomes of a single random variable, in order to estimate the spatial correlations \( \sigma_{ij} \) and \( \sigma_{kl} \), given a neighborhood system.

Note also that in this situation all information concerning the system’s global dependence structure (i.e., the information conveyed by \( \beta \), the inverse temperature) at a time instant \( t \), can be extracted directly from the covariance matrix of the contextual configuration patterns defined by the neighborhood system (as if each pattern were a vector sampled from a multivariate Gaussian density). In this sense, the proposed expression shows an explicit connection between the covariance matrix of the patterns and the system’s global behavior or inverse temperature (in other words, in a GMRF model, temperature is a quantity that is dependent only on second order statistics, as it would be expected).

Thus, to make computations easier, it is possible to rewrite equation (20) using Kronecker products, vectors and submatrices extracted from the covariance matrix of the contextual patterns, from now on denoted by \( \Sigma_p \). For example, considering a given neighborhood system (K-nearest-neighbors), the general form of \( \Sigma_p \) is given by:
\[ \Sigma_p = \begin{pmatrix}
\sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,K} \\
\sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,K} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{K,1} & \sigma_{K,2} & \cdots & \sigma_{K,K}
\end{pmatrix} \]  

(21)

Denoting by \(\Sigma_p^-\) the submatrix of dimensions \((K - 1) \times (K - 1)\) obtained by deleting the central row and central column (covariances between the central element and each one of the neighbors) and by \(\vec{\rho}\) the vector of dimensions \((K - 1) \times 1\) obtained by the elements of the central row (or column) of \(\Sigma_p\), excluding the central one (variance of the central variable of the pattern), equation (20) becomes (for \(n = 1\)):

\[
\Phi_\beta = \frac{1}{\sigma^4} \left[ \sigma^2 \| \Sigma^-_p \|_+ + 2 \| \vec{\rho} \otimes \vec{\rho}^T \|_+ - 6 \beta \| \vec{\rho}^T \otimes \Sigma^-_p \|_+ + 3 \beta^2 \| \Sigma^-_p \otimes \Sigma^-_p \|_+ \right]
\]

(22)

where \(\| A \|_+\) is the simple summation of all the entries of matrix \(A\) (not to be confused with a matrix norm) and \(\otimes\) denotes the Kronecker product. From an information geometry perspective, the presence of tensor products indicates the intrinsic differential geometry of a manifold in the form of the Riemman curvature tensor [16].

Following the same methodology, a closed form expression for \(\Psi_\beta\) can also be obtained using the second derivative of the pseudo-likelihood function. Plugging equation (2) into (13), leads us to the following expression. Note that unlike \(\Phi_\beta\), \(\Psi_\beta\) does not depend explicitly on \(\beta\) (inverse temperature) (\(\Phi_\beta\) is a quadratic function of the spatial dependence parameter).

\[
\Psi_\beta = \frac{1}{\sigma^2} \sum_{i=1}^{n} E \left\{ \left[ \sum_{x_j \in \eta_i} (x_j - \mu) \right]^2 \right\}
\]

\[
= \frac{1}{\sigma^2} \sum_{i=1}^{n} E \left[ \sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} (x_j - \mu)(x_k - \mu) \right] =
\]

\[
= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left\{ \sum_{x_j \in \eta_i} \sum_{x_k \in \eta_i} E [(x_j - \mu)(x_k - \mu)] \right\} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sigma_{jk}
\]

(23)
Assuming the observed data as the outcome of a complex system at a given time instant \( t \) \((n = 1)\), \( \Psi_\beta \) can be rewritten as:

\[
\Psi_\beta = \frac{1}{\sigma^2} \| \Sigma_p^- \|_+
\] (24)

Looking at the derived Fisher information expressions we note that a trivial condition for information equality, \( \Phi_\beta = \Psi_\beta \), is \( \beta = 0 \) and \( \sigma_{ij} = 0, \forall j \), which are, essentially, equivalent conditions to express no correlation between the MRF variables. Formally, the condition for information equilibrium (\( \Phi_\beta = \Psi_\beta \)) is achieved when:

\[
\Delta_\beta (\tilde{\rho}, \Sigma_p) = \frac{1}{\sigma^2} \left( 2 \| \tilde{\rho} \otimes \tilde{\rho}^T \|_+ - 6\beta \| \tilde{\rho}^T \otimes \Sigma_p^- \|_+ + 3\beta^2 \| \Sigma_p^- \otimes \Sigma_p^- \|_+ \right) = 0 \] (25)

Solving this simple quadratic equation lead us to the following values for \( \beta^* \):

\[
\beta^* = \frac{\| \tilde{\rho}^T \otimes \Sigma_p^- \|_+ + \sqrt{3} \sqrt{3 \| \tilde{\rho}^T \otimes \Sigma_p^- \|_+^2 - 2 \| \Sigma_p^- \otimes \Sigma_p^- \|_+ \| \tilde{\rho} \otimes \tilde{\rho}^T \|_+}}{3 \| \Sigma_p^- \otimes \Sigma_p^- \|_+} \] (26)

provided that \( 3 \| \tilde{\rho}^T \otimes \Sigma_p^- \|_+^2 \geq 2 \| \Sigma_p^- \otimes \Sigma_p^- \|_+ \| \tilde{\rho} \otimes \tilde{\rho}^T \|_+ \) and \( \| \Sigma_p^- \otimes \Sigma_p^- \|_+ \neq 0 \).

Experiments and empirical analysis with MCMC simulations shown in Section 6 reveal an interesting relationship between \( \Phi \) and \( \Psi \), indicating, among other things, that they can be useful in extracting relevant information from data, estimating and predicting the behavior of a dynamical system.

Before the experimental simulations, the next Section discusses another important information-theoretic connection which is the relation between Fisher information, Shannon entropy and the inverse temperature parameter (\( \beta \)) in pairwise GMRF’s.

**ENTROPY ON PAIRWISE GMRF’S**

In the classical situation of independent and identically distributed random variables \( X_1, X_2, \ldots, X_n \), the global Shannon entropy \( H \) is given in terms of the expected value of the self-information, \( H(X_i) = E[I(X)] \), where \( I(X) = -\log(p(X_i)) \), by simply multiplying
\(H(X_i)\) by the sample size \(n\). In this Section we derive an expression for the global entropy by considering the pseudo-likelihood function. Considering a pairwise GMRF model, the expression for the global entropy of \(n\) spatially dependent observations of a random variable is given by the following:

\[
H_{\beta} = -E \left[ \log \prod_{i=1}^{n} p(x_i|\eta_i, \theta) \right] = \\
\frac{n}{2} \log (2\pi \sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} E \left\{ \left[ x_i - \mu - \beta \sum_{x_j \in \eta_i} (x_i - \mu) \right]^2 \right\} = \\
\frac{n}{2} \log (2\pi \sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\{ E \left[ (x_i - \mu)^2 \right] - 2\beta E \left[ \sum_{x_j \in \eta_i} (x_i - \mu) (x_j - \mu) \right] + \beta^2 E \left\{ \sum_{x_j \in \eta_i} (x_i - \mu) \right\}^2 \right\} \\
\]

After some algebra we have:

\[
H_{\beta} = \frac{n}{2} \log (2\pi \sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\{ \sigma^2 - 2\beta \sum_{j \in \eta_i} \sigma_{ij} + \beta^2 \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sigma_{jk} \right\} = \\
\frac{n}{2} \log (2\pi) + \frac{n}{2} \log (\sigma^2) + \frac{n}{2} - \frac{\beta}{\sigma^2} \sum_{i=1}^{n} \left[ \sum_{j \in \eta_i} \sigma_{ij} \right] + \frac{\beta^2}{2\sigma^2} \sum_{i=1}^{n} \left[ \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sigma_{jk} \right] \quad (28) \\
\]

Analogous to the previous expressions, assuming a matrix-vector notation, we have:

\[
H_{\beta} = H_G - \left[ \frac{\beta}{\sigma^2} \| \bar{\rho} \|_+ - \frac{\beta^2}{2} \Psi_{\beta} \right] \quad (29)
\]

where \(H_G\) denotes the entropy of a Gaussian random variable.
Note that, Shannon entropy is a quadratic function of the spatial dependence parameter $\beta$. Since the coefficient of the quadratic term is strictly non-negative (it is the expected Fisher information), entropy is a convex function of $\beta$. Also, as expected, when $\beta = 0$, the resulting expression is the traditional equation for the entropy of a Gaussian random variable. Thus, calculating the value $\beta_{MH}$ that minimizes this form of entropy, leads to:

$$\frac{\partial H_\beta}{\partial \beta} = \frac{\beta}{\sigma^2} \| \Sigma_p \|_+ - \frac{1}{\sigma^2} \| \rho \|_+ = 0$$

$$\beta_{MH} = \frac{\| \rho \|_+}{\| \Sigma_p \|_+}$$

$$\frac{\beta_{MPL}}{\beta_{MH}} = 1$$  \hspace{1cm} (30)

showing that maximum pseudo-likelihood and minimum-entropy estimation are essentially equivalent in GMRF models. Moreover, using the derived equations we have the relation between $\Phi_\beta$, $\Psi_\beta$ and $H_\beta$ as:

$$\Phi_\beta - \Psi_\beta = \Delta_\beta (\rho, \Sigma_p)$$ \hspace{1cm} (31)

$$\frac{\partial^2 H_\beta}{\partial \beta^2} = \Psi_\beta$$ \hspace{1cm} (32)

where the functional $\Delta_\beta (\rho, \Sigma_p)$ that represents the difference between $\Phi_\beta$ and $\Psi_\beta$ is given by equation [25]. These equations relate the entropy and one form of Fisher information ($\Psi_\beta$) in GMRF models, showing that $\Psi_\beta$ can be roughly viewed as the curvature of $H_\beta$. In this sense, in an information equilibrium condition $\Psi_\beta = \Phi_\beta = 0$, the entropy’s curvature is null. Since $\Psi_\beta$ is strictly positive (Fisher information), $\beta_{MH}$ is actually a critical point. The results also suggest that an increase in the value of $\Phi_\beta$, which means stability (a measure of how the neighboring variables of a given point agree in their behavior), contributes to curve, and therefore to cause a change in the entropy of the system.

Experiments with both Markov Chain Monte Carlo simulation and real image data show that entropy can be quite useful in detecting the appearance of informative patterns. More
details can be found in Section 6. In the next Section, we will move forward to see how Fisher information is also related to measuring the uncertainty in the estimation of the inverse temperature parameter, $\beta$, which determines the global system behavior, by means of the asymptotic variance of maximum pseudo-likelihood estimators.

**ASYMPTOTIC VARIANCE OF MPL ESTIMATORS**

Unbiasedness is not granted by either ML or MPL estimation. Actually, there is no method that guarantees the existence of unbiased estimators for a fixed $n$-size sample. Often, in the exponential family, MLE coincide with UMVU (Uniform Minimum Variance Unbiased) estimators because they are functions of complete sufficient statistics (if MLE is unique, then it is a function of sufficient statistics). Also, there are several characteristics that make ML estimation a reference method \[13,15\]. Making the sample size grow infinitely ($N \to \infty$), MLE becomes asymptotically unbiased and efficient. Unfortunately, there is no result showing that the same occurs in MPL estimation. In this section, we propose an exact expression for the asymptotic variance of $\beta$ on an isotropic pairwise GMRF model in terms of both forms of expected Fisher information, $\Psi_\beta$ and $\Phi_\beta$.

**Asymptotic Variance of the Spatial Dependence parameter on the GMRF Model**

Asymptotic evaluations uncover the most fundamental properties of a mathematical procedure, providing a powerful and general tool for statistical analysis. In this section we derive an expression for the asymptotic variance of the MPL estimator of the pairwise GMRF spatial dependence parameter $\beta$. It is known from the statistical inference literature that both ML and MPL estimators share two important properties: consistency and asymptotic normality \[32,33\], making it possible to completely characterize their behavior in the limiting case. In other words, $\beta_{MPL} \approx N\left(\hat{\beta}_{MPL}, \upsilon_\beta\right)$, where $\upsilon_\beta$ denotes the asymptotic variance. It is known that the asymptotic covariance matrix of MPL estimators is given by \[34\]:

$$C(\vec{\beta}) = H^{-1}(\vec{\beta})J(\vec{\beta})H^{-1}(\vec{\beta})$$  \hspace{1cm} (36)

with
\[ H(\vec{\beta}) = E_\beta \left[ \nabla^2 \log PL(\vec{\beta}) \right] \] (37)

\[ J(\vec{\beta}) = \text{Var}_\beta \left[ \nabla \log PL(\vec{\beta}) \right] \] (38)

where \( H \) and \( J \) denotes the Jacobian and Hessian matrices regarding the log pseudo-likelihood function, respectively. Thus, in the uniparametric case we have the following definition for the asymptotic variance \( \upsilon_{\beta} \):

\[
\upsilon_{\beta} = \frac{\text{Var}_\beta \left[ \frac{\partial}{\partial \beta} \log PL(\beta) \right]}{E_\beta \left[ \frac{\partial^2}{\partial \beta^2} \log PL(\vec{\beta}) \right]} = \frac{E_\beta \left[ \left( \frac{\partial}{\partial \beta} \log PL(\beta) \right)^2 \right] - E_\beta \left[ \frac{\partial}{\partial \beta} \log PL(\beta) \right]^2}{E_\beta \left[ \frac{\partial^2}{\partial \beta^2} \log PL(\beta) \right]} \] (39)

Noting that the expected value of the derivative of the log pseudo-likelihood equation is zero:

\[
E \left[ \frac{\partial}{\partial \beta} \log PL (\beta) \right] = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left\{ E [x_i - \mu] - \beta \sum_{x_j \in \eta_i} E [x_j - \mu] \right\} = 0 \] (40)

we finally get the resulting expression for the asymptotic variance of \( \hat{\beta}_{MPL} \) as the ratio between \( \Phi_{\beta} \) and \( \Psi_{\beta}^2 \), given by:

\[
\upsilon_{\beta} = \frac{1}{n} \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sum_{l \in \eta_i} \sum_{m \in \eta_i} \sigma_{jk} \sigma_{lm} \left\{ \sum_{\beta \in \eta_i} \sum_{k \in \eta_i} \left[ \sigma^2 \sigma_{jk} + 2 \sigma_{ij} \sigma_{ik} \right] - 2 \beta \sum_{j \in \eta_i} \sum_{k \in \eta_i} \sum_{l \in \eta_i} \left[ \sigma_{ij} \sigma_{kl} + \sigma_{ik} \sigma_{jl} + \sigma_{il} \sigma_{jk} \right] \right\} \] (41)
Using the matrix-vector notation and considering the observed data as outcomes of a single random pattern, the above equation is simplified to:

\[
υ β = \frac{σ^2 \left\| \Sigma_p^- \right\|_+ + 2 \left\| \hat{ρ} \otimes \hat{ρ}^T \right\|_+ - 6β \left\| \hat{ρ}^T \otimes \Sigma_p^- \right\|_+ + 3β^2 \left\| \Sigma_p^- \otimes \Sigma_p^- \right\|_+}{\left\| \Sigma_p^- \right\|_+^2} + 2 \left\| \hat{ρ} \otimes \hat{ρ}^T \right\|_+ - 6β \left\| \hat{ρ}^T \otimes \Sigma_p^- \right\|_+ + 3β^2 \left\| \Sigma_p^- \otimes \Sigma_p^- \right\|_+ + \left\| \Sigma_p^- \right\|_+^2
\]

(42)

Note that, since \(υ β = Φ β / Ψ β\), if we had equivalence between the two forms of Fisher information, that is, \(Φ β = Ψ β \neq 0\), the expression for \(υ β\) would be simplified to the traditional Cramer-Rao lower bound. The interpretation of this equation indicates that the accuracy in \(β\) estimation depends essentially on two main factors: 1) minimization of \(Φ β\), which means the variance of the local log-likelihood functions is close to zero (concentration of patterns that are “aligned” to the expected global behavior, that is, patterns showing a high degree of agreement between the central element and its neighbors) and 2) maximization of \(Ψ β\), which essentially means that, in average, the local log-likelihood functions are not flat, that is, small variations on patterns cannot cause abrupt changes in the spatial dependence structure and therefore in the expected global behavior (that is, concentration of stable patterns showing a high degree of agreement among the neighboring elements themselves). Finally, once we cannot obtain a unique lower bound in terms of Fisher information, since information equality fails, it is not possible to conclude that the estimator is asymptotic efficient, as it would happen if we had independent observations (maximum-likelihood estimation).

**COMPUTATIONAL SIMULATIONS**

This Section presents some numerical experiments to illustrate some applications of the proposed tools, especially the two Fisher information measures \(Φ\) and \(Ψ\), in the context of both Markov Chain Monte Carlo simulations and real image data corrupted by additive random Gaussian noise. The empirical investigations were divided in two sets of experiments:

1. analysis of the local versions of the measures \(φ β\) and \(ψ β\) in both simulated and real images;

2. analysis of the global versions of the measures \(Φ β\), \(Ψ β\) and \(H β\) along a MCMC simulation in which the behavior of a complex system is disturbed;
FIG. 2. Example a white noise random field and the outcome of a second order pairwise isotropic GMRF model for $\beta = 0.125$.

**Analysis of local Fisher information**

To illustrate the application of both forms of local observed Fisher information, $\phi_\beta$ and $\psi_\beta$, we generated some synthetic images by using the Metropolis-Hastings algorithm with a white noise gaussian field as initialization. Figure 2 shows an example of input and an valid outcome for a second order (8 nearest neighbors) pairwise isotropic GMRF model using the following parameters setttings: $\mu = 0$, $\sigma^2 = 5$ and $\beta = 0.125$. The number of iterations used in the MCMC simulation was 1000.

Three Fisher information maps were generated from the resulting synthetic image. The first one was obtained by calculating the value of type-I local Fisher information (equation 9 for $n = 1$), $\phi_\beta$, for every element of the system. Similarly, the second one was obtained by using type-II local Fisher information (equation 11 for $n = 1$), $\psi_\beta$. For the last one, we used the ratio of type-I and type-II observed Fisher information, motivated by the fact that boundaries are mostly composed by patterns that are not expected to be “aligned” to the global behavior (high $\phi_\beta$) and also are somehow unstable (low $\psi_\beta$). We call this measure, $\phi_\beta/\psi_\beta$, *L-information*, since it is defined in terms of the two first derivatives of the local Log-PsudoLikelihood function. Figure 3 shows the visual results. Note that while $\phi_\beta$ has a strong response for boundaries, $\psi_\beta$ has a weak one, an evidence in favor of considering *L-information* in detection procedures.

The same experiment was repeated for grayscale images corrupted by additive gaussian noise. It is known from image processing and computer vision literatures that the problem
FIG. 3. Information maps obtained using both forms of Fisher information, $\phi_\beta$ and $\psi_\beta$, and also L-information, respectively.

of detecting edges in the presence of noise data is extremelly challenging, since typical edge detectors are based on differencial operators (which causes noise amplification). We believe the proposed method provides a reasonable solution to such cases since the obtained results were promissing. Figures 4 and 5 show noisy images, the result of the Laplacian edge detector, the result of canny edge detector and the respective L-information maps. Note that the response of L-information gives a good approximation to the image boundaries, even in the presence of random perturbation (noise). Note also that the L-information map retains the most of the relevant information without an excessive smoothing (loss of fine details).

Another result concerns the Boat image, which is known to have a complex structure due to the presence of fine details, thin lines and some texture. Basically, the same methodology described in the previous experiment was adopted here. Figure 5 shows the obtained results. Again, a reasonable amount of relevant information could be extracted with the proposed method in comparison with the Laplacian and Canny techniques.

To illustrate how entropy in GMRF models can be used to quantify and and measure variability in data, we show an example using some grayscale images. For this experiment we would like to know which of the four images - Baboon, Lena, Cameraman and a texture piece - contains higher entropy in terms of a GMRF model, that is, which one of them has more variation of contextual patterns. Figure 7 shows the values of Fisher information and entropy for each image. The results indicate that the Baboon image has the lowest entropy in comparison to the others and the texture piece has the highest value. This fact is possibly due to the presence of strong edges and some fine details making the distribution of patterns more uniform, defining a more unstable configuration.
Finally, to show how entropy is related to the distribution of patterns along the GMRF we compared its value for 4 versions of the Lena image from a very blurred one (low variability of contextual patterns) to a very noisy one (high variability of contextual patterns). Figure 6 shows the obtained results. As expected the entropy of the noisy images is higher.
FIG. 5. Edge detection performed in a noisy version of the Boat image. The results from top to bottom and left to right show the observed noise image and the Laplacian, Canny and L-information ($\beta_{MPL} = 0.1266$) maps, respectively.

Analysis of expected Fisher information and entropy

In order to study the behavior of both forms of the expected Fisher information and the entropy in GMRF models, a sequence of valid GMRF outcomes was generated by the Metropolis-Hastings algorithm. The purpose of this experiment is to observe what happens to $\Phi_\beta$, $\Psi_\beta$ and $H_\beta$ when the system evolves from a random initial state to other configurations.

To simulate a system in contact with a reservoir of heat that can control the temperature (or in other words the parameter $\beta$) we defined the value of $\beta$ in the current iteration as
FIG. 6. Information-theoretic measures for different versions of the Lena image, form blurred to noisy. The results for $H_\beta$ from left to right and top to bottom were: 4.6579, 4.7243, 4.9001 and 5.2040. The results for $\Phi_\beta$ were: 0.0580, 0.4776, 8.7643 and 17.7712. Finally, the results for $\Psi_\beta$ were: 62.7282, 60.9517, 52.6158, 39.3270.

FIG. 7. Grayscale images Baboon, Lena, Cameraman and a texture piece. The obtained values for $H_\beta$ were: 4.6363, 4.7857, 5.0825 and 5.1590, respectively. For $\Phi_\beta$ the values were: 8.112, 2.8516, 3.077 and 5.6932. Finally, the obtained values for $\Psi_\beta$ were: 51.101, 58.0462, 58.3909 and 19.2041. The usual entropy calculated by estimating the probabilities from the image histogram show the values: 7.2279, 7.4227, 7.0097 and 6.2418, respectively.
the previous one plus a small increment ($\Delta \beta = 0.001$) until it reaches a maximum value, where the next value of $\beta$ will be defined as the current one minus $\Delta \beta$ while $\beta > 0$. This process is repeated for a fixed number of iterations during the MCMC simulation. As a result of this algorithm, a sequence of GMRF samples is produced. We used this sequence to calculate $\Phi_\beta$, $\Psi_\beta$ and $H_\beta$. Figure 8 shows some of the system’s configurations along a MCMC simulation.

A plot of both forms of the expected Fisher information, $\Phi_\beta$ and $\Psi_\beta$, in each iteration of the MCMC simulation is shown in Figure 9. The graph produced by this experiment show some interesting results. First of all, regarding upper and lower bounds on these quantities, it is possible to note that when there is no induced spatial dependence structure ($\beta \approx 0$), $\Phi_\beta = \Psi_\beta$, showing the information equilibrium condition. This information equilibrium means that the variables are statistically independent in the sense that when we observe configuration patterns, they all have the same amount of information so it is hard to find different kinds of patterns: informative and non-informative (since they all behave in a similar way and there is no informative pattern to highlight). Moreover, in the information equilibrium, $\Psi_\beta$ reaches its lower bound indicating that this condition happens when the system is most susceptible to a global change in the sense that modification of the behavior of a small subset of patterns may guide the system to a different path of evolution. The
FIG. 9. Evolution of Fisher information along a MCMC simulation where the parameter $\beta$ is modified in order to control the system’s behavior. The higher the divergence between $\Phi_\beta$ and $\Psi_\beta$ the less susceptible to changes is the global behavior (*). In the information equilibrium condition (**) there is no induced spatial dependence structure and it is hard to find informative patterns.

The graph also shows that the difference between $\Phi_\beta$ and $\Psi_\beta$ is maximum when the system has large values of $\beta$, that is, when organization emerges and the system shows a strong dependence between its elements with clear visible clusters and boundaries. In such state, it is expected that the majority of patterns be aligned to the global behavior, which causes the appearance of few but highly informative patterns: those defining the boundaries between different regions. Also, the simulations suggest that it takes more time for the system to go from the information equilibrium state to organization than the opposite. From a statistical inference perspective, according to the asymptotic variance of $\beta$’s MPL, the uncertainty on the estimation of the inverse temperature is minimized when the system is more organized (for larger values of $\beta$), which is somehow intuitive since stability ($\Psi_\beta$) is higher and the universe of possible values for the temperature of the system is restricted (it is bounded by a critical value). Finally, the results suggest that both $\Phi_\beta$ and $\Psi_\beta$ are bounded by a superior value related to the size of the neighborhood system (with the minimum possible value being zero, obviously).

Despite showing a similar behavior, entropy can have larger values in comparison to both $\Phi_\beta$ and $\Psi_\beta$. Besides, from the graphs it is possible to note that knowledge of both Fisher information forms allows us to predict an increase in $H_\beta$. According to our experiments, the
FIG. 10. Evolution of entropy along a MCMC simulation where the parameter $\beta$ is modified in order to control the system’s behavior. $H_\beta$ start to grow when the system leaves the information equilibrium, where the entropy in the GMRF model is identical to the entropy of a simple Gaussian random variable.

Results suggest that a necessary condition to increase the entropy of a GMRF model, is a divergence in the values of $\Phi_\beta$ and $\Psi_\beta$, that is, the information equilibrium condition cannot prevail. Looking at Figures 9 and 10 we can see that $\Phi_\beta$ and $\Psi_\beta$ start to diverge around the iteration 100 while entropy start to grow a little bit later, around the iteration 170.

Another interesting measure is the global L-information, that is, the ratio between the two forms of expected Fisher information, $\Phi_\beta/\Psi_\beta$. An interesting property of this measure is that its values are always bounded by the $[0, 1]$ interval. With this single measurement it is possible to estimate and predict the global system behavior as shows Figure 11. A value close to one indicates the system is approximating the information equilibrium and a value close to zero indicates the system tends to a very organized and cohesive behavior (segmented), with clear informative patterns (boundaries).

To show the intrinsic non-linear relation between $\Phi_\beta$, $\Psi_\beta$ and $H_\beta$, we plotted the trajectories of the system’s state along the MCMC simulation in the information space defined by each one of these quantities. Each configuration of the system at a given iteration becomes a 3-D vector $\vec{I}_\beta = (\Phi_\beta, \Psi_\beta, H_\beta)$. Figure 12 shows the results. It is clear from the graph that which causes entropy to increase is a given combination of values of $\Phi_\beta$ and $\Psi_\beta$, more
FIG. 11. Evolution of L-information along a MCMC simulation where the parameter $\beta$ is modified in order to control the system’s behavior. When L-information approaches one the system tends to information equilibrium. For values close to zero, we have stable, organized and informative configuration patterns coexisting throughout the system (different classes of patterns are observable).

precisely when they start to diverge, that is, a decrease in the L-information value. Hence, the empirical analysis suggests that entropy and L-information are inversely proportional, as show Figure 13. Note also that, from a differential geometry perspective, the majority of the points of the parametric curve (those showing higher curvature values) are concentrated in the two main regions of the space: A) around the information equilibrium condition (desorganization) and B) around the maximum entropy value, where the divergence between the information values are maximum (organization emerges). The few points in the trajectory connecting these two regions possibly represent the system going through a phase transition. Its properties change rapidly and and in an assimetric way (the path from state A to state B is different from the path from state B to state A).

Finally, to show the influence of the system’s global dependence structure on the asymptotic variance of the inverse temperature MPL estimator (uncertainty about the $\beta$ parameter) we show a curve of its variances for different values of this parameter. As expected, the uncertainty is higher in states close to the information equilibrium since there is no relevant information in the observed contextual patterns (the patterns tend to behave sim-
FIG. 12. Trajectory of the system’s configurations in 2-D and 3-D information space defined by $(\Phi_\beta, \Psi_\beta)$ and $(\Phi_\beta, \Psi_\beta, H_\beta)$ along a MCMC simulation. The graph shows a parametric curve obtained by varying the $\beta$ parameter from 0 to 0.15 and back, repeatedly. Note that, from a differential geometry perspective, as the divergence between $\Phi_\beta$ and $\Psi_\beta$ increases, the torsion of the parametric curve becomes evident (the curve leaves the plane).

Similarly, Figure 14 illustrates this behavior with the evolution of $\beta$ (starting in 0 and with fixed increments until a maximum value is obtained and back) and $\beta^*$ (given by equation 26) along a MCMC simulation. It is interesting to mention that for larger values of $\beta$, the corresponding information equilibrium value $\beta^*$ is lower than $\beta$, but not null (greater than zero). In our MCMC simulations, the measured values of $\beta^*$ were always less than 0.0577, indicating an upper bound for the observation of the information equilibrium condition. During the experiments we verified that, given an observed global configuration at a time.
FIG. 13. Trajectory of the system’s configurations in the space defined by entropy and L-information along a MCMC simulation. The graph shows a parametric curve obtained by varying the $\beta$ parameter from 0 to 0.15 and back, repeatedly. As L-information increases, entropy decreases and vice-versa.

At instant $t$, a less abrupt method to perturb the system’s behavior in order to induce small changes and control the evolution of some components of the system is to set $\beta = \beta^*$ instead of zero directly.

Figures 15 and 16 shows the evolution of $\Phi_{\beta}$ and $\Psi_{\beta}$ after the system is disturbed by setting $\beta = 0$ and $\beta = \beta^*$ for the next consecutive five iterations of the MCMC simulation, returning to its original value in sequence. When the system is disturbed by making $\beta$ null the simulations indicate that the system is not successful in recovering components from its previous stable configuration (note that $\Phi_{\beta}$ and $\Psi_{\beta}$ clearly touch one another in the graph). When the same perturbation is applied but using the smallest of the two $\beta^*$ values (minimum solution of equation 26), after a short stabilization period, the system is able to recover parts of its previous stable state. This result suggests that this kind of perturbation is not enough to remove all the information encoded within the spatial dependence struture of the configuration patterns, preserving some short and long range correlations in data but slightly remodeling large cluster and removing smaller ones.

Figure 17 shows a sequence of outputs along this MCMC simulation. The first row shows an abrupt change in the stable configuration after the change on $\beta$ (regarding the simulation
FIG. 14. Asymptotic variances of the MPL estimator of $\beta$ and values of $\beta^*$ along a MCMC simulation. The uncertainty about the spatial dependence parameter increases as the system tends to information equilibrium since it is hard to extract information from the observed patterns.

depicted in the graph of Figure 15 while the second row indicates that the system was able to evolve to a stable configuration that is quite similar the previous one, recovering most of the original components (regarding the simulation depicted in the graph of Figure 16). Future works may include a detailed comparison on the evolution of complex systems considering two situations: normal evolution or evolution after an induced perturbation. This would allow us to study how similar or not the systems will behave in a future time, both in terms of cluster structure similarity (their major components remain similar) and information metrics.

CONCLUSIONS AND FINAL REMARKS

In this paper we addressed the problem of characterizing the spatial dependence structure of a pairwise isotropic Gaussian Markov Random Field defined on a lattice by means of information theoretic measures. Analytical expressions for observed and expected Fisher information regarding the spatial dependence parameter on a GMRF model were derived using the pseudo-likelihood function, elucidating the connection between the inverse temperature parameter $\beta$ and the covariance matrix of contextual patterns presented in data.
FIG. 15. Variation on $\Phi_\beta$ and $\Psi_\beta$ after the system is disturbed by a change on the value of $\beta$ to zero along a MCMC simulation. Note that $\Phi_\beta$ and $\Psi_\beta$ touch one another indicating that no residual information is kept as if the simulation had been restarted from a random configuration.

Intuitive physical interpretations for these information-theoretic quantities were discussed in the context of MRF models. However, to allow the computation of these measures, a proper $\beta$ parameter estimative is required. Maximum Pseudo-Likelihood estimators for the GMRF model parameters were derived, indicating that $\hat{\beta}_{MPL}$ is completely defined in terms of spatial correlation coefficients. Using the same methodology, an expression for the Shannon entropy of a GMRF was derived, showing its relation with Fisher information and maximum pseudo-likelihood estimation. Finally, using the derived expressions for the Fisher information, an exact expression for the asymptotic variance of $\hat{\beta}_{MPL}$ in the GMRF model was presented, allowing a complete characterization of the system’s behavior in the limiting case. The proposed experiments and numerical evaluations obtained with MCMC simulations supported the theoretical background by showing practical results in analysis of both static and dynamic data. Future works may include an information-theoretic study about the spatial dependence structure on Gaussian, Ising and Potts pairwise MRF models defined on lattices and non-regular graphs, Markov Chain Monte Carlo simulation for analysis of temporal correlations in the study of phase-transitions in complex systems, information geometry based approaches on MRF models and also applications on signal, image and video processing, such as the development of coding, compression, denoising, segmentation and data analysis techniques.
FIG. 16. Variation on $\Phi_\beta$ and $\Psi_\beta$ after the system is disturbed by a change on the value of $\beta$ to $\beta^*$ along a MCMC simulation. Note that the perturbation is not enough to remove all the information within the spatial patterns allowing the system to recover a significant part of its original configuration.

ACKNOWLEDGMENTS

The author would like to thank CNPQ (Brazilian Council for Research and Development) for the financial support through the research grant number 475054/2011-3.

*alexandre@dc.ufscar.br http://www.dc.ufscar.br/~alexandre

[1] C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Chicago, IL & London, 1949).

[2] A. Rényi, in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* (1960) pp. 547–561.

[3] C. Tsallis, Journal of Statistical Physics 52, 479 (1988).

[4] A. Bashkirov, Theoretical and Mathematical Physics 149, 1559 (2006).

[5] E. Jaynes, Physical Review 106, 620 (1957).

[6] H. Grad, Communications in Pure and Applied Mathematics 14, 323 (1961).

[7] R. Adler, A. Konheim, and A. McAndrew, Transactions of the American Mathematical Society 114, 309 (1965).
FIG. 17. Sequence of outputs along the MCMC simulation. The first row ($\beta$ to zero) shows a completely different stable configuration after the perturbation while the second row ($\beta$ to $\beta^*$) indicates that the system was able to evolve to a stable configuration similar to its previous one.

[8] L. Goodwyn, American Journal of Mathematics 94, 366 (1972).
[9] P. A. Samuelson, The American Economic Review 62, 249 (1972).
[10] M. Costa, IEEE Transactions on Information Theory 29, 439 (1983).
[11] A. Dembo, T. Cover, and J. Thomas, IEEE Transactions on Information Theory 37, 1501 (1991).
[12] T. Cover and Z. Zhang, IEEE Transactions on Information Theory 40, 1244 (1994).
[13] E. L. Lehmann, Theory of Point Estimation (Wiley, New York, 1983).
[14] P. J. Bickel, Mathematical Statistics (Holden Day, New York, 1991).
[15] G. Casella and R. L. Berger, Statistical Inference, 2nd ed. (Duxbury, New York, 2002).
[16] N. H. Amari, S., Methods of information geometry (Translations of mathematical monographs v. 191) (American Mathematical Society, 2000).
[17] R. E. Kass, Statistical Science 4, 188 (1989).
[18] A. Anandkumar, L. Tong, and A. Swami, IEEE Trans. on Information Theory 55, 816 (2009).
[19] M. A. Gómez-Villegas, P. Main, and R. Susi, Information Sciences 222, 439 (2013) including Special Section on New Trends in Ambient Intelligence and Bio-inspired Systems.
[20] J. Moura and N. Balram, IEEE Trans. on Information Theory 38, 334 (1992).
[21] J. Moura and S. Goswami, IEEE Trans. on Information Theory 43, 1560 (1997).
[22] J. Besag, Journal of the Royal Statistical Society - Series B 36, 192 (1974).

[23] J. Besag, The Statistician 24, 179 (1975).

[24] A. L. M. Levada, N. D. A. Mascarenhas, and A. Tannús, in Proceedings of the 19th International Conference on Pattern Recognition (ICPR) (IEEE, Tampa/FL, 2008) pp. 1–4.

[25] J. Hammersley and P. Clifford, “Markov field on finite graphs and lattices,” (1971), unpublished.

[26] R. Chellappa, “Progress in pattern recognition 2,” (North Holland Pub. Co., 1985) Chap. Two-dimensional discrete gaussian Markov random field models for image processing, pp. 79–112.

[27] B. F. Efron and D. V. Hinkley, Biometrika 65, 457 (1978).

[28] A. L. M. Levada, N. D. A. Mascarenhas, and A. Tannús, Pattern Recognition Letters 31, 1795 (2010).

[29] A. L. M. Levada and D. C. Corrêa, in Proceedings of the 24th International Symposium on Circuits and Systems (ISCAS 2011) (2011) pp. 125–128.

[30] D. H. P. Salvadeo, N. D. A. Mascarenhas, and A. L. M. Levada, in Proceedings..., edited by C. M. D. S. Freitas, S. Sarkar, R. Scopigno, and L. Silva, Conference on Graphics, Patterns and Images, 25. (SIBGRAPI) (IEEE Computer Society Conference Publishing Services, Los Alamitos, 2012).

[31] L. Isserlis, Biometrika 12, 134 (1918).

[32] J. Jensen and H. Künsch, Annals of the Institute of Statistical Mathematics 46, 475 (1994).

[33] G. Winkler, Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).

[34] G. Liang and B. Yu, IEEE Trans. on Signal Processing 51, 2043 (2003).