Noname manuscript No.
(will be inserted by the editor)

Pairwise likelihood estimation for multivariate mixed Poisson models generated by Gamma intensities

Florent Chatelain · Sophie Lambert-Lacroix · Jean-Yves Tourneret

Abstract Estimating the parameters of multivariate mixed Poisson models is an important problem in image processing applications, especially for active imaging or astronomy. The classical maximum likelihood approach cannot be used for these models since the corresponding masses cannot be expressed in a simple closed form. This paper studies a maximum pairwise likelihood approach to estimate the parameters of multivariate mixed Poisson models when the mixing distribution is a multivariate Gamma distribution. The consistency and asymptotic normality of this estimator are derived. Simulations conducted on synthetic data illustrate these results and show that the proposed estimator outperforms classical estimators based on the method of moments. An application to change detection in low-flux images is also investigated.

Keywords Pairwise likelihood estimation · multivariate mixed Poisson models · multivariate Gamma distributions · negative multinomial distributions

1 Introduction

Univariate mixed Poisson distributions have received much attention in statistics and image processing applications (see for instance **??**, and the references therein). These applications include active imaging, where the image is obtained from a scene illuminated with laser light (**?**), or astronomy, where low-flux images are recorded by using

F. Chatelain
IRIT/ENSEEIHT/Tésa, 2 rue Charles Camichel, BP 7122, 31071 Toulouse cedex 7, France
E-mail: florent.chatelain@sophia.inria.fr

S. Lambert-Lacroix (**corresponding author**)
Laboratoire Jean Kuntzmann, Université de Grenoble et CNRS, 51 rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9, France
Tel: +33 4.76.51.45.47; Fax: +33 4.76.63.12.63
E-mail: Sophie.Lambert@imag.fr

J.-Y. Tourneret
IRIT/ENSEEIHT/Tésa, 2 rue Charles Camichel, BP 7122, 31071 Toulouse cedex 7, France
E-mail: jean-yves.tourneret@enseeiht.fr

photocounting cameras (**?**). A univariate mixed Poisson distribution is the distribution of a random variable N such that the conditional distribution of $N|\lambda$ is a Poisson distribution with parameter λ (denoted as $N|\lambda \sim \mathcal{P}(\lambda)$). The parameter λ is also a random variable (called intensity) whose distribution is referred to as structure distribution (**?**) or mixing distribution. When λ has an absolutely continuous distribution defined on $\mathbb{R}^+$ (whose probability density function (pdf) is denoted as $f_1(\lambda)$), the probability masses of N can be written:

$$\begin{aligned}\mathbb{P}(N=n) &= \int_0^\infty \mathbb{P}(N=n|\lambda) f_1(\lambda) d\lambda, \\ &= \int_0^\infty \frac{\lambda^n}{n!} \exp(-\lambda) f_1(\lambda) d\lambda. \end{aligned} \tag{1}$$

Multivariate extensions of mixed Poisson distributions are naturally constructed from a joint intensity pdf $f_d(\boldsymbol{\lambda})$ defined on $\mathbb{R}_+^d$. The corresponding masses of the d-multivariate variable $\boldsymbol{N} = (N_1, ..., N_d)$, can be computed as follows:

$$\mathbb{P}(\boldsymbol{N}=\boldsymbol{n}) = \int_{\mathbb{R}_+^d} \cdots \int \prod_{\ell=1}^{d} \frac{(\lambda_\ell)^{n_\ell}}{n_\ell!} \exp(-\lambda_\ell) f_d(\boldsymbol{\lambda}) d\boldsymbol{\lambda}, \tag{2}$$

where $\boldsymbol{n} = (n_1, ..., n_d)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d)$. Some properties of multivariate mixed Poisson distributions (MMPDs) have been recently reported in **?** and **?**. For instance, conditions ensuring that MMPDs belong to an exponential family have been derived. These conditions ensure that the parameters of MMPDs can be estimated easily using the maximum likelihood (ML) principle. Unfortunately, they are not satisfied in practical image processing applications. As a consequence, estimating the parameters of MMPDs is still a challenging problem.

? recently studied methods of moments to estimate the parameters of MMPDs. However, likelihood based methods are often preferred since they usually provide estimates with lower variances. Additional reasons for preferring likelihood-type inference to the method of moments include the invariance to reparameterization and the better performance of likelihood ratio test statistics with respect to Wald-type statistics. This paper studies a maximum composite likelihood (MCL) approach to estimate the parameters of MMPDs when the mixing distribution is a multivariate Gamma distribution. A composite likelihood (CL) is a weighted sum of likelihoods associated to marginal or conditional events. The concept of CL has been widely studied in the literature (see **???**, and the references therein) since the seminal paper of **?**. Usual CLs include the composite marginal likelihood, the pairwise likelihood (**?**) and the Besag's pseudolikelihood (**?**). The maximum composite likelihood estimator (MCLE) is obtained by maximizing the corresponding CL. The advantage of using a CL instead of a standard likelihood is to reduce the computational complexity of the optimization procedure. As a consequence, it allows one to handle very complex models, even if the full likelihood cannot be expressed in a closed form. This is the case when multivariate mixed Poisson distributions are studied since the corresponding joint masses cannot be generally computed easily by using (2).

This paper is organized as follows: Section 2 presents some important results on MMPDs. Section 3 introduces the maximum pairwise likelihood estimator (MPLE) which will be considered in this paper. The consistency and asymptotical normality of the proposed MPLE are also demonstrated. Simulation results on synthetic data are

provided in Section 4. These simulations clearly show the advantage of the MPLE with respect to moment estimators. Section 5 addresses the important problem of detecting changes in synthetic aperture radar (SAR) images. The correlation coefficient of pixels belonging to images affected by a natural disaster is estimated by the maximum pairwise likelihood (MPL) method. A comparison of this estimate with an appropriate threshold (depending on the level of significance of the test) allows one to detect whether a given pixel has been affected by the disaster. The proofs of theorems are reported in the appendices.

2 MMPDs with multivariate Gamma mixing distributions

An MMPD with multivariate Gamma mixing distribution is defined by the masses (2), where $f_d(\boldsymbol{\lambda})$ is the pdf associated to a multivariate gamma distribution. For any $L \geq 0$ and for any affine polynomial $P(\boldsymbol{z})$[1], a multivariate Gamma distribution on $\mathbb{R}_+^d$ with shape parameter L and scale parameter $P(\boldsymbol{z})$, denoted as $\boldsymbol{\lambda} \sim \gamma_{L,P}$, is defined through its Laplace transform (see **?**):

$$\mathcal{L}_{\gamma_{L,P}}(\boldsymbol{z}) = \mathbb{E}\left(e^{-\boldsymbol{z}^T\boldsymbol{\lambda}}\right) = [P(\boldsymbol{z})]^{-L}, \tag{3}$$

on an appropriate domain of existence, with the obvious condition $P(0) = 1$. The main properties of multivariate gamma distributions have been reported in several recent studies including **???**. In particular, all marginal distributions of $\boldsymbol{\lambda}$ are multivariate gamma distributions. The moment generating function of an MMPD $\boldsymbol{N}$ expresses as (**?**):

$$\begin{aligned} G_{\boldsymbol{N}}(\boldsymbol{z}) = \mathbb{E}\left(\prod_{k=1}^{d} z_k^{N_k}\right) &= \mathbb{E}\left(\prod_{k=1}^{d} \mathbb{E}(z_k^{N_k}|\lambda_k)\right), \\ &= \mathbb{E}\left(\prod_{k=1}^{d} \exp\left[-\lambda_k(1-z_k)\right]\right), \qquad (4) \\ &= \mathcal{L}_{\gamma_{L,P}}(1-z_1,\ldots,1-z_d), \qquad (5) \\ &= [P(1-z_1,...,1-z_d)]^{-L}, \qquad (6) \end{aligned}$$

where $\mathcal{L}_{\gamma_{L,P}}(\boldsymbol{z})$ is the Laplace transform of the intensity distribution defined in (3) (note that the generating function of a Poisson distribution has been used to obtain (4)). Since $P(1-z_1,...,1-z_d)$ is an affine polynomial, the results of **?** allow one to conclude that the distribution of $\boldsymbol{N}$ is a negative multinomial distribution. This multinomial distribution is fully characterized by the affine polynomial $P(\boldsymbol{z})$ and by the shape parameter L. This result is an extension of the following well known property: a mixed Poisson distribution generated by a gamma intensity is a negative binomial distribution (see, for instance, **?**, chap. 8, p. 328).

Bivariate mixed Poisson distributions correspond to the particular case $d = 2$ and will be used intensively in this paper. When $d = 2$, the affine polynomial defining the Laplace transform of (λ_1, λ_2) can be written as $P(z_1, z_2) = 1 + p_1 z_1 + p_2 z_2 + p_{12} z_1 z_2$.

[1] A polynomial $P(\boldsymbol{z})$ with respect to $\boldsymbol{z} = (z_1, \ldots, z_d)$ is said to be affine if the one variable polynomial $z_j \mapsto P(\boldsymbol{z})$ can be written as $A^{(-j)}z_j + B^{(-j)}$ (for any $j = 1, \ldots, d$), where $A^{(-j)}$ and $B^{(-j)}$ are polynomials of z_i's with $i \neq j$.

Straightforward computations allow one to express the generating function of $\boldsymbol{N} = (N_1, N_2)$ as follows:

$$G_{\boldsymbol{N}}(z_1, z_2) = \left[\frac{(1-a)(1-b)-c}{1-az_1-bz_2+(ab-c)z_1z_2}\right]^L, \tag{7}$$

where

$$\begin{aligned} a &= \frac{p_1+p_{12}}{1+p_1+p_2+p_{12}}, \\ b &= \frac{p_2+p_{12}}{1+p_1+p_2+p_{12}}, \\ c &= \frac{p_1p_2-p_{12}}{(1+p_1+p_2+p_{12})^2}. \end{aligned} \tag{8}$$

In the bivariate case, there are necessary and sufficient conditions regarding p_1, p_2, p_{12} ensuring that $[P(\boldsymbol{z})]^{-L}$ is the Laplace transform of a probability distribution defined on $[0,\infty[^2$:

$$p_1 > 0,\ p_2 > 0,\ p_{12} > 0,\ p_1p_2 - p_{12} \geq 0. \tag{9}$$

Moreover, the set of triplets (a, b, c) defined above belongs to the following set:

$$\Delta = \{(a,b,c) \in [0,1[^3\ ,\ (1-a)(1-b)-c > 0\}. \tag{10}$$

It is important to note that the set Δ defined above corresponds to the necessary and sufficient conditions for which the expression (7) is the generating function of a bivariate negative multinomial distribution (see Appendix A). In the bivariate case, the distribution of $\boldsymbol{N} = (N_1, N_2)$ is characterized by the affine polynomial coefficients (p_1, p_2, p_{12}) and the shape parameter L, or equivalently by (a, b, c, L). The appropriate parameterization depends on the application and will be discussed in Sections 4 and 5.

Of course, closed form expressions for the masses defined in (2) are generally difficult to obtain. However, in the bivariate case, a tractable expression of these probability masses is given by the following theorem:

Theorem 1 *The probability masses of a bivariate negative multinomial distribution* $\boldsymbol{N} = (N_1, N_2)$ *are*

$$\mathbb{P}(N_1 = m, N_2 = n) = a^m b^n\ [(1-a)(1-b)-c]^L \sum_{k=0}^{\min(m,n)} C_{L,k}^{m,n} \left(\frac{c}{ab}\right)^k,$$

for $(m, n) \in \mathbb{N}^2$*, where*

$$C_{L,k}^{m,n} = \frac{(L)_k}{k!}\frac{(L+k)_{m-k}}{(m-k)!}\frac{(L+k)_{n-k}}{(n-k)!},$$

and $(p)_k$ *is the Pochhammer symbol such that* $(p)_0 = 1$ *and* $(p)_{k+1} = (p+k)(p)_k$ *for any positive integer* k*.*

The proof of this theorem is given in Appendix B. Note that this result allows one to obtain tractable expressions for the joint probabilities of the pair (N_k, N_l), $1 \leq k < l \leq d$ associated to an MMPD $\boldsymbol{N}$. This property will be used for estimating the parameters of MMPDs using an MPL method. It is also interesting to note that similar derivations could be used to derive higher order marginal distributions of $\boldsymbol{N}$. However, the masses of $N_{k_1}, ..., N_{k_l}$ with $(k_1, ..., k_l) \in \mathbb{N}^l$, $l > 2$, are expressed as functions of $(l-1)$-dimensional summations whose computational complexity is an increasing function of l.

3 Maximum Pairwise likelihood method

Let $\boldsymbol{N}^i = (N_1^i, \ldots, N_d^i)$, $i = 1, \ldots, n$, be an independent sample of the d-multivariate random vector $\boldsymbol{N}$ distributed according to an MMPD generated with a multivariate Gamma mixing distribution $\gamma_{L,P}$. We assume that the affine polynomial P is parameterized by an unknown parameter vector $\boldsymbol{\theta_0}$. The definition of $\boldsymbol{\theta_0}$ is very problem dependent and will be explained carefully in Section 4. We denote by $p(\boldsymbol{n}, \boldsymbol{\theta_0})$, $\boldsymbol{n} \in \mathbb{N}^d$, the joint probability of $\boldsymbol{N}$, and by $p_{k,l}(n_k, n_l, \boldsymbol{\theta_0})$, $(n_k, n_l) \in \mathbb{N}^2$, $1 \leq k < l \leq d$, the joint probability of the pair (N_k, N_l). This section studies an MPLE of $\boldsymbol{\theta_0}$ based on the n-sample $(\boldsymbol{N}^1, \ldots, \boldsymbol{N}^n)$. After recalling the principle of MCL methods, we establish the asymptotic properties of the resulting estimator.

3.1 MCL methods

MCL methods are interesting estimation methods which can be used when the standard maximum likelihood estimator (MLE) is difficult to implement. To construct a CL, one starts with a set of conditional or marginal events for which the likelihood is tractable. The choice of these events is motivated by the following two points: 1) the CL method must identify all the parameters, and 2) the loss of efficiency of the MCL estimators should be acceptable and balanced by the computational ease.

Since it is difficult here to have a tractable expression of the joint masses $p(\boldsymbol{n}, \boldsymbol{\theta})$ in terms of $\boldsymbol{\theta}$ for $\boldsymbol{n} \in \mathbb{N}^d$, we propose to estimate $\boldsymbol{\theta}$ by using the probabilities of the pairs (N_k, N_l), for $1 \leq k < l \leq d$. These probabilities have tractable expressions provided by Theorem 1. We define the pairwise log-likelihood (PL) of the random vector $\boldsymbol{N}$ as

$$l(\boldsymbol{n}, \theta) = \sum_{1 \leq k < l \leq d} \log p_{k,l}(n_k, n_l, \boldsymbol{\theta}). \tag{11}$$

The MPLE $\widehat{\boldsymbol{\theta}}_n$ is the value of $\boldsymbol{\theta}$ which minimizes

$$U_n(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{N}^i, \boldsymbol{\theta}). \tag{12}$$

Applications of MPL methods are numerous in multivariate statistics. These applications include the analysis of correlated binary data (**??**), binary spatial data (**?**) and random set models for binary images (**?**). More recent applications include serially correlated count data (**?**), estimation of recombination rates from pairs of loci in gene sequences (**?**), stochastic geometry for a variety of spatial point process (**?**) and analysis of ordinal categorical time series (**?**). **?** also considered the case of a fixed sample size n and provided conditions for the consistency of the MPL estimators when the dimension d of the vectors, and thus the number of pairs, increases. Note that these conditions are not satisfied in our application where the vector size is fixed (to the number of images) and where the sample size increases (when the size of the estimation window increases).

Many other CL functions have been considered in the statistical literature (**???**). These CL include the composite marginal log-likelihood and the pseudo log-likelihood whose main properties are recalled below.

– The composite marginal log-likelihood

$$l^{\text{marg}}(\boldsymbol{n};\theta) = \sum_{1\leq j\leq n} \log \Pr(N_j = n_j).$$

is the sum of the log-likelihoods associated to the univariate marginal distributions. The composite marginal log-likelihood is generally easy to compute. It corresponds to the full likelihood when the different components of $\boldsymbol{N}$ are independent. Consequently, this CL does not contain any information regarding the dependence structure between the marginal distributions of $\boldsymbol{N}$. The composite marginal log-likelihood is not appropriate to the change detection problem since we are precisely trying to estimate correlations between the pixels of different images. An hybrid method based on both univariate composite marginal and pairwise log-likelihoods was recently proposed in **?**. A two-stage iterative procedure was proposed for estimating jointly the parameters of the marginal distributions, and the parameters associated to the correlation structure between the pairs. This method improved the performance of the marginal parameter estimators with respect to the corresponding pairwise likelihood estimators. However, no significant improvement was observed for the correlation parameters. This hybrid approach was not considered in this paper since we are precisely trying to estimate the correlation coefficients for image change detection.

– The pseudo log-likelihood: **?** introduced another famous variety of composite log-likelihood, often referred to as *Besag's pseudo log-likelihood* or just *pseudo log-likelihood*, defined by:

$$l^{\text{Besag}}(\boldsymbol{n};\theta) = \sum_{1\leq j\leq n} \log \mathbb{P}(N_j = n_j|\boldsymbol{N}_{[j]}),$$

where $\boldsymbol{N}_{[j]}$ denote all the components of $\boldsymbol{N}$ except the jth one. The probability $\mathbb{P}(N_j = n_j|\boldsymbol{N}_{[j]})$ provides the distribution of the jth component of $\boldsymbol{N}$ conditioned upon the other components of $\boldsymbol{N}$. This composite log-likelihood has been introduced for the analysis of lattice data and has received much attention for Markov random fields (see for instance **?**). In the case of Markov random fields, the pseudo-likelihood has a tractable closed-form expression, up to a normalizing constant. The original pseudo-likelihood has been extended to spatial point processes in **?** and **?**. Pseudo-likelihood estimators for spatial point processes have then been studied from both a theoretical (**??**) and a practical (**??**) point of view. However there is no simple tractable expression for the pseudo-likelihood of MMPDs (contrary to the pairwise log-likelihood), which precludes its use for our image change detection problem.

Based on the above discussion, the rest of this paper focuses on the MPLE for the parameters of MMPDs.

3.2 Asymptotic properties

This section studies the consistency and asymptotical normality of the MPLE $\widehat{\boldsymbol{\theta}}_n$ for the model introduced above, i.e. for an MMPD with multivariate Gamma mixing distribution parameterized by L and $\boldsymbol{\theta_0}$. These asymptotic properties are derived in the

particular case where L is known. This assumption is in agreement with the image processing application considered in Section 5.

Assumptions

1. The space parameter Θ is a compact subset of $\mathbb{R}^p$. The point $\boldsymbol{\theta_0}$ belongs to the interior of the space Θ,
2. Let $F_{k,l}$ be functions from Θ to $\Delta = \{(a,b,c) \in [0,1[^3;\ (1-a)(1-b) > c\}$ that give the relation between $\boldsymbol{\theta}$ and $(a_{k,l}, b_{k,l}, c_{k,l})$, $1 \leq k < l \leq d$. The function $F(\boldsymbol{\theta}) = \left(F_{1,2}(\boldsymbol{\theta})^T, \ldots, F_{d-1,d}(\boldsymbol{\theta})^T\right)^T$ is an injective map from Θ to $\Delta^{d(d-1)/2}$.
3. The functions $F_{k,l}$ are twice continuously differentiable,

Theorem 2 *The maximum pairwise log-likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ converges almost surely to $\boldsymbol{\theta_0}$. Furthermore $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ converges in distribution to a centered normal distribution with covariance matrix equal to $I_U(\boldsymbol{\theta_0})^{-1}\Gamma_U(\boldsymbol{\theta_0})I_U(\boldsymbol{\theta_0})^{-1}$, where*

$$I_U(\boldsymbol{\theta_0})_{u,v=1,\ldots,p} = -\sum_{1\leq k<l\leq d} \mathbb{E}_{\boldsymbol{\theta_0}}\left(\frac{\partial}{\partial\theta_u}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\frac{\partial}{\partial\theta_v}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\right),$$

and

$$\Gamma_U(\boldsymbol{\theta_0})_{u,v=1,\ldots,p} =$$

$$\mathbb{E}_{\boldsymbol{\theta_0}}\left(\sum_{1\leq k<l\leq d}\frac{\partial}{\partial\theta_u}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\sum_{1\leq r<s\leq d}\frac{\partial}{\partial\theta_v}\log p_{r,s}(N_r,N_s,\boldsymbol{\theta_0})\right),$$

and where the subscript U means that the corresponding matrices depend on the negative pairwise log-likelihood defined in (12).

Note that the matrix $-I_U(\boldsymbol{\theta_0})$ is the sum of Fisher information matrices associated to the pairs (N_k, N_l). For the MLE of $\boldsymbol{\theta_0}$, the matrix $\Gamma_U(\boldsymbol{\theta_0})$ reduces to $I_U(\boldsymbol{\theta_0})$. However, this is not the case for the proposed MPLE. Theorem 2 has been proved by showing that the first and second order moments of a bivariate negative multinomial distribution exist and are finite, and by using the results of **?** (see Appendix C). An alternative to prove Theorem 2 would be to use the results of **?** and **?** for negative multinomial distributions.

4 Effectiveness of the proposed MPL method

Many simulations have been conducted to validate the previous theoretical results. This section studies the performance of the MPLE of $\boldsymbol{\theta_0}$ for synthetic data.

4.1 Generation of MMPDs

We first consider MMPDs which have been used to model longitudinal count data on patient-controlled analgesia in **?**. However, it is important to note that the asymptotic properties of the resulting MPLE were not provided in **?**. The generation of random vectors distributed according to MMPDs has been performed as follows:

– Simulate $2L$ independent multivariate centered Gaussian vectors of $\mathbb{R}^d$ denoted as $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^{2L}$ with the following $d \times d$ covariance matrix:

$$\boldsymbol{\mathcal{C}} = \left(c_{i,j}\right)_{1 \leq i,j \leq d} = \frac{\sigma}{2}\left(\rho^{\frac{|i-j|}{2}}\right)_{1 \leq i,j \leq d},$$

where $\sigma/2$ is the variance of each component of $\boldsymbol{X}^i$ ($\sigma > 0$) and ρ is the correlation coefficient between any pair of components extracted from $\boldsymbol{X}^i$.

– Compute the kth component of the intensity vector as $\lambda_k = \sum_{1 \leq i \leq 2L} (X_k^i)^2$, where X_k^i is the kth component of $\boldsymbol{X}^i$.

The random vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ generated above is distributed according to a multivariate Gamma distribution whose margins are univariate Gamma distributions $\gamma_{L,\sigma}$. Moreover the pair (λ_k, λ_l) is distributed according to a bivariate gamma distribution with shape parameter L and the following scale parameter:

$$P_{k,l}^{\mathcal{C}}(z_k, z_l) = 1 + \sigma z_k + \sigma z_l + \sigma^2(1 - \rho^{l-k}) z_k z_l, \qquad 1 \leq k < l \leq d. \tag{13}$$

Indeed, since the vectors $\boldsymbol{X}^i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\mathcal{C}})$ are mutually independent for all $1 \leq i \leq 2L$, a classical result (see for instance **?**) states that the matrix

$$\boldsymbol{A} = \sum_{i=1}^{2L} \boldsymbol{X}^i \left(\boldsymbol{X}^i\right)^T,$$

is distributed according to a Wishart distribution with Laplace transform

$$\mathcal{L}_{\boldsymbol{A}}(\boldsymbol{S}) = \mathbb{E}\left(e^{-\mathrm{tr}(\boldsymbol{S}\boldsymbol{A})}\right) = \det\left(\boldsymbol{I}_d + 2\,\boldsymbol{S}\boldsymbol{\mathcal{C}}\right)^{-L}, \tag{14}$$

for all symmetric matrix $\boldsymbol{S}$ such that $\boldsymbol{I}_d + 2\boldsymbol{S}\mathcal{C}$ is definite positive, where $\boldsymbol{I}_d$ is the identity matrix of size $d \times d$ and $\mathrm{tr}\,(\cdot)$ is the matrix trace. By noting that the vector $\boldsymbol{\lambda}$ is the diagonal of the Wishart matrix $\boldsymbol{A}$ and by using the relation:

$$\mathrm{tr}\left(\boldsymbol{S}_{\boldsymbol{z}}\boldsymbol{A}\right) = \sum_{i=1}^{d} z_i A_{ii} = \boldsymbol{z}^T \boldsymbol{\lambda},$$

where $\boldsymbol{S}_{\boldsymbol{z}}$ denotes the following $d \times d$ diagonal matrix:

$$\boldsymbol{S}_{\boldsymbol{z}} = \begin{pmatrix} z_1 & & \\ 0 & \ddots & 0 \\ & & z_d \end{pmatrix},$$

the Laplace transform of $\boldsymbol{\lambda}$ can be finally expressed as:

$$\mathbb{E}\left[e^{-\boldsymbol{z}^T\boldsymbol{\lambda}}\right] = \left[\det\left(\boldsymbol{I}_d + 2\,\boldsymbol{\mathcal{C}}\boldsymbol{S}_{\boldsymbol{z}}\right)\right]^{-L}.$$

The multilinearity property of the determinant ensures that the function $\boldsymbol{z} \mapsto \det(\boldsymbol{I}_d + 2\,\boldsymbol{C}\boldsymbol{S}_{\boldsymbol{z}})$ defines an affine polynomial with respect to $\boldsymbol{z}$, denoted as $P^{\boldsymbol{C}}(\boldsymbol{z})$. Consequently, thanks to the definition (3), the vector $\boldsymbol{\lambda}$ is distributed according to a multivariate gamma distribution with shape parameter L and scale parameter $P^{\boldsymbol{C}}$. Furthermore, the distribution of (λ_k, λ_l) is a bivariate gamma distribution with shape

parameter L. The corresponding affine polynomial denoted as $P^{\boldsymbol{C}}_{k,l}$ is obtained by setting to zero all the z_i's such that $i \neq k, l$ in $P^{\boldsymbol{C}}(\boldsymbol{z})$. By expanding the determinant all along its columns $i \neq k, l$, the following result is obtained:

$$P^{\boldsymbol{C}}_{k,l}(z_k, z_l) = \begin{vmatrix} 1+\sigma z_k & \sigma\rho^{(l-k)/2} z_l \\ \sigma\rho^{(l-k)/2} z_k & 1+\sigma z_l \end{vmatrix} = 1 + \sigma z_k + \sigma z_l + \sigma^2(1-\rho^{l-k}) z_k z_l, \quad (15)$$

for all $1 \leq k < l \leq d$. Note that the first moments of $(X_1, X_2) \sim \gamma_{L,P}$ can be obtained as follows:

$$\begin{aligned} &\mathbb{E}(X_i) = Lp_i, \quad \mathsf{Var}\,(X_i) = Lp_i^2, \qquad \text{for } i \in \{1, 2\}, \\ &\mathsf{Cov}\,(X_1, X_2) = L\,(p_1 p_2 - p_{12})\,. \end{aligned} \quad (16)$$

These last properties and (15) can be used to show that the covariance between λ_k and λ_l is $\mathrm{cov}(\lambda_k, \lambda_l) = L\sigma^2\rho^{l-k}$ for all $1 \leq k < l \leq d$. It is then possible to generate the MMPD vector $\boldsymbol{N}$ conditionally upon $\boldsymbol{\lambda}$, since $\boldsymbol{N}|\boldsymbol{\lambda} \sim \mathcal{P}(\boldsymbol{\lambda})$.

4.2 Estimation (known shape parameter L)

The MMPDs introduced in the previous section are parameterized by the shape parameter L and by $\boldsymbol{\theta} = (\sigma^2, \rho)^T$. This section assumes that the shape parameter L is known. This is a classical assumption in synthetic aperture radar (SAR) imagery since L corresponds to the so-called number of looks which is known by the radar (see for instance **?**, p. 93). When L is known, the convergence and asymptotic normality of the maximum pairwise log-likelihood estimator of $\boldsymbol{\theta} = (\sigma^2, \rho)^T$ are guarantied by Theorem 2. Indeed, there is a functional relation between $\boldsymbol{\theta} = (\sigma^2, \rho)^T$ and $(a_{k,l}, b_{k,l}, c_{k,l})$ denoted as $F_{k,l}(\boldsymbol{\theta}) = (a_{k,l}, b_{k,l}, c_{k,l})^T$, where

$$\begin{aligned} a_{k,l} = b_{k,l} &= \frac{\sigma + \sigma^2\left(1-\rho^{l-k}\right)}{1 + 2\sigma + \sigma^2\left(1-\rho^{l-k}\right)}, \\ c_{k,l} &= \frac{\sigma^2\rho^{l-k}}{\left(1 + 2\sigma + \sigma^2\left(1-\rho^{l-k}\right)\right)^2}. \end{aligned} \quad (17)$$

For all $\boldsymbol{\theta} = (\sigma^2, \rho) \in \Xi =]0, +\infty[\times[0, 1[\subset \mathbb{R}^p$, the function $F_{k,l}(\boldsymbol{\theta}) = (a_{k,l}, b_{k,l}, c_{k,l})^T$ takes its values in Δ defined in (10). Moreover from (17), it is easy to show that $F_{k,l}$ is a twice continuously differentiable injective map from Ξ to Δ for all $1 \leq k < l \leq d$.

Note that the MPLE defined by (12) and (11) (which will be used in our simulations) corresponds to a uniform weighting between the different pairwise log-likelihoods. However, it would be possible to introduce a set of weights $\left(\omega_{k,l}\right)_{1 \leq k < l \leq d}$ modifying the pairwise log-likelihood as follows:

$$l(\boldsymbol{n}, \boldsymbol{\theta}) = \sum_{1 \leq k < l \leq d} \omega_{k,l} \log p_{k,l}(n_k, n_l, \boldsymbol{\theta}).$$

Such weighting can be recommended to mitigate the influence of pairs between non-neighboring observations (which should be less informative on the correlation structure in the framework of spatial data). This strategy might also reduce the optimization complexity in particular applications. **?** have proposed an optimal weighting strategy

for the MPL method. These weights depend on the theoretical values of the parameters and thus have to be estimated. This estimation is achieved by a bootstrap based algorithm due to **?**, namely the window subsampling method. However such method appears computationally prohibitive in our image processing applications, since it has to be applied to each pixel of the image (whose size is 200×100 in our simulations). As a consequence, only simple weighting strategies are investigated in this section. Moreover, we will show that weighting does not provide significant performance improvement in our simulations, especially for the change detection problem. The optimization procedure used to minimize the negative pairwise log-likelihood is the direct geometrical Nelder Mead Simplex method (MATLAB function fminsearch.m).

In order to appreciate the interest of the proposed MPL method, the unknown parameters σ^2 and ρ have also been estimated by the classical method of moments. This method is based on the following equations, derived from the expression of MMPD moments as function of the intensity moments (**?**) and equations (15,16):

$$\begin{aligned} \mathbb{E}[N_i] &= L\sigma, & \forall\ 1 \leq i \leq d, \\ \mathsf{Cov}(N_i, N_j) &= L\sigma\rho^{|i-j|}, & \forall\ 1 \leq i \neq j \leq d. \end{aligned}$$

The first equation allows us to estimate σ whereas the parameter ρ can be estimated from the covariances $\mathsf{Cov}(N_i, N_j)$. Note that several methods of moments have been implemented to estimate ρ. Methods of moments based on all the pairs $(N_i, N_j)_{1 \leq i < j \leq d}$ do not yield better estimation than estimates constructed only from the lag-one pairs $(N_i, N_{i+1})_{1 \leq i \leq d-1}$. This can be explained by the fact that non-neighboring observations are less informative in our model. As a result, giving too much importance to non-neighboring pairwise leads to bad estimations. An alternative is to compute a weighted least squares estimator, whose weights are defined from the inverse covariance matrix of first and second order moments (the reader is invited to consult **?**, for more details). However, no significant improvement has been observed with this strategy. As a consequence, this paper will focus on the moment estimator based on lag-one pairs $(N_i, N_{i+1})_{1 \leq i \leq d-1}$.

The empirical bias, standard deviations ("std") and mean square errors (MSEs) of the estimated parameters σ^2 and ρ are reported in Table 1 for a correlation structure $\rho = 0.8$ and for different sample sizes n. The number of Monte Carlo runs is 1000. The other parameters for this example are $\sigma^2 = 2$, $L = 4$ (shape parameter) and $d = 12$ (dimension of the observations). Figures 1 and 2 also show the log MSEs of the estimated parameters σ^2 and ρ as a function of the logarithm of the sample size n (logarithmic scales are preferred since the log MSEs are classically linear functions of $\log(n)$). The circle curve corresponds to the estimator of moments whereas the triangle curve corresponds to the MPLE. Figure 1 shows that the performances obtained with both methods are similar for the estimation of σ^2. However the MPL approach is much more efficient for the estimation of ρ as illustrated in fig. 2. The theoretical asymptotic log variances of the MPLE provided by Theorem 2 are also displayed in Figures 1 and 2. Note that all mathematical expectations appearing in this theorem have been computed by Monte-Carlo averages for the true values of the parameters. These theoretical asymptotic variances are clearly in good agreement with the empirical MSEs, computed from 1000 Monte Carlo runs, for large values of n.

The frequency polygon of the estimates $\hat{\rho}$ and $\widehat{\sigma^2}$ are displayed in Figures 3 and 4, as well as the theoretical asymptotic distribution (dashed line) and 95% confidence intervals. The frequency polygon is based on the histogram estimates obtained from

1000 Monte Carlo runs with 50 bins (the other parameters are $n = 5000$ and $\rho = 0.8$). This polygon connects the midpoints at the top of the bars of the histogram with line segments. Confidence intervals are obtained by noting that the number of estimates belonging to each bin of the histogram is distributed according to a binomial distribution $\mathcal{B}(N, p)$, where N is the total number of estimates (i.e. the number of Monte-Carlo runs in this simulation) and p is the theoretical probability that an estimate belongs to the considered bin. By using the theoretical asymptotic normality of the MPLE (see Theorem 2), confidence bounds are then obtained for each bin thanks to the Clopper-Pearson expression for calculating exact binomial confidence intervals (**?**). These figures show that the asymptotic Gaussian distribution derived in Theorem 2 is very close to its estimation for this sample size.

The last experiments study the performance of the MPLE as a function of the number of neighbors considered in the PL. Figure 5 shows the logarithm of the MPLE asymptotic theoretical variance versus the maximal lag τ when the likelihoods of the following neighboring pairs $(N_i, N_j)_{|i-j|\leq\tau}$ are considered in the PL. Considering these lags is equivalent to introduce dummy weights: these weights are zero for the non-neighboring pairs, and 1 for the neighboring pairs. As expected, the MPLE performance for ρ first increases when the number of considered pairs increases. However, the performance slightly decreases after an extremum (obtained for $\tau = 5$ in this simulation). This simulation emphasizes that non-neighboring observations are less informative in our model and can deteriorate the estimation performance. Moreover the gain obtained in using only neighboring observations is not very important and should be balanced with the computational cost to estimate the optimal set of weights, as proposed for example in **?**.

4.3 Estimation (unknown shape parameter L)

This section presents some simulation results obtained for the joint estimation of $\boldsymbol{\theta} = (L, \sigma^2, \rho)^T$. Note that Theorem 2 does not apply here since the shape parameter L is unknown. As previously, to appreciate the interest of the proposed MPL method, the unknown parameters σ^2, ρ and L have also been estimated by the classical method of moments. This method is based on the following equations, derived from the expression of MMPD moments as function of the intensity moments (**?**) and equations (15,16):

$$\begin{aligned}
\mathbb{E}[N_i] &= L\sigma, && \forall\ 1 \leq i \leq d,\\
\mathsf{Var}[N_i] &= L\sigma(1+\sigma), && \forall\ 1 \leq i \leq d,\\
\mathsf{Cov}(N_i, N_j) &= L\sigma^2\rho^{|i-j|}, && \forall\ 1 \leq i \neq j \leq d.
\end{aligned}$$

The first and second equations allow us to estimate L and σ whereas the parameter ρ can be estimated from the covariances $\mathsf{Cov}(N_i, N_j)$. This section focuses on the lag-one pairs $(N_i, N_{i+1})_{1\leq i\leq d-1}$ as previously.

Figures 6, 7 and 8 show the MSEs of the estimated parameters σ^2, ρ and L, for a correlation structure $\rho = 0.8$, as a function of the sample size n. The number of Monte Carlo runs is 1000. The other parameters for this example are $L = 4$, $\sigma^2 = 2$ and $d = 12$. The circle curve corresponds to the estimator of moments whereas the triangle curve corresponds to the MPLE. The empirical bias, standard deviations ("std") and MSEs are also reported in Table 2. These results illustrate the interest of the MPL

approach, which is much more efficient for this problem than the method of moments for the three parameters σ^2, ρ and L.

Note that the optimization procedure used for the maximization of the PL does not yield necessarily integer values for L. However, it has been observed that non integer values of L can be appropriate when the averaged images (looks) are correlated. This remark has even motivated the definition of an equivalent number of looks (see **?**, p. 95). The proposed estimation strategy (which allows one to estimate the parameter L) can be useful in this context.

5 Application to Change detection in real radar images

5.1 Change Detection Problem

This section considers a fundamental problem in image processing referred to as change detection problem. Consider several co-registered images acquired at different dates before and after a change, here a natural disaster. The objective of change detection is to produce a map representing the changes affecting the scene due to this natural disaster. This paper considers three one look (i.e. $L = 1$) 200×100 low-flux images displayed in Fig. 9: a reference image I of the Nyaragongo volcano in Congo before an eruption and two secondary images J and K of the same scene acquired after the eruption. Figure 9(d) indicates the pixels of the image which have been affected by the eruption (white pixels). These images have been obtained from real power radar images corresponding to low-flux scenarios. Low-flux scenarios correspond to very short exposure times or images with low intensity objects (to be detected). In this case, the image intensities cannot be measured directly. Thus, the observed data are the numbers of photons collected at each pixel of the image (**?**). The distribution of these numbers of photons is classically a mixed Poisson distribution. In the case of power radar images, it is well known that the intensities are marginally distributed according to gamma distributions (**?**, p. 95). Therefore, multivariate gamma distributions seem good candidates to model the distribution of intensities collected at a given location in the three images (see **??**). By using this multivariate gamma distribution as mixing distribution in (2), the joint distribution of the numbers of photons received in the three images at a given location is an MMPD whose margins are negative multinomial distributions according to Section 2.

Change detection algorithms produce an indicator of change for each pixel location. For each pixel location, we observe three numbers of photons denoted as (N_I, N_J, N_K), where N_I is the number of photons corresponding to the reference image I, and (N_J, N_K) are the numbers of photons corresponding to the secondary images J and K potentially affected by the disaster. The detection of a change at a given pixel location is classically achieved by the following binary hypothesis test (**?**):

$$\begin{array}{ll} H_0 & \text{(absence of change)}, \\ H_1 & \text{(presence of change)}, \end{array} \tag{18}$$

where H_0 is the null hypothesis and H_1 the alternative hypothesis. The images J and K have been both registered after the volcano eruption. Thus, it is natural to assume that the correlation coefficients between the reference image I and the secondary images J and K, denoted as r_{IJ} and r_{IK}, are equal, i.e. $r_{IJ} = r_{IK} = r$. The presence of a

change (hypothesis H_1) at a given pixel location can then be detected by comparing the estimated correlation coefficient r to an appropriate threshold t. More precisely, the change detection strategy for a given pixel location can be written

$$H_0 \text{ rejected if } \widehat{r} < t, \tag{19}$$

where $\widehat{r}$ denotes the estimated correlation coefficient and t is a threshold depending of the significance level of the test (also referred to as probability of false alarm in image processing). As a consequence, the change detection problem mainly consists of estimating the correlation coefficient locally for each pixel position. Since only one pixel is available for each image at a given location, the images are supposed to be locally stationary and ergodic, allowing us to make estimates using several neighboring pixels. This neighborhood is the so-called estimation window. A classical assumption is that the neighbors of a given pixel are independent and have the same statistical properties. If we denote as $\boldsymbol{N}^i = (N_I^i, N_J^i, N_K^i)$ the numbers of photons of the three images corresponding to the location i, we want to estimate the correlation coefficient r from n independent triplets $\boldsymbol{N}^i$, $i = 1, .., n$ belonging to the estimation window. The stationarity and ergodicity assumptions are valid for small estimation windows. On the other hand, robust statistical estimates require a high number of samples. Therefore, the key point of the estimation of the correlation coefficient r is to perform high quality estimates with a small number of samples n belonging to the estimation window. This section proposes to estimate r from pixels belonging to the estimation window using the MPLE strategy studied in this paper.

5.2 Statistical model for $\boldsymbol{N} = (N_I, N_J, N_K)$

The intensity vector $\boldsymbol{\lambda} = (\lambda_I, \lambda_J, \lambda_K)^T$ is supposed to be distributed according to a multivariate gamma distribution whose Laplace transform can be written:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\lambda}}(z_I, z_J, z_K) = (1 + p_I z_I \; + p_J z_J + p_K z_K + p_{IJ} z_I z_J \\ + p_{IJ} z_I z_K + p_{JK} z_J z_K + p_{IJK} z_I z_J z_K)^{-L},\end{aligned} \tag{20}$$

(here $L = 1$). Straightforward computations allow one to express the correlation coefficient between the images l and m (denoted as r_{lm}) as functions of p_l, p_m and p_{lm}

$$r_{lm} = 1 - \frac{p_{lm}}{p_l p_m}, \tag{21}$$

where $(l, m) \in \{(I, J), (I, K), (J, K)\}$. Thus the correlation between the images l and m is controlled by the parameter $p_{lm} = p_l p_m (1 - r_{lm})$. As explained previously, the images J and K have been both registered after the volcano eruption. Thus, it is natural to assume that the correlation coefficients between the reference image I and the secondary images J and K are equal, i.e. $r_{IJ} = r_{IK} = r$. Moreover the Laplace transform of the pair (N_J, N_K) can be obtained by setting $z_I = 0$ in the joint Laplace transform (20). It shows that the distribution of the pair (N_J, N_K) only depends on the parameters p_J, p_K and r_{JK}. Since this distribution does not depend on r, which is the parameter of interest for our change detection problem, this pair is not taken into account in the PL. Therefore, the studied PL is formed by the likelihood of the two pairs (N_I, N_J) and (N_I, N_K). This is equivalent to introducing a dummy weight

$w_{JK} = 0$ in the PL. The advantage of this strategy is to reduce the computational cost of the PL evaluation.

The previous statistical model implies that the pairwise distributions of the intensity vector $\boldsymbol{\lambda}$ are characterized by $\boldsymbol{\theta} = (p_I, p_J, p_K, r)^T$. It is important to note here that the correlation structure between the different components of $\boldsymbol{\lambda}$ are not proportional to that of an autoregressive process of order one as in **?** since the pairwise correlation coefficient are identical ($r_{IJ} = r_{IK} = r$). Moreover, this remark emphasizes that the two pairs (N_I, N_J) and (N_I, N_K) have the same importance in order to estimate the parameter r. Therefore a weighting strategy controlling the contributions of each pair in the pairwise likelihood should not improve the estimation performance.

According to Section 2, the joint probabilities of the two pairs (N_I, N_J) and (N_I, N_K) associated to the MMPD vector $\boldsymbol{N} = (N_I, N_J, N_K)^T$ (whose multivariate mixing Gamma distribution has been described above) are distributed according to bivariate negative multinomial distributions having the same shape parameter L. The parameters of the affine polynomial corresponding to the pairs (N_l, N_m), with $(l, m) \in \{(I, J), (I, K)\}$ can be expressed as follows:

$$\big(a_{l,m}, b_{l,m}, c_{l,m}\big)^T = F_{lm}(\boldsymbol{\theta}), \tag{22}$$

where

$$\begin{aligned} a_{l,m} &= \frac{p_l + p_l p_m (1 - r_{l,m})}{1 + p_l + p_m + p_l p_m (1 - r_{lm})}, \\ b_{l,m} &= \frac{p_m + p_l p_m (1 - r_{l,m})}{1 + p_l + p_m + p_l p_m (1 - r_{lm})}, \\ c_{l,m} &= \frac{r_{lm}}{(1 + p_l + p_m + p_l p_m (1 - r_{lm}))^2}, \end{aligned} \tag{23}$$

and where $\boldsymbol{\theta} = (p_I, p_J, p_K, r)^T$ is the parameter vector to be estimated. Note that for all $\boldsymbol{\theta} \in \Xi =]0, +\infty[^3 \times [0, 1[$, the function $F(\boldsymbol{\theta}) = (F_{IJ}(\boldsymbol{\theta})^T, F_{IK}(\boldsymbol{\theta})^T, F_{JK}(\boldsymbol{\theta})^T)^T$ takes its values in Δ^2, where Δ is defined in (10). Then from (23), it is easy to show that F is a twice continuously differentiable injective map from Ξ to Δ^3. The convergence and asymptotic normality of the MPLE of $\boldsymbol{\theta}$ are then guarantied by the Theorem 2.

5.3 Performance of change detection algorithms

In order to appreciate the performance of the detector based on the MPLE of r, denoted as $\widehat{r}_{\mathrm{MPLE}}$, estimators based on the method of moments are also investigated. More precisely, we consider the following classical estimator based on empirical averages:

$$\widehat{r}_{\mathrm{MOM}} = \frac{1}{2}\left(\frac{\sum_{i=1}^n N_I^i N_J^i - \overline{N}_I \overline{N}_J}{\sqrt{\sum_{i=1}^n (N_I^i)^2 - \overline{N}_I^2}\sqrt{\sum_{i=1}^n (N_J^i)^2 - \overline{N}_J^2}} + \frac{\sum_{i=1}^n N_I^i N_K^i - \overline{N}_I \overline{N}_K}{\sqrt{\sum_{i=1}^n (N_I^i)^2 - \overline{N}_I^2}\sqrt{\sum_{i=1}^n (N_K^i)^2 - \overline{N}_K^2}} \right),$$

where n is the size of the estimation window and $\overline{N}_k = \frac{1}{n}\sum_{i=1}^n N_k^i$ is the sample mean, for $k = I, J, K$.

The MLE of the correlation coefficient based on only two images, I and J, is also studied in order to appreciate the gain obtained by using 3 images instead of 2. In the case of two images, the likelihood reduces to the product of the bivariate masses associated with the pairs $(N_I^i, N_J^i)_{1\leq i\leq n}$. The MLE of r based on two images can be easily computed by a numerical optimization since a tractable expression of the bivariate masses is available. It is important to note that the log-likelihood based on the two images I and J is the term associated with the pair (N_I, N_J) in the PL based on the three images I, J and K. As a consequence, this bivariate log-likelihood can be seen as a special case of the PL when the weights associated to the pairs (N_I, N_K) and (N_J, N_K) are zeros.

The detection performance obtained for the three estimators of r analyzed here is studied in terms of their receiver operating characteristics (ROCs). The ROCs express the power of the test (also referred to as probability of detection) π as a function of the significance level α (**?**, p. 38) where:

$$\begin{aligned} \pi =& \mathbb{P}\left[\text{rejecting } H_0 \mid H_1 \text{ is true}\right], \\ \alpha =& \mathbb{P}\left[\text{rejecting } H_0 \mid H_0 \text{ is true}\right]. \end{aligned} \tag{24}$$

The ROCs obtained for the MPLE (continuous line), the estimator based on the method of moments (dashed line) and the MLE based on two images (dots) are depicted on Figure 10 for several estimation window sizes ($n = 3 \times 3$, $n = 5 \times 5$ and $n = 7 \times 7$). It is important to mention here that the power of the test π and the level of significance α have been estimated for each value of the threshold t by counting the number of estimates $\widehat{r}$ below t for all pixels of the image associated to hypotheses H_1 and H_0 respectively. Note also that the pixels of the image have been associated to hypotheses H_1 and H_0 by using the ground truth given by the mask shown in Figure 9(d). The performances of the correlation coefficient estimators $\widehat{r}$ are reported in Table 3 for the two classes "Presence of Change" and "Absence of Change". As expected, the detector based on the MPLE provides the best performance. It is interesting to note that the gain in detection performance when using 3 images with respect to 2 images is less significant for large estimation window sizes. On the other hand, the MPLE and MLE outperform the estimators of moments in all cases. In conclusion, one reviewer mentioned that it would be interesting to extend the proposed algorithm to more sophisticated models that would account for spatial correlations among adjacent pixels of the image. The resulting algorithms might improve the change detection performance.

Acknowledgements

The authors would like to thank Gérard Letac for fruitful discussions regarding multivariate gamma distributions and André Ferrari for interesting discussions on composite likelihoods. They are also very grateful to Professor Petar Djuric from Stony Brook university for helping them to fix the English grammar. Finally, the authors would like to thank the two reviewers as well as the AE for the careful and thoughtful comments about this paper. Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24.

A Proof of conditions (9)

The necessary conditions (9) are obtained by noting that the probability masses $p_{m,n} = \mathbb{P}(N_1 = m, N_2 = n)$ expressed in (1) satisfy $0 \leq p_{m,n} \leq 1$ for all positive integers m, n.

1. $p_{0,0} = [(1-a)(1-b)-c]^L$ yields $(1-a)(1-b)-c > 0$,
2. $p_{1,0} = La[(1-a)(1-b)-c]^L$ yields $a \geq 0$ and $b \geq 0$ by symmetry,
3. $p_{1,n} = [(1-a)(1-b)-c]^L \frac{(L)_n}{n!} b^{n-1}(Lab+nc)$ leads to $c \geq 0$.Indeed, for $c < 0$, $Lab + nc$ would be < 0 for large values of n.
4. Since $p_{m,0} = ((1-a)(1-b)-c)^L a^m \frac{(L)_m}{m!}$, we have for a given value of $L > 0$, $p_{m,0} > ((1-a)(1-b)-c)^L \frac{a^m}{m}$. This lower bound goes to infinity as m goes to infinity if $a > 1$. Moreover the case $a = 1$ is not possible since $c \geq 0$ and $(1-a)(1-b)-c > 0$. Thus, we have $a < 1$. Note that this last constraint implies that $b < 1$.

Proving that the conditions above are sufficient requires to show that 1) the coefficients of all the monomials $z_1^m z_2^n$ in the Taylor series (26), denoted as $c_{m,n}$, are positive and 2) their sum is equal to one. Thanks to the conditions (9), it is obvious that all the coefficients $c_{m,n}$ are positive. Moreover these conditions ensure that $\left|\frac{cz_1z_2}{(1-az_1)(1-bz_2)}\right| < 1$, $|az_1| < 1$ and $|bz_2| < 1$ for all $-1 \leq z_1, z_2 \leq 1$. Consequently, the Taylor series expansion (26) is valid for all (z_1, z_2) in $[-1,1]^2$. In particular, we obtain that $\sum_{m,n\geq 0} c_{m,n} = G_{\boldsymbol{N}}(1,1) = 1$.

B Proof of Theorem 1

From the definition of the generating function, the following Taylor series expansion with respect to the two variables z_1 and z_2 is obtained:

$$G_{\boldsymbol{N}}(z_1, z_2) = \sum_{n_1, n_2 \geq 0} \mathbb{P}(N_1 = n_1, N_2 = n_2)\ z_1^{n_1} z_2^{n_2}, \tag{25}$$

for all $-1 \leq z_1, z_2 \leq 1$. Thus the probability masses $\mathbb{P}(N_1 = n_1, N_2 = n_2)$ can be identified from the Taylor series expansion of the bivariate negative multinomial generating function (7). As $g(z_1, z_2) = \frac{cz_1z_2}{(1-az_1)(1-bz_2)}$ is a continuous function such that $g(0,0) = 0$, there exists a non empty neighborhood of $(0,0)$ denoted by U_1 such that $|g(z_1, z_2)| < 1$ for all (z_1, z_2) in U_1. Therefore, for all (z_1, z_2) in U_1, (7) yields:

$$\begin{aligned}
\left[\frac{(1-a)(1-b)-c}{1-az_1-bz_2+(ab-c)z_1z_2}\right]^L &= \left[\frac{(1-a)(1-b)-c}{(1-az_1)(1-bz_2)(1-\frac{cz_1z_2}{(1-az_1)(1-bz_2)}}\right]^L, \\
&= \left[\frac{(1-a)(1-b)-c}{(1-az_1)(1-bz_2)}\right]^L \sum_{k=0}^{\infty} \frac{(L)_k}{k!} \frac{c^k z_1^k z_2^k}{(1-az_1)^k (1-bz_2)^k}, \\
&= ((1-a)(1-b)-c)^L \sum_{k=0}^{\infty} \frac{(L)_k}{k!} \frac{c^k z_1^k z_2^k}{(1-az_1)^{(L+k)}(1-bz_2)^{(L+k)}}.
\end{aligned}$$

Similarly, there exists a non empty neighborhood of $(0,0)$, denoted as U_2, such that for all (z_1, z_2) in U_2, $|az_1| < 1$ and $|bz_2| < 1$. Therefore for all (z_1, z_2) in U_2, the following series expansions are obtained:

$$\frac{1}{(1-az_1)^{L+k}} = \sum_{r=0}^{\infty} \frac{(L+k)_r}{r!} a^r z_1^r, \quad \frac{1}{(1-bz_2)^{L+k}} = \sum_{s=0}^{\infty} \frac{(L+k)_s}{s!} b^s z_2^s.$$

As U_1 and U_2 are non empty neighborhoods of $(0,0)$, $U = U_1 \cap U_2$ is also non empty. For all (z_1, z_2) in U the following expression is finally obtained:

$$G_{\mathbf{N}}(z_1,z_2) = ((1-a)(1-b)-c)^L \sum_{k,r,s=0}^{\infty} \frac{(L)_k}{k!}\frac{(L+k)_r}{r!}\frac{(L+k)_s}{s!} a^r b^s c^k z_1^{r+k} z_2^{s+k},$$

$$= ((1-a)(1-b)-c)^L \sum_{m,n=0}^{\infty} a^m b^n \left(\sum_{k=0}^{\min(m,n)} C_{L,k}^{m,n} \left(\frac{c}{ab}\right)^k \right) z_1^m z_2^n. \qquad (26)$$

The Taylor series (26) is defined on the non empty set U. Therefore by unicity of the Taylor series expansion, the coefficients of the monomials $z_1^m z_2^n$ in (26) are the masses $\mathbb{P}(N_1 = m, N_2 = n)$.

C Proof of Theorem 2

To show consistency and asymptotical normality of the composite log-likelihood estimator, we can use more general results over minimum contrast estimators (see **??**). Let us recall that $\widehat{\boldsymbol{\theta}}_n$ is the $\boldsymbol{\theta}$ value for which $U_n(\boldsymbol{\theta})$ given by (12) is minimum. By the weak law of large numbers, as n goes to ∞, $U_n(\boldsymbol{\theta})$ converges in $\mathbb{P}_{\boldsymbol{\theta_0}}$-probability to

$$K(\boldsymbol{\theta_0}, \boldsymbol{\theta}) = -\sum_{1 \le k < l \le d} \int \log(p_{k,l}(n_k, n_l, \boldsymbol{\theta})) p_{k,l}(n_k, n_l, \boldsymbol{\theta_0}) d\mu(n_k, n_l),$$

where μ is the counting measure. When the function $\boldsymbol{\theta} \to K(\boldsymbol{\theta_0}, \boldsymbol{\theta})$ (from Θ to $\mathbb{R}^+$) has a strict minimum at $\boldsymbol{\theta} = \boldsymbol{\theta_0}$, U_n defines a contrast relative to $\boldsymbol{\theta_0}$ and K. Consequently, $\widehat{\boldsymbol{\theta}}_n$ is called a minimum contrast estimator (see **?**, p. 92). Note that minimizing $K(\cdot, \boldsymbol{\theta})$ is equivalent to minimize

$$\sum_{1 \le k < l \le d} \int \log\left(\frac{p_{k,l}(n_k, n_l, \boldsymbol{\theta_0})}{p_{k,l}(n_k, n_l, \boldsymbol{\theta})}\right) p_{k,l}(n_k, n_l, \boldsymbol{\theta_0}) d\mu(n_k, n_l).$$

By the properties of Kullback-Leibler distance, $K(\boldsymbol{\theta_0}, \boldsymbol{\theta})$ is minimum for $\boldsymbol{\theta} = \boldsymbol{\theta_0}$. Moreover, this minimum is unique if and only if

$$\mathbf{A_0}: \quad \forall k < l; \quad p_{k,l}(\cdot,\cdot,\boldsymbol{\theta}) = p_{k,l}(\cdot,\cdot,\boldsymbol{\theta_0}) \quad \text{almost everywhere (a.e)} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta_0}.$$

C.1 Consistency for minimum contrast estimator

To obtain the consistency of the minimum contrast estimator, we need the following two assumptions (see **?**, p. 93).

$\mathbf{A_1}$: Θ is a compact subset of $\mathbb{R}^p$. The functions $U_n(\boldsymbol{\theta})$ and $K(\boldsymbol{\theta_0}, \boldsymbol{\theta})$ are continuous for $\boldsymbol{\theta} \in \Theta$.

$\mathbf{A_2}$: For $\eta > 0$, let $w(n, \eta) = \sup\{|U_n(\alpha) - U_n(\beta)|;\ \|\alpha - \beta\| \le \eta\}$, where $\|\cdot\|$ is the Euclidian norm. There exists one sequence $(\varepsilon_K)_{K \in \mathbb{N}}$, decreasing to zero such that for any K :

$$\lim_{n \to \infty} \mathbb{P}_{\boldsymbol{\theta_0}} \left(w\left(n, \frac{1}{K}\right) \ge \varepsilon_K \right) = 0.$$

C.2 Asymptotical normality for the minimum contrast estimator

The following additional assumptions are required for the asymptotic normality:

$\mathbf{A_3}$: The point $\boldsymbol{\theta_0}$ belongs to the interior of the space Θ. The function $U_n(\boldsymbol{\theta})$ is twice continuously differentiable on a neighborhood V of $\boldsymbol{\theta_0}$.

$\mathbf{A_4}$: $\sqrt{n}\nabla U_n(\boldsymbol{\theta_0})$ converges in distribution to a centered normal distribution whose covariance matrix is $\Gamma_U(\boldsymbol{\theta_0})$.

$\mathbf{A_5}$: For $r > 0$ and $1 \leq u, v \leq p$,

$$1_{|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}| \leq r} \left[\int_0^1 \frac{\partial^2}{\partial\theta_u \partial\theta_v} U_n(\boldsymbol{\theta_0} + s(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}))ds - \frac{\partial^2}{\partial\theta_u \partial\theta_v} U_n(\boldsymbol{\theta_0}) \right]$$

converges in $\mathbb{P}_{\boldsymbol{\theta_0}}$-probability to zero.

$\mathbf{A_6}$: There exists an invertible matrix $I_U(\boldsymbol{\theta_0})$ such that $\left(\frac{\partial^2}{\partial\theta_u \partial\theta_v} U_n(\boldsymbol{\theta_0}) \right)_{u,v=1,\ldots,p}$ converges in $\mathbb{P}_{\boldsymbol{\theta_0}}$-probability to $I_U(\boldsymbol{\theta_0})$.

Under $\mathbf{A_{3\cdot 6}}$ and if the minimum contrast estimator is consistent, it can be shown that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ converges in distribution to a zero mean Gaussian distribution with covariance matrix $I_U(\boldsymbol{\theta_0})^{-1} \Gamma_U(\boldsymbol{\theta_0}) I_U(\boldsymbol{\theta_0})^{-1}$ (**?**, p. 104).

For contrasts of the form (12) and as for MLEs, we can replace $\mathbf{A_4}$ and $\mathbf{A_5}$ by,

$\tilde{\mathbf{A}}_\mathbf{4}$: Derivation and integration relating to μ can be permuted for $p_{k,l}$. The covariance matrix of the random vector $\sum_{1 \leq k < l \leq d} \nabla \log p_{k,l}(N_k, N_l, \boldsymbol{\theta_0})$ exist.

$\tilde{\mathbf{A}}_\mathbf{5}$: There exist some functions h_{kl} in $\mathcal{L}^1(\mathbb{P}_{\boldsymbol{\theta_0}})$ such that for all $\boldsymbol{\theta} \in V$ and $u, v = 1, \ldots, p$,

$$\left| \frac{\partial^2}{\partial\theta_u \partial\theta_v} \log p_{k,l}(n_k, n_l, \boldsymbol{\theta}) \right| \leq h_{k,l}(n_k, n_l), \ \forall (n_k, n_l) \in \mathbb{N}^2.$$

C.3 Properties of the proposed estimator

In order to prove Theorem 2, we must show that the assumptions $\mathbf{A_0}$, $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$, $\tilde{\mathbf{A}}_\mathbf{4}$, $\tilde{\mathbf{A}}_\mathbf{5}$ and $\mathbf{A_6}$ are satisfied for the proposed model under the Assumptions 1-3. Let us recall that for this model, we have (see Theorem 1)

$$p_{k,l}(n_k, n_l, \boldsymbol{\theta}) = a_{k,l}(\boldsymbol{\theta})^{n_k} b_{k,l}(\boldsymbol{\theta})^{n_l} \left((1 - a_{k,l}(\boldsymbol{\theta}))(1 - b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta}) \right)^L \times \sum_{j=0}^{\min(n_k, n_l)} C_{L,j}^{n_k, n_l} \left(\frac{c_{k,l}(\boldsymbol{\theta}))}{a_{k,l}(\boldsymbol{\theta}) b_{k,l}(\boldsymbol{\theta})} \right)^j.$$

For all $k < l$, $F_{k,l}(\boldsymbol{\theta}) = (a_{k,l}(\boldsymbol{\theta}), b_{k,l}(\boldsymbol{\theta}), c_{k,l}(\boldsymbol{\theta}))^T$ where $F_{k,l}$ are functions from Θ to $\Delta = \{(a_{k,l}, b_{k,l}, c_{k,l}) \in [0,1[^3;\ (1-a_{k,l})(1-b_{k,l}) > c_{k,l}\}$, and $F(\boldsymbol{\theta}) = \left(F_{1,2}(\boldsymbol{\theta})^T, \ldots, F_{d-1,d}(\boldsymbol{\theta})^T \right)^T$ is an injective map from Θ to $\Delta^{p(p-1)/2}$ (Assumption 2). Furthermore, the functions $F_{k,l}$ are twice continuously differentiable (Assumption 3).

$\mathbf{A_0}$: Since for all $k < l$, $p_{k,l}(\cdot, \cdot, \boldsymbol{\theta}) = p_{k,l}(\cdot, \cdot, \boldsymbol{\theta_0})$ almost everywhere, in particular we have for $(n_k, n_l) = (0, 0)$,

$$\left((1 - a_{k,l}(\boldsymbol{\theta}))(1 - b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta}) \right)^L = \left((1 - a_{k,l}(\boldsymbol{\theta_0}))(1 - b_{k,l}(\boldsymbol{\theta_0})) - c_{k,l}(\boldsymbol{\theta_0}) \right)^L,$$

for $(n_k, n_l) = (1, 0)$,

$$a_{k,l}(\boldsymbol{\theta}) \left((1 - a_{k,l}(\boldsymbol{\theta}))(1 - b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta}) \right)^L = a_{k,l}(\boldsymbol{\theta_0}) \left((1 - a_{k,l}(\boldsymbol{\theta_0}))(1 - b_{k,l}(\boldsymbol{\theta_0})) - c_{k,l}(\boldsymbol{\theta_0}) \right)^L,$$

and for $(n_k, n_l) = (0, 1)$,

$$b_{k,l}(\boldsymbol{\theta}) \left((1 - a_{k,l}(\boldsymbol{\theta}))(1 - b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta}) \right)^L = b_{k,l}(\boldsymbol{\theta_0}) \left((1 - a_{k,l}(\boldsymbol{\theta_0}))(1 - b_{k,l}(\boldsymbol{\theta_0})) - c_{k,l}(\boldsymbol{\theta_0}) \right)^L.$$

So $a_{k,l}(\boldsymbol{\theta}) = a_{k,l}(\boldsymbol{\theta_0})$, $b_{k,l}(\boldsymbol{\theta}) = b_{k,l}(\boldsymbol{\theta_0})$ and $c_{k,l}(\boldsymbol{\theta}) = c_{k,l}(\boldsymbol{\theta_0})$, *i.e.* $F_{k,l}(\boldsymbol{\theta}) = F_{k,l}(\boldsymbol{\theta_0})$ for all $1 \leq k < l \leq d$. Thus $F(\boldsymbol{\theta}) = F(\boldsymbol{\theta_0})$, which involves $\boldsymbol{\theta} = \boldsymbol{\theta_0}$ since F is an injective map.

$\mathbf{A_1}$: From Assumption 1, Θ is a compact subset of $\mathbb{R}^p$. Clearly the function $U_n(\boldsymbol{\theta})$ (as sum of continuous functions) is continuous for $\boldsymbol{\theta} \in \Theta$. For $K(\boldsymbol{\theta_0}, \boldsymbol{\theta})$, we can apply the continuity

theorem for integrals defined by a parameter (corollary of Lebesgue's dominated convergence theorem). Denoting as

$$A = \frac{a_{k,l}(\boldsymbol{\theta})^{n_k} b_{k,l}(\boldsymbol{\theta})^{n_l} \left((1-a_{k,l}(\boldsymbol{\theta}))(1-b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta})\right)^L}{(1+n_k)(1+n_l)},$$
$$B = (1+n_k)(1+n_l) \sum_{j=0}^{\min(n_k,n_l)} C_{L,j}^{n_k,n_l} \left(\frac{c_{k,l}(\boldsymbol{\theta}))}{a_{k,l}(\boldsymbol{\theta}) b_{k,l}(\boldsymbol{\theta})}\right)^j,$$

we obtain $p_{k,l}(n_k, n_l, \boldsymbol{\theta}) = AB$ that leads to $0 \leq AB \leq 1$. Due to the constraints over $(a_{k,l}, b_{k,l}, c_{k,l})$ (Assumption 2) and $(1+n_k)(1+n_l)C_{L,0}^{n_k,n_l} = (1+n_k)\frac{(L)_{n_k}}{n_k!}(1+n_l)\frac{(L)_{n_l}}{n_l!} > 1$ for all $L > 0$, we have $A < 1$ and $B > 1$. As a consequence, $\log(A) \leq \log(AB) \leq 0$, $|\log(AB)| \leq |\log(A)|$, which implies

$$\begin{aligned}
\left|\log(p_{k,l}(n_k, n_l, \boldsymbol{\theta}))\right| \leq & n_k \left|\log(a_{k,l}(\boldsymbol{\theta}))\right| + n_l \left|\log(b_{k,l}(\boldsymbol{\theta}))\right| + \\
& L \left|\log\left((1-a_{k,l}(\boldsymbol{\theta}))(1-b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta})\right)\right| + \log(1+n_k) + \log(1+n_l), \\
\leq & n_k \left(1 + \left|\log(a_{k,l}(\boldsymbol{\theta}))\right|\right) + n_l \left(1 + \left|\log(b_{k,l}(\boldsymbol{\theta}))\right|\right) + \\
& L \left|\log\left((1-a_{k,l}(\boldsymbol{\theta}))(1-b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta})\right)\right|.
\end{aligned}$$

Since the functions $F_{k,l}$ are uniformly continuous (as continuous functions over a compact set), we have

$$\left|\log(p_{k,l}(n_k, n_l, \boldsymbol{\theta}))\right| \leq C_1 n_k + C_2 n_l + C_3,$$

where C_1, C_2 and C_3 are positive constants. The dominated function is $\mathbb{P}_{\boldsymbol{\theta_0}}$-integrable since all order moments of variables N_k, $k = 1, \ldots, d$, exist. Using the continuity of the function $p_{k,l}(n_k, n_l, \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, we can conclude that $K(\boldsymbol{\theta_0}, \boldsymbol{\theta})$ is continuous for $\boldsymbol{\theta} \in \Theta$.

$\mathbf{A_2}$: Denoting as $p_{k,l}(n_k, n_l, \boldsymbol{\theta}) = g_{k,l}(\boldsymbol{\theta})\tilde{p}_{k,l}(n_k, n_l, \boldsymbol{\theta})$, where

$$g_{k,l}(\boldsymbol{\theta}) = \left((1-a_{k,l}(\boldsymbol{\theta}))(1-b_{k,l}(\boldsymbol{\theta})) - c_{k,l}(\boldsymbol{\theta})\right)^L,$$

$$\tilde{p}_{k,l}(n_k, n_l, \boldsymbol{\theta}) = \sum_{j=0}^{\min(n_k,n_l)} C_{L,j}^{n_k,n_l} a_{k,l}(\boldsymbol{\theta})^{n_k-j} b_{k,l}(\boldsymbol{\theta})^{n_l-j} c_{k,l}(\boldsymbol{\theta})^j,$$

we obtain

$$|U_n(\alpha) - U_n(\beta)| \leq \underbrace{\sum_{1\leq k<l\leq d} \left|\log\left(\frac{g_{k,l}(\alpha)}{g_{k,l}(\beta)}\right)\right|}_{P_1} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \sum_{1\leq k<l\leq d} \left|\log\left(\frac{\tilde{p}_{k,l}(N_k^i, N_l^i, \alpha)}{\tilde{p}_{k,l}(N_k^i, N_l^i, \beta)}\right)\right|}_{P_2}.$$

The first quantity P_1 is composed of continuous functions over the compact set Θ and consequently is uniformly continuous. Thus, for $\|\alpha - \beta\| \leq \frac{1}{K}$, there exist ε_K^1 such that $P_1 \leq \varepsilon_K^1$, where ε_K^1 is a sequence of numbers decreasing to zero as K goes to ∞. For the second term P_2, we have

$$P_2 \leq \frac{1}{n}\sum_{i=1}^{n} \sum_{1\leq k<l\leq d} \sup_{\boldsymbol{\theta}\in\Theta} \left\|\nabla \log\left(\tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta})\right)\right\| \|\alpha - \beta\|.$$

As

$$\begin{aligned}
\frac{\partial}{\partial a_{k,l}} \tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta}) &\leq \frac{N_k^i}{a_{k,l}(\boldsymbol{\theta})} \tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta}), \\
\frac{\partial}{\partial b_{k,l}} \tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta}) &\leq \frac{N_l^i}{b_{k,l}(\boldsymbol{\theta})} \tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta}), \\
\frac{\partial}{\partial c_{k,l}} \tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta}) &\leq \frac{N_k^i + N_l^i}{c_{k,l}(\boldsymbol{\theta})} \tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta}),
\end{aligned}$$

we have ($F_{k,l}$ is continuously differentiable over a compact set) for $u = 1, \ldots, p$,

$$\begin{aligned}\left|\frac{\partial}{\partial\theta_u}\log\left(\tilde{p}_{k,l}(N_k^i, N_l^i, \boldsymbol{\theta})\right)\right| &\leq \frac{\partial\, a_{k,l}}{\partial\theta_u}(\boldsymbol{\theta})\frac{N_k^i}{a_{k,l}(\boldsymbol{\theta})} + \frac{\partial\, b_{k,l}}{\partial\theta_u}(\boldsymbol{\theta})\frac{N_l^i}{b_{k,l}(\boldsymbol{\theta})} + \frac{\partial\, c_{k,l}}{\partial\theta_u}(\boldsymbol{\theta})\frac{N_k^i + N_l^i}{c_{k,l}(\boldsymbol{\theta})} \\ &\leq C_{k,l}(N_k^i + N_l^i),\end{aligned}$$

where $C_{k,l}$ is a positive constant. By denoting as C the maximum constant $C_{k,l}$, $1 \leq k < l \leq d$, the following result can be obtained:

$$\begin{aligned}P_2 &\leq \sqrt{p}C\|\alpha - \beta\|\frac{1}{n}\sum_{i=1}^{n}\sum_{1\leq k<l\leq d}(N_k^i + N_l^i) \\ &\leq \frac{\sqrt{p}}{K}\frac{C}{n}\sum_{i=1}^{n}\sum_{1\leq k<l\leq d}(N_k^i + N_l^i) \\ &= \frac{1}{K}W_n.\end{aligned}$$

By the weak law of large numbers, as n goes to ∞, W_n converges in $\mathbb{P}_{\boldsymbol{\theta_0}}$-probability to $l_W = \sqrt{p}C\sum_{1\leq k<l\leq d}\mathbb{E}(N_k + N_l) < \infty$. Let denote $\varepsilon_K = \varepsilon_K^1 + \frac{2}{K}l_W$, which goes to zero when K goes to ∞. Finally, since $w(n, 1/K) \leq P_1 + P_2 \leq \varepsilon_K^1 + \frac{1}{K}W_n$, we obtain

$$\mathbb{P}_{\boldsymbol{\theta_0}}\left(w\left(n, \frac{1}{K}\right) \geq \varepsilon_K\right) \leq \mathbb{P}_{\boldsymbol{\theta_0}}\left(W_n - l_W > l_W\right),$$

which converges to zero as n goes to ∞ since $l_W > 0$.

$\mathbf{A_3}$: Assumption 1 involves that the point $\boldsymbol{\theta_0}$ belongs to the interior of the space Θ. The function $U_n(\boldsymbol{\theta})$ is twice continuously differentiable on Θ as sum of twice continuously differentiable functions.

$\tilde{\mathbf{A}}_\mathbf{4}$: To prove that derivation and integration relating to μ can be permuted for $p_{k,l}$, we can use the differentiability properties of integrals defined by a parameter. Following the same way as for $\mathbf{A_2}$, we use an upper bound for the partial derivatives of $p_{k,l}$,

$$\begin{aligned}\left|\frac{\partial}{\partial\theta_u}p_{k,l}(n_k, n_l, \boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial\theta_u}g_{k,l}(\boldsymbol{\theta})\tilde{p}_{k,l}(n_k, n_l, \boldsymbol{\theta}) + g_{k,l}(\boldsymbol{\theta})\frac{\partial}{\partial\theta_u}\tilde{p}_{k,l}(n_k, n_l, \boldsymbol{\theta})\right|, \\ &\leq (C_1 + C_2(n_k + n_l))\tilde{p}_{k,l}(n_k, n_l, \boldsymbol{\theta}^\star),\end{aligned}$$

where $u = 1, \ldots, p$, C_1 and C_2 are positive constants and $\boldsymbol{\theta}^\star$ is the maximum argument of the continuous function $\tilde{p}_{k,l}$ over the compact set Θ. So the dominated function is μ-integrable. Since $p_{k,l}$ is differentiable, derivation and integration relating to μ can be permuted for $p_{k,l}$. In particular, that implies the random vector $\sum_{1\leq k<l\leq d}\nabla\log p_{k,l}(N_k, N_l, \boldsymbol{\theta_0})$ is centered. To prove the existence of its covariance matrix, we can show that for all $u, v = 1, \ldots, p$ and for all $k < l$:

$$\mathbb{E}_{\boldsymbol{\theta_0}}\left(\left|\frac{\partial}{\partial\theta_u}\log p_{k,l}(N_k, N_l, \boldsymbol{\theta_0})\frac{\partial}{\partial\theta_v}\log p_{k,l}(N_k, N_l, \boldsymbol{\theta_0})\right|\right) < \infty.$$

As above, there exist positive constants C_1, C_2 and C_3 such that,

$$\left|\frac{\partial}{\partial\theta_u}\log p_{k,l}(N_k, N_l, \boldsymbol{\theta_0})\frac{\partial}{\partial\theta_v}\log p_{k,l}(N_k, N_l, \boldsymbol{\theta_0})\right| \leq C_1 + C_2(N_k + N_l) + C_3(N_k + N_l)^2,$$

which is of finite expectation since all order moments of variables N_k, $k = 1, \ldots, d$, exist.

$\tilde{\mathbf{A}}_\mathbf{5}$: We have

$$\left|\frac{\partial^2}{\partial\theta_u\partial\theta_v}\log p_{k,l}(n_k, n_l, \boldsymbol{\theta})\right| \leq \left|\frac{\frac{\partial^2}{\partial\theta_u\partial\theta_v}p_{k,l}(n_k, n_l, \boldsymbol{\theta})}{p_{k,l}(n_k, n_l, \boldsymbol{\theta})}\right| + \left|\frac{\frac{\partial}{\partial\theta_u}p_{k,l}(n_k, n_l, \boldsymbol{\theta})\frac{\partial}{\partial\theta_v}p_{k,l}(n_k, n_l, \boldsymbol{\theta})}{p_{k,l}(n_k, n_l, \boldsymbol{\theta})^2}\right|.$$

As above ($F_{k,l}$ is twice continuously differentiable), straightforward computations leads to the following results:

$$\begin{aligned}\left|\frac{\partial^2}{\partial\theta_u\partial\theta_v}\log p_{k,l}(n_k,n_l,\boldsymbol{\theta})\right| &\leq C_1+C_2(n_k+n_l)+C_3(n_k+n_l)^2,\\ &= h_{k,l}(n_k,n_l),\end{aligned}$$

where C_1, C_2 and C_3 are positive constants. For the same raisons as previously, $h_{k,l}$ is $\mathbb{P}_{\boldsymbol{\theta_0}}$-integrable.

$\mathbf{A_6}$: From $\tilde{\mathbf{A}}_\mathbf{5}$, the random variables $\frac{\partial^2}{\partial\theta_u\partial\theta_v}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})$ are of finite expectation and by the weak law of large numbers, $\left(\frac{\partial^2}{\partial\theta_u\partial\theta_v}U_n(\boldsymbol{\theta_0})\right)_{u,v=1,\dots,p}$ converges in $\mathbb{P}_{\boldsymbol{\theta_0}}$-probability to

$$I_U(\boldsymbol{\theta_0})_{u,v=1,\dots,p}=\mathbb{E}_{\boldsymbol{\theta_0}}\left(\sum_{1\leq k<l\leq d}\frac{\partial^2}{\partial\theta_u\partial\theta_v}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\right).$$

Furthermore, from $\tilde{\mathbf{A}}_\mathbf{5}$ derivation and integration can be permuted twice and from $\tilde{\mathbf{A}}_\mathbf{4}$ the random vector $\sum_{1\leq k<l\leq d}\nabla\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})$ is centered; that leads to

$$\begin{aligned}I_U(\boldsymbol{\theta_0})_{u,v=1,\dots,p}&=-\sum_{1\leq k<l\leq d}\mathbb{E}_{\boldsymbol{\theta_0}}\left(\frac{\partial}{\partial\theta_u}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\frac{\partial}{\partial\theta_v}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\right),\\ &=-\sum_{1\leq k<l\leq d}J_{F_{k,l}}(\boldsymbol{\theta_0})^T\times\\ &\mathbb{E}_{\boldsymbol{\theta_0}}\left[\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})^T\right]J_{F_{k,l}}(\boldsymbol{\theta_0}),\end{aligned}\tag{27}$$

where $J_{F_{k,l}}(\boldsymbol{\theta_0})$ is the Jacobian matrix (of size $3\times p$) at the point $\boldsymbol{\theta_0}$. Note that the matrix $I_U(\boldsymbol{\theta_0})$ is the opposite of a sum of covariance matrices. Denoting

$$I_U^{k,l}(\boldsymbol{\theta_0})=\mathbb{E}_{\boldsymbol{\theta_0}}\left[\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})^T\right],$$

for all $1\leq k<l\leq d$, Eq. (27) leads to the following expression:

$$I_U(\boldsymbol{\theta_0})=-J_F(\boldsymbol{\theta_0})^T\begin{pmatrix}I_U^{1,2}(\boldsymbol{\theta_0})&&\\&\ddots&0\\0&&I_U^{d-1,d}(\boldsymbol{\theta_0})\end{pmatrix}J_F(\boldsymbol{\theta_0}),$$

where $J_F(\boldsymbol{\theta_0})=\left(J_{F_{1,2}}(\boldsymbol{\theta_0})^T\dots J_{F_{d-1,d}}(\boldsymbol{\theta_0})^T\right)^T$ is the $\frac{3}{2}d(d-1)\times p$ Jacobian matrix of $F(\boldsymbol{\theta_0})$. As F is an injective map on Θ (assumption 2), the matrix $J_F(\boldsymbol{\theta_0})$ has rank p. Therefore $I_U(\boldsymbol{\theta_0})$ is invertible if the diagonal matrix composed of the matrices $I_U^{k,l}(\boldsymbol{\theta_0})$ is invertible. Thus, we must only show that the matrices $I_U^{k,l}(\boldsymbol{\theta_0})$ are invertible for all $1\leq k<l\leq d$. However the property

$$\det\left(I_U^{k,l}(\boldsymbol{\theta_0})\right)=$$

$$\det\left(\mathbb{E}_{\boldsymbol{\theta_0}}\left(\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})^T\right)\right)=0\,,$$

is equivalent to the existence of some constants α_1, α_2 and α_3 (not all zero) such that

$$(\alpha_1,\alpha_2,\alpha_3)\nabla_{\{a_{k,l},b_{k,l},c_{k,l}\}}\log p_{k,l}(N_k,N_l,\boldsymbol{\theta_0})=0\tag{28}$$

almost surely. Eq. (28) involves in particular when $(N_k,N_l)=(0,0)$,

$$\alpha_1(1-b_{k,l}(\boldsymbol{\theta_0}))+\alpha_2(1-a_{k,l}(\boldsymbol{\theta_0}))+\alpha_3=0,$$

when $(N_k, N_l) = (1, 0)$,

$$-\alpha_1 \frac{(1 - a_{k,l}(\boldsymbol{\theta_0}))(1 - b_{k,l}(\boldsymbol{\theta_0})) - c_{k,l}(\boldsymbol{\theta_0})}{L a_{k,l}(\boldsymbol{\theta_0})} + \alpha_1(1 - b_{k,l}(\boldsymbol{\theta_0})) + \alpha_2(1 - a_{k,l}(\boldsymbol{\theta_0})) + \alpha_3 = 0,$$

and when $(N_k, N_l) = (0, 1)$,

$$-\alpha_2 \frac{(1 - a_{k,l}(\boldsymbol{\theta_0}))(1 - b_{k,l}(\boldsymbol{\theta_0})) - c_{k,l}(\boldsymbol{\theta_0})}{L b_{k,l}(\boldsymbol{\theta_0})} + \alpha_1(1 - b_{k,l}(\boldsymbol{\theta_0})) + \alpha_2(1 - a_{k,l}(\boldsymbol{\theta_0})) + \alpha_3 = 0.$$

Due to the constraints on $a_{k,l}(\boldsymbol{\theta_0})$, $b_{k,l}(\boldsymbol{\theta_0})$ and $c_{k,l}(\boldsymbol{\theta_0})$ (see Assumption 2), that leads to $\alpha_1 = \alpha_2 = \alpha_3 = 0$, so $I_U^{k,l}(\boldsymbol{\theta_0})$ is invertible for all $1 \leq k < l \leq d$. Consequently $I_U(\boldsymbol{\theta_0})$ is invertible.

List of Figures

List of Tables

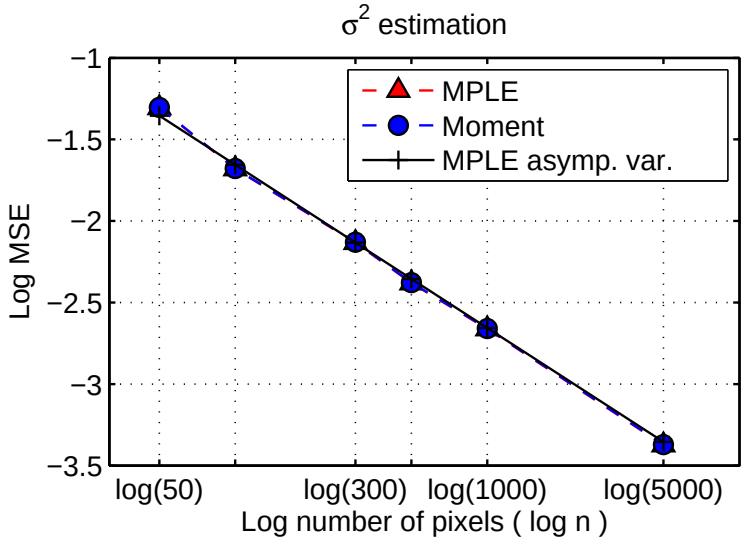


Fig. 1 log MSEs for parameter σ^2 ("MPLE": Maximum Pairwise likelihood estimator, "Moment": Moment estimator).

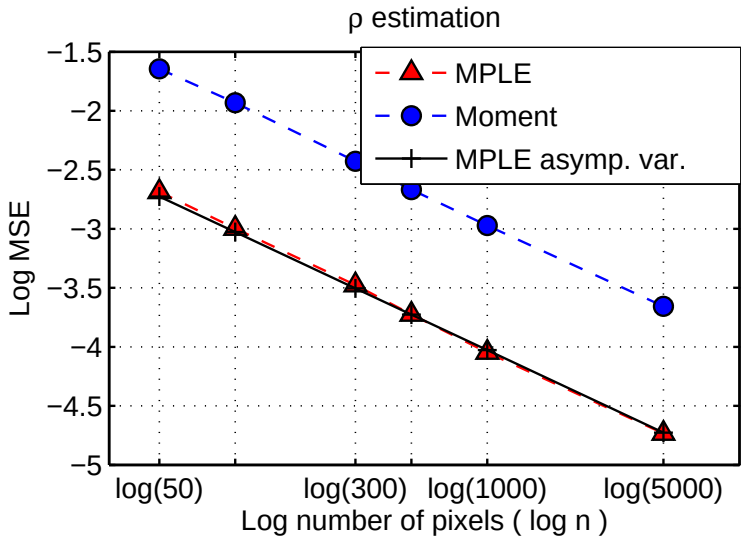


Fig. 2 log MSEs for parameter ρ ("MPLE": Maximum Pairwise likelihood estimator, "Moment": Moment estimator).

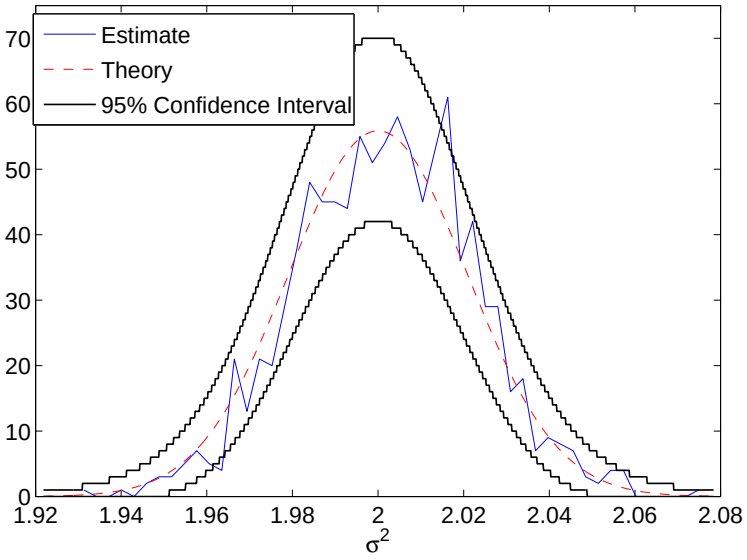


Fig. 3 Frequency polygon and theoretical asymptotic frequency distribution (denoted respectively as "Estimate" and "Theory") of $\hat{\sigma^2}$ with 95% confidence intervals.

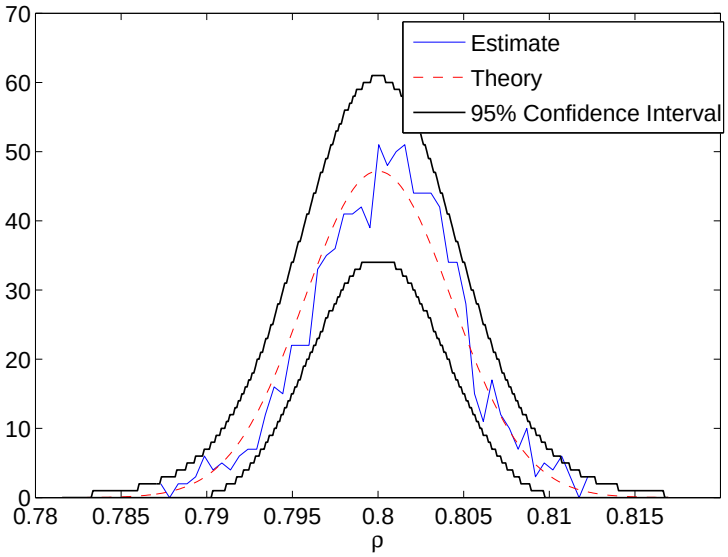


Fig. 4 Frequency polygon and theoretical asymptotic frequency distribution (denoted respectively as "Estimate" and "Theory") of $\hat{\rho}$ with 95% confidence intervals.

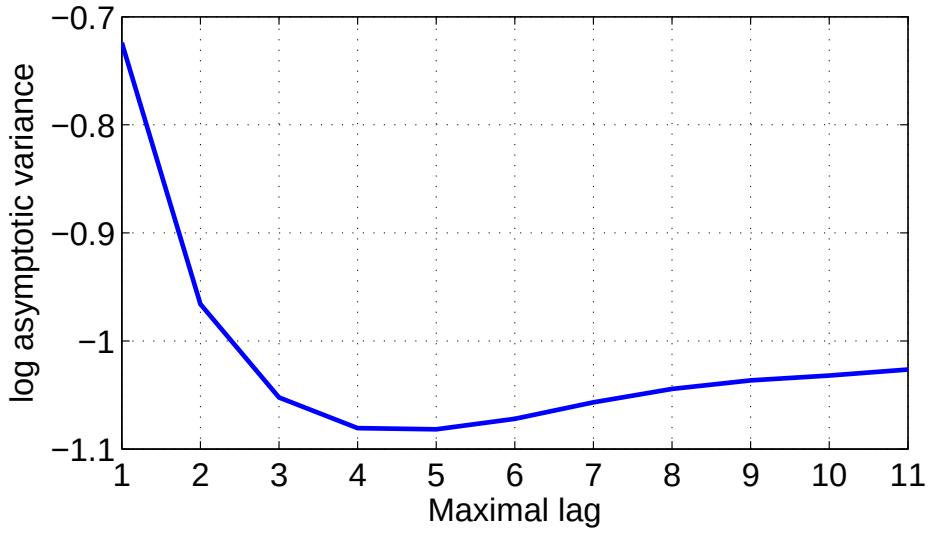


Fig. 5 Logarithm of the asymptotic variance for the weighted MPLE of ρ vs the maximal lag τ between the considered pairs

.

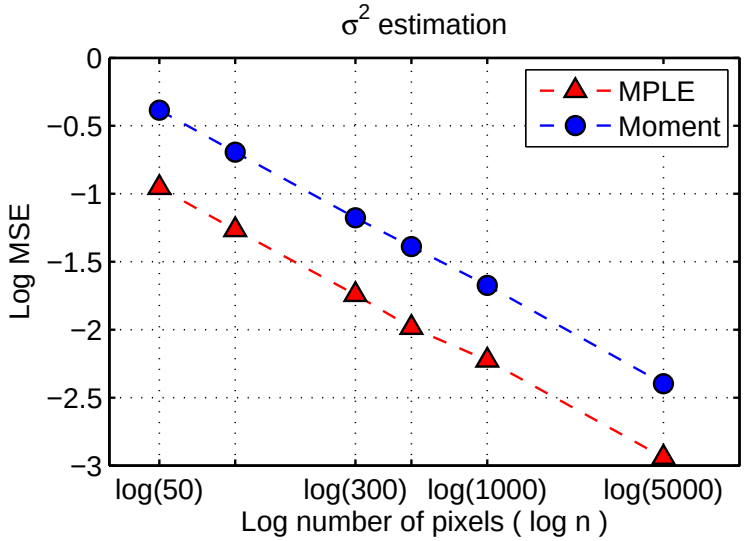


Fig. 6 log MSEs for parameter σ^2 ("MPLE": Maximum Pairwise likelihood estimator, "Moment": moment estimator).

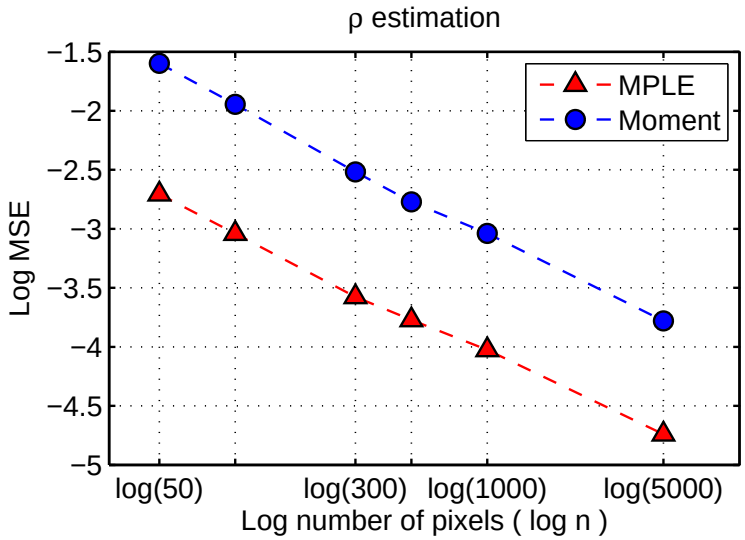


Fig. 7 log MSEs for parameter ρ ("MPLE": Maximum Pairwise likelihood estimator, "Moment": Moment estimator).

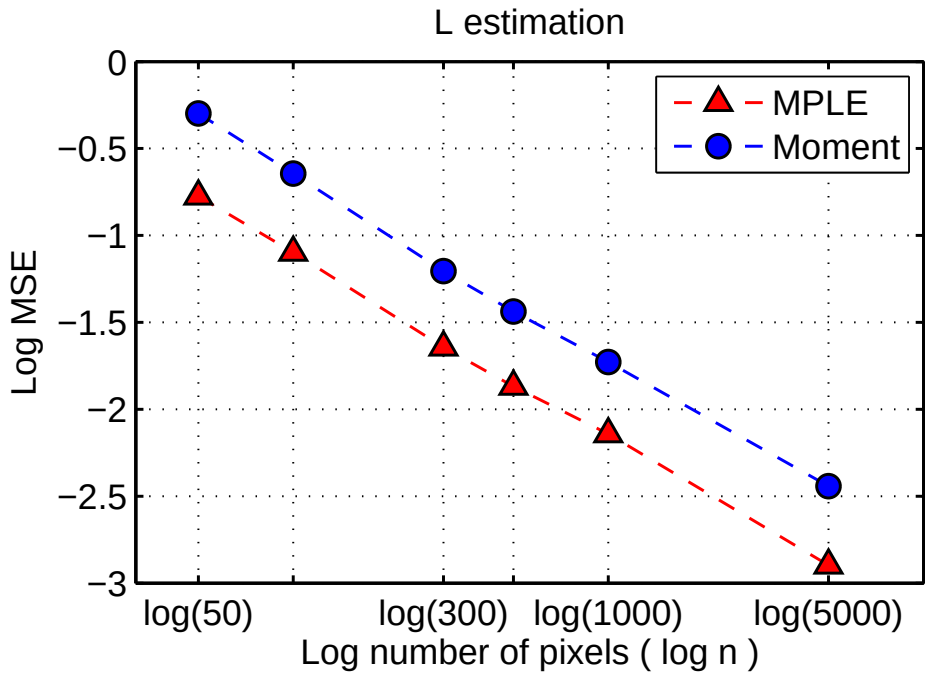
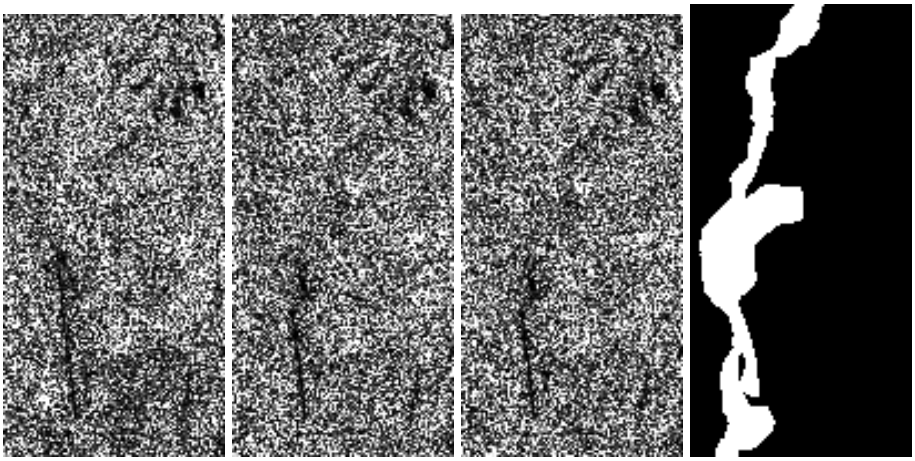
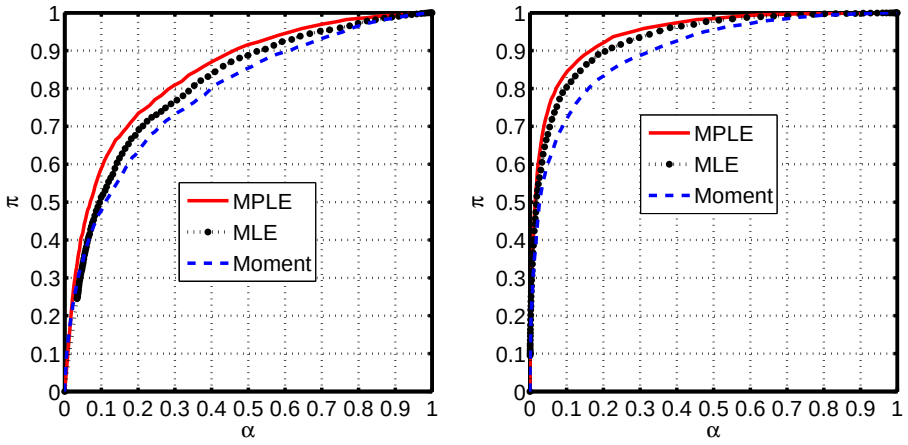


Fig. 8 log MSEs for parameter L ("MPLE": Maximum Pairwise likelihood estimator, "Moment": Moment estimator).

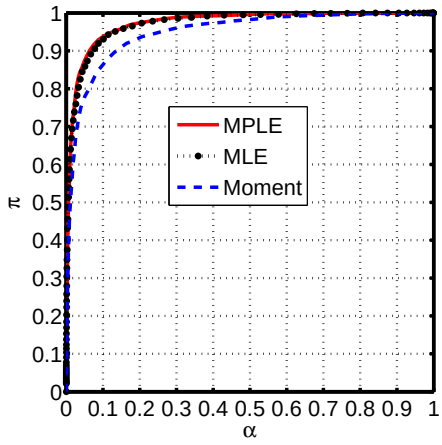


(a) I (before) (b) J (after) (c) K (after) (d) mask

Fig. 9 Low-flux 200×100 radarsat images of the Nyiragongo volcano before and after an eruption.



(a) $n = 3 \times 3$ (b) $n = 5 \times 5$



(c) $n = 7 \times 7$

Fig. 10 ROCs for Nyiragongo volcano images for different window sizes.

Table 1 Simulation results for the estimation of $\boldsymbol{\theta} = (\sigma^2, \rho)$ obtained from 1000 Monte-Carlo runs ($\sigma^2 = 2$, $\rho = 0.8$ and $L = 4$)

		σ^2			ρ		
n		bias	std	MSE	bias	std	MSE
50	MPLE	-1.80e-03	2.22e-01	4.93e-02	-7.72e-03	4.47e-02	2.06e-03
	Mom	-1.63e-03	2.23e-01	4.96e-02	-2.21e-02	1.49e-01	2.27e-02
100	MPLE	1.91e-03	1.44e-01	2.08e-02	-3.99e-03	3.15e-02	1.01e-03
	Mom	1.94e-03	1.45e-01	2.10e-02	-1.10e-02	1.08e-01	1.17e-02
300	MPLE	1.95e-03	8.59e-02	7.37e-03	-1.33e-03	1.83e-02	3.35e-04
	Mom	1.97e-03	8.61e-02	7.41e-03	-3.97e-03	6.10e-02	3.73e-03
500	MPLE	1.98e-03	6.45e-02	4.17e-03	-1.24e-04	1.37e-02	1.88e-04
	Mom	1.95e-03	6.47e-02	4.19e-03	2.87e-04	4.64e-02	2.15e-03
1000	MPLE	3.00e-03	4.66e-02	2.18e-03	-4.71e-04	9.47e-03	8.98e-05
	Mom	3.07e-03	4.67e-02	2.19e-03	3.60e-04	3.27e-02	1.07e-03
5000	MPLE	1.09e-03	2.06e-02	4.23e-04	-1.08e-04	4.30e-03	1.85e-05
	Mom	1.09e-03	2.06e-02	4.26e-04	-1.28e-05	1.49e-02	2.21e-04

Table 2 Simulation results for the estimation of $\boldsymbol{\theta} = (\sigma^2, \rho, L)$ obtained from 1000 Monte-Carlo runs ($\sigma^2 = 2$, $\rho = 0.8$ and $L = 4$)

			σ^2	
n		bias	std	MSE
50	MPLE	-1.35e-02	3.33e-01	1.11e-01
	Mom	-8.55e-02	6.37e-01	4.12e-01
100	MPLE	-3.60e-03	2.33e-01	5.43e-02
	Mom	-4.98e-02	4.48e-01	2.03e-01
300	MPLE	6.20e-03	1.34e-01	1.81e-02
	Mom	-1.60e-02	2.58e-01	6.66e-02
500	MPLE	9.81e-03	1.01e-01	1.04e-02
	Mom	-4.87e-03	2.02e-01	4.09e-02
1000	MPLE	9.53e-03	7.66e-02	5.95e-03
	Mom	2.03e-03	1.46e-01	2.12e-02
5000	MPLE	3.34e-03	3.37e-02	1.15e-03
	Mom	1.43e-03	6.33e-02	4.01e-03
			ρ	
n		bias	std	MSE
50	MPLE	-4.28e-03	4.41e-02	1.96e-03
	Mom	5.61e-02	1.49e-01	2.53e-02
100	MPLE	-2.58e-03	3.01e-02	9.10e-04
	Mom	3.30e-02	1.01e-01	1.14e-02
300	MPLE	-3.03e-04	1.63e-02	2.65e-04
	Mom	9.27e-03	5.44e-02	3.04e-03
500	MPLE	5.50e-04	1.30e-02	1.69e-04
	Mom	5.15e-03	4.08e-02	1.69e-03
1000	MPLE	1.02e-03	9.66e-03	9.43e-05
	Mom	1.98e-03	3.02e-02	9.16e-04
5000	MPLE	4.67e-04	4.24e-03	1.82e-05
	Mom	5.98e-04	1.29e-02	1.66e-04
			L	
n		bias	std	MSE
50	MPLE	5.59e-02	4.06e-01	1.68e-01
	Mom	2.44e-01	6.66e-01	5.02e-01
100	MPLE	3.25e-02	2.81e-01	7.98e-02
	Mom	1.34e-01	4.57e-01	2.27e-01
300	MPLE	4.12e-03	1.51e-01	2.27e-02
	Mom	4.19e-02	2.46e-01	6.23e-02
500	MPLE	-3.05e-03	1.16e-01	1.35e-02
	Mom	2.04e-02	1.90e-01	3.64e-02
1000	MPLE	-5.56e-03	8.48e-02	7.22e-03
	Mom	5.85e-03	1.37e-01	1.87e-02
5000	MPLE	-8.95e-04	3.56e-02	1.27e-03
	Mom	1.60e-03	6.01e-02	3.61e-03

Table 3 Means and standard deviations of the estimated correlation coefficients for the two classes "Pixels affected by a change" (white pixels in the mask shown in fig. 9(d)) and "Pixels not affected by a change" (black pixels in the mask)

		Pixels affected by a change		Pixels not affected by a change	
n		mean	std	mean	std
3×3	MPLE	0.355	0.260	0.669	0.181
	ML	0.355	0.276	0.661	0.212
	Mom	0.286	0.337	0.621	0.242
5×5	MPLE	0.332	0.184	0.663	0.108
	ML	0.324	0.191	0.658	0.123
	Mom	0.314	0.215	0.647	0.147
7×7	MPLE	0.338	0.146	0.658	0.084
	ML	0.332	0.144	0.654	0.091
	Mom	0.327	0.164	0.653	0.113