



**HAL**  
open science

# A randomized pairwise likelihood method for complex statistical inferences

Gildas Mazo, Dimitris Karlis, Andrea Rau

► **To cite this version:**

Gildas Mazo, Dimitris Karlis, Andrea Rau. A randomized pairwise likelihood method for complex statistical inferences. 2022. hal-03126620v2

**HAL Id: hal-03126620**

**<https://hal.inrae.fr/hal-03126620v2>**

Preprint submitted on 24 Feb 2022 (v2), last revised 13 Sep 2023 (v5)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A randomized pairwise likelihood method for complex statistical inferences

Gildas Mazo<sup>1</sup>, Dimitris Karlis<sup>2</sup>, and Andrea Rau<sup>3,4</sup>

<sup>1</sup> MaIAGE, INRAE, Université Paris Saclay, 78350 Jouy-en-Josas, France

<sup>2</sup> Athens University of Economics and Business

<sup>3</sup> Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France

<sup>4</sup> BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille,

Université de Picardie Jules Verne, 80200 Estrées-Mons, France

## Abstract

Pairwise likelihood methods are commonly used for inference in parametric statistical models in cases where the full likelihood is too complex to be used, such as multivariate count data. Although pairwise likelihood methods represent a useful solution to perform inference for intractable likelihoods, several computational challenges remain. The pairwise likelihood function still requires the computation of a sum over all pairs of variables and all observations, which may be prohibitive in high dimensions. Moreover, it may be difficult to calculate confidence intervals of the resulting estimators, as they involve summing all pairs of pairs and all of the four-dimensional marginals. To alleviate these issues, we consider a randomized pairwise likelihood approach, where only summands randomly sampled across observations and pairs are used for the estimation. In addition to the usual tradeoff between statistical and computational efficiency, it is shown that, under a condition on the sampling parameter, this two-way random sampling mechanism makes the individual bivariate likelihood scores become asymptotically independent, allowing more computationally efficient confidence intervals to be constructed. The proposed approach is illustrated in tandem with copula-based models for multivariate count data in simulations, and in real data from a transcriptome study.

*Keywords:* composite likelihood; randomization; confidence intervals; multivariate count data; computational challenges

# 1 Introduction

Multivariate models represent a valuable framework to explore and estimate interrelationships among variables in large and complex datasets, such as high-throughput count data collected in molecular biology. However, regardless of the considered multivariate model, the corresponding likelihood is often complex, costly to evaluate, or even intractable. To overcome this issue, one can maximize a sum of lower-dimensional likelihoods, called a composite likelihood, instead of the full likelihood (Lindsay, 1988). Often, bivariate marginals are used and the composite likelihood is called the pairwise likelihood. The advantage is computational, since it obviates the need to compute the full likelihood. In many models, the information retained is sufficient to estimate the parameters of interest. The corresponding price to pay is a loss of efficiency of the resulting estimator, which is nonetheless guaranteed to be asymptotically normal under mild conditions (Varin et al., 2011); we note that variational methods do not have this guarantee in general (Blei et al., 2017). In addition, we remark that composite likelihood methods are agnostic to data type and not limited to multivariate count data, although such models may particularly benefit from their use (Zhao and Joe, 2005).

Pairwise likelihood methods have been successfully used in many applications (Varin et al., 2011). Many variants have been derived to accommodate specific models, data or tasks; for instance, spatial data are addressed in Bai et al. (2014). Also, several authors have proposed ways to improve the efficiency of composite likelihood methods (Ferrari et al., 2016), primarily by adding weights to the component likelihoods (see, e.g. Joe and Lee, 2009). It appears, however, that finding and estimating the optimal weights in general is a very difficult problem which may not have a solution (Lindsay et al., 2011). In the following, we shall focus on the pairwise likelihood, the most popular version of composite

likelihoods.

In high dimensions, applying the pairwise likelihood method may be cumbersome. With  $d$  variables, the number of pairs is of order  $d^2$ . To get confidence intervals, one needs to compute a double sum over pairs of pairs of order  $d^4$  and all of the four-dimensional marginals. All this can be time-consuming and burdensome. Although there is little literature on how to address these computational issues, several research directions have been proposed. For instance, instead of taking all of the pairs, one can consider a small subset (Huang and Ferrari, 2021; Papageorgiou and Moustaki, 2019), although selecting a good subset is a difficult problem. Some heuristics were proposed in Papageorgiou and Moustaki (2019), but no theoretical justification was provided and the asymptotic properties of the estimators are unknown. In Huang and Ferrari (2021), pair selection was performed by regularization, but as this approach depends on the existence of a consistent estimator with rate  $\sqrt{n}$ , the computational issue is unresolved. In the context of conditional random fields, a stochastic combination of low-dimensional conditional likelihoods was proposed in Dillon and Lebanon (2010). This allows a reduction in the number of times the conditional log-densities of the model are evaluated, but it does not solve the problem for the construction of confidence intervals.

To alleviate the computational issues of the pairwise likelihood method, we consider a randomized pairwise likelihood approach. Only summands randomly sampled across observations and pairs are used for the estimation of the parameters. To implement this strategy, one draws, for each sample size  $n$ , i.i.d. Bernoulli weights  $W_{ni}^{(a)}$ ,  $i = 1, \dots, n$ ,  $a \in \{\{1, 2\}, \dots, \{d - 1, d\}\}$ , with parameter  $\pi_n$ ; all summands for which  $W_{ni}^{(a)} = 0$  are discarded. A fundamental point here is that we allow the sampling parameter  $\pi_n$  to decrease with  $n$ —we shall come back to this later. The sampling parameter controls the tradeoff

between the computational complexity and the statistical efficiency. An intuitive way to see this is to notice that the average number of summands needed to compute the randomized pairwise likelihood is equal to  $n\pi_n d(d-1)/2$ . However, there is an additional reason why  $\pi_n$  permits a reduction of the computational cost. By letting  $\pi_n \rightarrow 0$ , the bivariate log density gradients become asymptotically independent, leading to the disappearance of the term of computational complexity  $d^4$  in the estimator's asymptotic variance. In practice, this suggests that one may be able to approximate confidence intervals at a much lower cost than in the standard pairwise likelihood method.

The remainder of the paper proceeds as follows: Section 2 reviews the pairwise likelihood method. The theory is presented in a rigorous way not yet achieved in the literature. In particular, the conditions for consistency, that is, the ability to estimate the full distribution from its bivariate marginals alone, are made explicit. Computational problems are discussed in more detail. Then, Section 3 introduces the randomized pairwise likelihood method. A first asymptotic result is given when  $\pi_n = \pi$  is fixed. The impact of  $\pi$  on the estimator's asymptotic variance is discussed in detail. A second asymptotic result is given when  $\pi_n \rightarrow 0$ . It is explained why, in this setting, inexpensive approximation of confidence intervals may be possible. Section 4 focuses on the exchangeable Gaussian model, for which explicit calculations are possible. Section 5 reviews copula models and explains why the randomized pairwise likelihood may benefit them. Section 6 reports the results of simulation experiments carried out to assess performance of the randomized pairwise likelihood method on multivariate count data. Section 7 illustrates how the approach can be applied to a set of transcriptome data with multivariate count data models based on Poisson marginals and Gaussian copulas. Concluding remarks may be found in Section 8. The proofs and additional simulations can be found in the Supplementary Material.

## 2 Maximum pairwise likelihood inference

Pairwise likelihood methods permit the estimation of unknown parameters of a statistical model without the need to specify the complete joint density (or probability mass) function of the model. The idea is to replace the full likelihood by a sum of marginal likelihoods, which is useful when the full likelihood is complex, such as the case of discrete data. Pairwise likelihood is a particular case of the so-called composite likelihood, which is based on likelihoods conditioned on certain events (Lindsay, 1988; Varin et al., 2011; Varin and Vidoni, 2005). For simplicity and because it is most widely used, we shall focus on the pairwise likelihood, but the theoretical results extend straightforwardly to composite likelihoods.

### 2.1 Definition, assumptions and asymptotic properties

Let  $X_i := (X_{i1}, \dots, X_{id})$ ,  $i = 1, \dots, n$ , be independent random vectors with a common density  $f_0$  with respect to some “base measure”—typically the Lebesgue measure or the counting measure—on the Euclidean space  $\mathbf{R}^d$ . The density  $f_0$  is assumed to be square integrable and lie in an identifiable parametric family  $\{f(\bullet; \theta), \theta = (\theta_1, \dots, \theta_q) \in \Theta\}$  for some open subset  $\Theta$  of  $\mathbf{R}^q$ . Let  $\theta_0$  denote the element of  $\Theta$  such that  $f_0(\bullet) = f(\bullet; \theta_0)$ . Let  $\mathcal{A}$  be the set of all pairs of variables. Its cardinal is  $d(d-1)/2$ . The pairs in  $\mathcal{A}$  are ordered in the lexicographical order. Denote by  $f_a(\cdot, \cdot; \theta)$  the marginal density corresponding to the pair  $a$  and write  $\ell_a(\cdot, \cdot; \theta)$  for  $\log f_a(\cdot, \cdot; \theta)$ . Whenever it exists, denote by  $\dot{\ell}_a(\cdot, \cdot; \theta)$  the gradient of  $\ell_a(\cdot, \cdot; \theta)$  with respect to  $\theta$ . Whenever a function is encountered with a bullet symbol, it means that the argument it replaces is a vector with three components or more. Otherwise, there are as many dot symbols as there are components. If  $a = \{j, j'\}$  is a pair

then  $(X_{ij}, X_{i'j'})$  is also denoted by  $X_i^{(a)}$ .

The pairwise log-likelihood function is given by

$$L_n^{\text{PL}}(\theta) = \frac{1}{n} \sum_{a \in \mathcal{A}} \sum_{i=1}^n \ell_a(X_i^{(a)}; \theta), \quad \theta \in \Theta. \quad (1)$$

The population version of the pairwise log-likelihood function is  $\sum_a L_a(\theta)$ , where  $L_a(\theta)$  stands for  $E \ell_a(X_1^{(a)}; \theta)$ . As usual, the goal is to estimate the maximizer of the population pairwise log-likelihood by maximizing the pairwise log-likelihood function. From the viewpoint of M-estimation theory, the population pairwise likelihood is the objective criterion function, the maximizer of which is the parameter of interest. In this case the objective criterion is the sum of “bivariate” Kullback-Leibler information criteria. This is the viewpoint we shall adopt throughout the paper. Wang and Wu (2014) provide a different view. According to them, maximizing the pairwise likelihood function can also be seen as maximizing the full Kullback-Leibler information under some information constraints.

We call the maximum pairwise likelihood estimator (MPLE) every element  $\hat{\theta}_n^{\text{MPL}}$  of  $\Theta$  that satisfies  $L_n^{\text{PL}}(\hat{\theta}_n^{\text{MPL}}) \geq L_n^{\text{PL}}(\theta)$  for all  $\theta$  in some compact subset of  $\Theta$ . Maximization over compact subsets ensures the existence of MPLEs under minimal smoothness assumptions. Whenever we refer to MPLEs, it is implicitly understood that the compact subset over which  $\theta$  is estimated contains  $\theta_0$ .

**Assumption 1.** *The first, second and third derivatives of  $\ell_a(X_1^{(a)}; \theta)$  with respect to the components of  $\theta$  exist and are square integrable. Moreover, there exist square integrable functions  $\Psi_a$ ,  $a \in \mathcal{A}$ , such that  $\sup_{\theta \in \Theta} |\partial^3 \ell_a(X_1^{(a)}; \theta) / (\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3})| \leq \Psi_a(X_1^{(a)})$ , for all  $1 \leq i_1 \leq i_2 \leq i_3 \leq q$ . Finally, if  $\mathbf{m}_a$  stands for the base measure of which  $f_a(\cdot, \cdot; \theta)$  is the density then  $\int f_a(\cdot, \cdot; \theta) d\mathbf{m}_a$  and  $\int (\partial / \partial \theta_{i_1}) f_a(\cdot, \cdot; \theta) d\mathbf{m}_a$  can be differentiated under the integral sign.*

Assumption 1 is standard. It is mild enough to encompass many models and yet enable simple proofs. Under Assumption 1, the pairwise log-likelihood function is differentiable and hence MPLEs always exist. Assumption 1 could be weakened but at the expense of much more complicated proofs, and thus we keep this assumption.

When  $d = 2$ , MPLEs and maximum likelihood estimators coincide. In this case, Assumption 1 suffices to get the consistency and the asymptotic normality of these estimators. In general, however, we cannot expect MPLEs to be consistent without further assumptions, because a family of multivariate distributions cannot always be described by its pairs. There is, therefore, no reason for the map  $\theta \mapsto \sum_a L_a(\theta)$  to admit a unique maximizer, and we need to impose this as a condition.

**Assumption 2.** *The maximizer of  $\theta \mapsto \sum_a L_a(\theta)$  is unique.*

It is easy to see that each  $L_a$  is maximized at  $\theta_0$  and hence so is the mapping  $\sum_a L_a(\theta)$ . Thus, we deduce from Assumption 2 that  $\theta_0$  is the only maximizer of  $\sum_a L_a(\theta)$ .

**Remark 1.** *Even if  $\theta_0$  is the only maximizer of  $\sum_a L_a(\theta)$ , it does not mean that  $\theta_0$  is the only maximizer of  $L_a$ . Let  $d = 3$  and let  $(X_{11}, X_{12}, X_{13})$  be a Gaussian random vector with mean  $\mu_{01}, \mu_{02}, \mu_{03}$ , variances equal to one and correlation parameter  $\rho_0$ , so that  $\theta_0 = (\mu_{01}, \mu_{02}, \mu_{03}, \rho_0)$ . Then not only is  $L_{12}$  maximized at  $\theta_0$ , but also at  $(\mu_{01}, \mu_{02}, \mu, \rho_0)$  for any  $\mu$ .*

Assumption 2 is critical to ensure the consistency of pairwise likelihood methods. Sufficient conditions can be found in Proposition 1 below.

**Proposition 1.** *If, for every  $a \in \mathcal{A}$ , there is a function  $v_a$  on  $\Theta$  into a Euclidean space and a family of bivariate densities  $\{\tilde{f}_a(\cdot, \cdot; \vartheta_a), \vartheta_a \in \text{range } v_a\}$  such that*



- (i) the family  $\{\tilde{f}_a(\cdot, \cdot; \vartheta_a), \vartheta_a \in \text{range } v_a\}$  is identifiable
- (ii) the distributions  $\tilde{f}_a(\cdot, \cdot; v_a(\theta)) = f_a(\cdot, \cdot; \theta)$  coincide for all  $\theta$
- (iii) the mapping  $V(\theta) := (v_a(\theta))_{a \in \mathcal{A}}$  is one-to-one

then Assumption 2 holds.

These conditions will be useful to check Assumption 2 for the copula models of Section 5.

Assumptions 1 and 2 together imply that the MPLE is asymptotically normal, that is, we have that  $\sqrt{n}(\hat{\theta}_n^{\text{MPL}} - \theta_0)$  converges in distribution to a centered Gaussian random vector with some variance-covariance matrix, called the *asymptotic variance-covariance matrix*—or simply the *asymptotic variance*—of the estimator, given by  $S^{-1}CS^{-1} + S^{-1}$ , where  $S = \sum_{a \in \mathcal{A}} \text{E} \dot{\ell}_a \dot{\ell}_a^\top$ , and  $C = \sum_{a \neq b \in \mathcal{A}} \text{E} \dot{\ell}_a \dot{\ell}_b^\top$  is the between-scores correlation matrix. Here  $\text{E} \dot{\ell}_a \dot{\ell}_b^\top$  is a shorthand for  $\text{E} \dot{\ell}_a(X_1^{(a)}; \theta_0) \dot{\ell}_b(X_1^{(b)}; \theta_0)^\top$ . This result is standard and known since at least Lindsay (1988) but, as it turns out, it is difficult to find in the literature precise conditions under which this result is true.

To improve efficiency, weights could be added to the pairwise log-likelihood (Lindsay, 1988; Joe and Lee, 2009; Lindsay et al., 2011), leading to the maximization of

$$L_n^{\text{WPL}}(\theta) = \frac{1}{n} \sum_{a \in \mathcal{A}} w_a \sum_{i=1}^n \ell_a(X_i^{(a)}; \theta), \quad (2)$$

for some weights  $w_a \geq 0$ . In this case, Assumption 2 must be changed to “The maximizer of  $\theta \mapsto \sum_a w_a L_a(\theta)$  is unique” and Proposition 1 still holds.

The problem of choosing the optimal weights is difficult. In the one-dimensional case, that is, when the parameter is a scalar, a formula for the optimal weights exists but it

requires the computation of the between-scores correlation matrix  $C$ . This can be computationally challenging, as we shall see next. In the more realistic multivariate case, according to Lindsay (1988), a solution may not exist, and if it existed it would be difficult to compute.

## 2.2 Computational issues in higher dimensions

When the number of variables is large, the pairwise likelihood method may be burdensome to apply. Indeed, the computation of the pairwise log-likelihood requires up to  $O(nd^2)$  evaluations of a potentially complex function. Perhaps less apparent but not less important in applications is the computation of confidence intervals for the parameters. These are also difficult to get because the between-scores correlation matrix  $C$  is a double sum over pairs of order up to  $O(d^4)$ . Moreover, computing confidence intervals requires dealing with distributions in four dimensions, which were assumed to be quite complex in the first place.

To reduce the computational burden, a natural approach consists of choosing a small subset of pairs and computing the pairwise log-likelihood based on that subset alone. This method can be seen as a particular case of the weighted pairwise likelihood method, in which some weights are set to zero and the others equal to one. The performance of the estimator depends on the chosen subset. Choosing a good subset is a difficult problem. To the best of our knowledge, it appears that little work on this area exists in the literature. Some algorithms are given in Papageorgiou and Moustaki (2019) but no theory is provided. In Huang and Ferrari (2021), the mean squared error between the maximum log-likelihood score and the weighted pairwise log-likelihood score is minimized, and a penalty term is added to shrink some weights to zero. However, for this method to work, an initial consistent estimator is needed, and we are back to our initial problem.

Finally, it should be noted that subset selection methods are not always applicable. Removing a pair can invalidate the method, as the conditions for consistency are no longer met. As an example, consider a trivariate Gaussian distribution with three free correlation parameters. Removing any pair leads to the impossibility of estimating the corresponding correlation parameter.

### 3 The randomized pairwise likelihood method

We introduce a new estimator of  $\theta_0$  based on a randomized version of the pairwise log-likelihood function and thus cheaper to compute. Interestingly, confidence intervals can be computed with no more than  $O(d^2)$  computations.

#### 3.1 Definition and first results

The randomized pairwise likelihood method consists of taking at random only some of the pairs  $a$  and observations  $i$  in (1) to carry out the summation. Formally, the randomized pairwise log-likelihood function is defined as

$$L_n^{\text{RPL}}(\theta) = \frac{1}{n\pi_n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} W_{ni}^{(a)} \ell_a(X_i^{(a)}; \theta), \quad (3)$$

where, for each  $n$ ,  $W_{ni}^{(a)}$ ,  $i = 1, \dots, n$ ,  $a \in \mathcal{A}$ , are independent Bernoulli random variables with parameter  $0 < \pi_n \leq 1$ . They are assumed to be independent of  $X_1, \dots, X_n$ . The unknown parameter  $\theta_0 = (\theta_{01}, \dots, \theta_{0q})$  is estimated by maximizing the function in (3). In practice, one first draws the Bernoulli weights, which allows certain terms to be excluded from the pairwise log-likelihood function, and then maximizes the sum of the remaining

terms. If  $\pi_n = 1$  then  $\Pr(W_{ni}^{(a)} = 1) = 1$  and hence the functions (3) and (1) coincide.

**Definition 1.** *Every element  $\hat{\theta}_n^{MRPL}$  of  $\Theta$  that satisfies  $L_n^{RPL}(\hat{\theta}_n^{MRPL}) \geq L_n^{RPL}(\theta)$  for all  $\theta$  in some compact subset of  $\Theta$  is called a maximum randomized pairwise likelihood estimator (MRPLE).*

As before, it is implicitly understood that the compact subset has been taken large enough to contain  $\theta_0$ . The parameter  $\pi_n$  controls the computational cost. For clarity, suppose that  $\mathcal{A}$  is the set of all pairs. Since there are  $n$  observations and  $d(d-1)/2$  pairs, the expected number of terms in the randomized pairwise log-likelihood function is  $nd(d-1)\pi_n/2$ . For instance, if  $\pi_n = 1/6$ ,  $d = 3$  and  $n = 10000$  then one needs to sum 5000 terms on average to compute the randomized pairwise likelihood, and 30000 to compute the standard pairwise likelihood method.

The difference between the criterion functions (2) and (3) is that in the former, the weights do not depend on  $i$  and, hence, when a pair is dropped out, one removes all of the observations corresponding to it. With the randomized pairwise log-likelihood function, at least some partial observations will be included for any given pair and hence all parameters can be estimated, even in unstructured models. The probability that all pairs pick out at least one observation is  $[1 - (1 - \pi_n)^n]^{d(d-1)/2}$ . For instance, with  $\pi_n = 9/10$ ,  $n = 50$  and  $d = 10$ , this probability is about 0.793; with  $n = 100$  it is already 0.999.

We now turn to asymptotic properties. In general we let the parameter  $\pi_n$  vary with  $n$ . (The reason will be explained later.) For the time being, however, suppose that  $\pi_n$  is equal to some  $\pi \in (0, 1]$  for all  $n$ .

**Theorem 1.** *Suppose that Assumptions 1 and 2 hold. Assume that  $\pi_n$  is a constant sequence, that is,  $\pi_n = \pi \in (0, 1]$  for all  $n$ . If  $\hat{\theta}_n^{MRPL}$  is a MRPLE such that  $L_n^{RPL}(\hat{\theta}_n^{MRPL}) \geq$*

$L_n^{RPL}(\theta)$  for all  $\theta \in \Lambda$ , where  $\Lambda$  is a compact subset of  $\Theta$  and  $\theta_0$  is an interior point of  $\Lambda$ , then  $\sqrt{n} \left( \hat{\theta}_n^{MRPL} - \theta_0 \right)$  converges in distribution to a Gaussian random vector with mean zero and variance-covariance matrix  $S^{-1}CS^{-1} + \pi^{-1}S^{-1}$ .

**Remark 2.** Without the last sentence of Assumption 1, asymptotic normality still holds but with a different variance-covariance matrix.

Theorem 1 implies  $\hat{\theta}_n^{MRPL} \rightarrow \theta_0$  in probability. Choosing  $\pi = 1$  allows us to recover the results of Section 2.

### 3.2 Statistical versus computational efficiency

The randomized pairwise likelihood method sacrifices statistical efficiency (measured by asymptotic variance) for computational efficiency (measured by the expected number of times the function  $\ell_a(X_i^{(a)}; \theta)$  needs to be evaluated to compute the randomized pairwise log-likelihood). If one chooses, say,  $\pi = 1/k$ ,  $k \geq 1$ , then the expected number of needed evaluations will be divided by  $k$ , and hence the maximization of the randomized pairwise log-likelihood, and thus the computation of the estimate  $\hat{\theta}_n^{MRPL}$  will be greatly facilitated.

The price to pay, however, is that the asymptotic variance-covariance matrix of the estimator will be multiplied by some inflation factor. To emphasize the dependence on  $\pi$ , denote temporarily by  $\hat{\theta}_n^{MRPL}(\pi)$  the MRPLE based on  $\pi$ . For simplicity, assume that  $\hat{\theta}_n^{MRPL}(\pi)$  is a scalar and denote by  $V(\pi)$  its asymptotic variance. The factor by which the MRPLE's asymptotic variance will be multiplied, should one consider  $\hat{\theta}_n^{MRPL}(\pi')$  instead of  $\hat{\theta}_n^{MRPL}(\pi)$ , is referred to as *the inflation factor from  $\pi$  to  $\pi'$* . By definition, the inflation

factor is given by

$$\text{IF}(\pi'|\pi) := \frac{V(\pi')}{V(\pi)} = \frac{\pi S^{-1} C S^{-1}}{\pi S^{-1} C S^{-1} + S^{-1}} + \frac{S^{-1}}{\pi S^{-1} C S^{-1} + S^{-1}} \frac{\pi}{\pi'}.$$

For instance, if one sets  $\pi' = \pi/k$ ,  $k > 1$ , thus dividing the number of evaluations by  $k$ , then the asymptotic variance of the estimator will be multiplied by  $\text{IF}(k^{-1}\pi|\pi)$ .

We say that the inflation factor is subhomogeneous of order -1, or simply subhomogeneous, if  $\text{IF}(k^{-1}\pi|\pi) \leq k \text{IF}(\pi|\pi) = k$  for every  $\pi$ . If the inequality is replaced by an equality, we say that the inflation factor is homogeneous of order -1, or simply homogeneous. Arguably, the compromise between statistical and computational efficiency is acceptable when the inflation factor is subhomogeneous. In this case, dividing the number of evaluations by  $k$  yields an inflation of the variance by a factor less than  $k$ .

**Proposition 2.** *The inflation factor is subhomogeneous if and only if the matrix  $S^{-1} C S^{-1}$  is nonnegative definite.*

From Proposition 2, a satisfactory compromise occurs when  $S^{-1} C S^{-1}$  is nonnegative definite, that is, when the scores  $\dot{\ell}_a, \dot{\ell}_b$ ,  $a \neq b$ , tend to be positively correlated. In the real world, are the scores positively correlated? Intuitively, it can be argued that this is to be expected if the variables tend to be positively correlated. More often than not, this should be the case. To see this, note that in the multivariate Gaussian model of dimension  $d$  with a common correlation parameter, the common correlation cannot be less than  $-1/(d-1)$ , which is essentially zero as soon as the number of variables is more than a few.

### 3.3 Consequences of letting the sampling parameter vanish

The MRPLE depends on the sampling parameter  $\pi$ . If  $\pi$  is too small, there would be too little of the data and we would expect poor performance. Thus it is of interest to understand how small  $\pi$  can be. Also, intriguingly, multiplying  $\sqrt{n}(\hat{\theta}_n^{\text{MRPL}} - \theta_0)$  by  $\sqrt{\pi}$  in Theorem 1 yields the asymptotic variance  $\pi S^{-1}CS^{-1} + S^{-1}$ , suggesting that, by letting  $\pi = \pi_n \rightarrow 0$  as  $n \rightarrow \infty$ , we may simply get  $S^{-1}$ : this would allow one to get rid of the costly matrix  $C$ .

**Theorem 2.** *Suppose that Assumptions 1 and 2 hold. Let  $\hat{\theta}_n^{\text{MRPL}}$  be a MRPLE. If  $\pi_n \rightarrow 0$  such that  $n\pi_n \rightarrow \infty$ , then  $\hat{\theta}_n^{\text{MRPL}} \rightarrow \theta_0$  in probability as  $n \rightarrow \infty$ .*

**Theorem 3.** *Suppose that Assumptions 1 and 2 hold. Let  $\hat{\theta}_n^{\text{MRPL}}$  be a MRPLE such that  $L_n^{\text{RPL}}(\hat{\theta}_n^{\text{MRPL}}) \geq L_n^{\text{RPL}}(\theta)$  for all  $\theta \in \Lambda$ , where  $\Lambda$  is a compact subset of  $\Theta$  and  $\theta_0$  is an interior point of  $\Lambda$ . If  $\pi_n \rightarrow 0$  such that, for all  $\kappa > 0$  and all  $a \in \mathcal{A}$ ,*

$$\frac{1}{\pi_n} \mathbb{E} \Phi_a(X_1^{(a)}; \theta_0)^4 \exp\left(\frac{-n\pi_n\kappa}{\sum_{a \in \mathcal{A}} \Phi_a(X_1^{(a)}; \theta_0)^2}\right) \rightarrow 0, \quad (4)$$

where  $\Phi_a(X_1^{(a)}; \theta)$  is the maximum of  $|\partial \ell_a(X_1^{(a)}; \theta) / \partial \theta_{i_1}|$ ,  $|\partial^2 \ell_a(X_1^{(a)}; \theta) / (\partial \theta_{i_1} \partial \theta_{i_2})|$  and  $\Psi_a(X_1^{(a)})$  over all possible indices  $1 \leq i_1, i_2 \leq q$ , then, as  $n \rightarrow \infty$ ,  $\sqrt{n\pi_n}(\hat{\theta}_n^{\text{MRPL}} - \theta_0)$  converges to a centered Gaussian distribution with variance-covariance matrix given by  $S^{-1}$ .

Subject to conditions on  $n$  and  $\pi_n$  (discussed in more detail next), Theorem 3 predicts that a reasonable approximation of the MRPLE's variance is given by  $S^{-1}/(n\pi_n)$ . In comparison with the previous formula  $S^{-1}CS^{-1}/n + S^{-1}/(n\pi)$ , the term  $S^{-1}CS^{-1}/n$ , which is the only term that involves the correlations between the scores, has disappeared. This

can be exploited to build approximate confidence intervals without the need to estimate the onerous matrix  $C$ .

Let us come back to the conditions on  $n$  and  $\pi_n$ . First, notice that Theorems 2 and 3 are consistent with each other, because the condition (4) implies  $n\pi_n \rightarrow \infty$ . To benefit from the approximation suggested by Theorem 3, the sampling parameter  $\pi_n$  must be small, but not too much; the meaning of “not too small” is captured by the condition (4), which in particular implies that  $n\pi_n$  must be large enough.

**Remark 3.** *In practice, how can one choose a value for  $\pi$ ? To benefit from the variance approximation of Theorem 3, we know that  $\pi$  must be small and  $n\pi$  must be large enough. The smaller  $\pi$ , the better the approximation, but then the MRPLE’s variance will increase. Simulations can help to choose  $\pi$ . One can identify a region  $(\pi, n\pi) \in (0, a] \times [b, \infty)$  for some constants  $a$  and  $b$ , where the approximation seems to be good enough. Next, we can choose  $\pi$  to lie within this region and adjust it according to what we consider to be an acceptable value for the standard errors. In other words, we can set  $\pi$  arbitrarily within the identified region, calculate the standard errors with the approximation  $S^{-1}/(n\pi)$  and determine if they are acceptable. If not, we may increase  $\pi$  a little (as long as it stays within the region) to decrease the standard errors.*

Translating the condition (4) into a more transparent condition on  $\pi_n$  is not always easy. A simple case is that of smooth models with a compact support, because the derivatives are bounded.

**Proposition 3.** *Suppose that, in Assumption 1, the first and second derivatives and the functions  $\Psi_\alpha$  are bounded in absolute value by some constant. If  $\pi_n \rightarrow 0$  such that  $n\pi_n^2 \rightarrow \infty$ , then (4) is satisfied.*



Under the conditions of Proposition 3, choosing  $\pi_n = n^{-\alpha}$ ,  $0 < \alpha < 1/2$ , makes  $n^{(1-\alpha)/2}(\hat{\theta}_n^{\text{MRPL}} - \theta_0)$  go to a Gaussian limit. The sampling parameter  $\pi_n$  can decrease almost as fast as  $1/\sqrt{n}$ . Another example that satisfies (4) is given in Section 4.

## 4 Example of the exchangeable standard Gaussian model

The exchangeable standard Gaussian model (Cox and Reid, 2004) is a model where explicit calculations are feasible and hence facilitates our understanding of the randomized pairwise likelihood method. The density of the Gaussian model with a common correlation parameter and standard Gaussian univariate margins is proportional to

$$f(x; \theta) \propto |\Sigma_\theta|^{-1/2} \exp\left(-\frac{1}{2}x^\top \Sigma_\theta^{-1}x\right), \quad (5)$$

$x \in \mathbf{R}^d$ ,  $\theta \in (-1/(d-1) + \epsilon, 1 - \epsilon) =: \Theta$ ,  $\epsilon > 0$ , where  $\Sigma_\theta$  is a variance-covariance matrix with 1s on its diagonal and  $\theta$  elsewhere. The matrix  $\Sigma_\theta$  is always positive-definite because  $\Theta \subset (-1/(d-1), 1)$ . The addition of  $\pm\epsilon$  at both ends of  $\Theta$  allows it to be enlarged to a compact interval on which continuous functions can be bounded, which helps to satisfy the assumptions. If  $a = \{i, j\}$  then

$$\ell_a(x_i, x_j; \theta) = -\frac{\log(1 - \theta^2)}{2} - \frac{x_i^2 + x_j^2}{2(1 - \theta^2)} + \frac{\theta x_i x_j}{1 - \theta^2} + \text{constant},$$

where  $x_i$  and  $x_j$  are the  $i$ th and  $j$ th components of  $x$ , respectively.

**Proposition 4.** *If  $\{f(\bullet; \theta), \theta \in \Theta\}$  is the Gaussian model (5) then Assumptions 1 and 2 hold.*

Proposition 4 is trivial. In the proof, the assumptions are checked directly.

## 4.1 A class of asymptotically normal estimators

Let  $\pi_n = n^{-\alpha}$ ,  $\alpha > 0$ , and let  $\hat{\theta}_n^{\text{MRPL}}(\alpha)$  be a MRPLE. In this setting MRPLEs depend on  $\alpha$  because they are maximizers of the randomized pairwise likelihood, which depends on  $\pi_n$  through the weights. Clearly,  $\alpha < 1$ ; otherwise the estimator has no chance to be consistent. Hence a class of estimators  $\{\hat{\theta}_n^{\text{MRPL}}(\alpha), 0 < \alpha < 1\}$  has been defined and we may wonder whether all members of this class are asymptotically normal.

**Proposition 5.** *If  $\{f(\bullet; \theta), \theta \in \Theta\}$  is the Gaussian model (5) and  $\pi_n = n^{-\alpha}$ ,  $0 < \alpha \leq 1/4$ , then (4) is satisfied.*

Proposition 5 gives the precise rate at which the estimators go to a limit distribution. Corollary 1 below is an immediate consequence.

**Corollary 1.** *If  $\{f(\bullet; \theta), \theta \in \Theta\}$  is the Gaussian model (5) and  $\hat{\theta}_n^{\text{MRPL}}(\alpha)$  is a MRPLE with  $0 < \alpha \leq 1/4$  then  $n^{(1-\alpha)/2}(\hat{\theta}_n^{\text{MRPL}}(\alpha) - \theta_0) \rightarrow N(0, 2/[d(d-1)E\dot{\ell}_{12}^2])$ , as  $n \rightarrow \infty$ , where  $E\dot{\ell}_{12}^2 = E[\partial\ell_{12}(X_{11}, X_{12}, \theta)/\partial\theta]_{\theta=\theta_0}^2 = (\theta_0^6 - \theta_0^4 - \theta_0^2 + 1)/(1 - \theta_0^2)^4$ .*

The parameter  $\alpha$  controls the compromise between the computational cost and the statistical efficiency of the estimator. If  $\alpha$  is large then the computational burden will be reduced but there will be a loss of statistical efficiency. If  $\alpha$  is small the reverse is true. In any case,  $\pi_n$  cannot go to zero too fast. Compare the admissible range of values for  $\alpha$  in Corollary 1 with the range  $0 < \alpha \leq 1/2$  found in Proposition 3. In Proposition 3 the sampling parameter was allowed to go to zero faster because the assumed model had lighter (in fact, bounded) tails than the Gaussian model. The formulas for the cross-correlations are given by the equations  $(1 - \theta_0^2)^4 E\dot{\ell}_{12}\dot{\ell}_{13} = \theta_0^2(1 - \theta_0^2)^2 - 4\theta_0^2(1 - \theta_0^2) + 2\theta_0^2(1 + \theta_0^2)(1 - \theta_0^2) + 6\theta_0^2(1 + \theta_0^2) - 2\theta_0^2(1 + \theta_0^2)(4 + 2\theta_0) + \theta_0(1 + \theta_0^2)^2(1 + 2\theta_0)$  and  $(1 - \theta_0^2)^4(E\dot{\ell}_{12}(\dot{\ell}_{13} - \dot{\ell}_{34})) = (1 + \theta_0^2)\theta_0(1 - \theta_0)(1 + \theta_0^2 - 4\theta_0) + 2\theta_0^2(1 - \theta_0^2)$ .

## 4.2 Comparison to the subset selection method

Remember that the subset selection method consists of choosing a subset of pairs  $\mathcal{B} \subset \mathcal{A}$ , and makes the inference rest on those pairs, taking all of the observations. On the contrary, the randomized pairwise likelihood method draws at random both observations and pairs, and makes the inference rest on those “(observation, pair)” couples for which both the observation and the pair have been selected.

Next, the two methods are compared for the exchangeable standard Gaussian model. For simplicity, put  $L_{ij,kl} = E \dot{\ell}_{ij} \dot{\ell}_{kl}$ ,  $L_{ij} = E \dot{\ell}_{ij}^2$ ,  $|\mathcal{B}| = B \leq A = |\mathcal{A}|$ . To make the methods comparable, set  $\pi = B/A$ , so that, on average, both the randomized pairwise likelihood and the pairwise likelihood based on the set of pairs  $\mathcal{B}$  have the same computational cost, measured by the number of times the density of a bivariate Gaussian distribution is evaluated. As in Section 3.2, let  $V(\pi) = V(B/A)$  be the asymptotic variance of the MRPLE. Denote by  $W(B) = W(A\pi)$  the asymptotic variance of the estimator obtained from the subset selection method.

According to Theorem 1, for  $B \geq 2$ , we have

$$\begin{aligned} W(B) &= \left( \sum_{a \in \mathcal{B}} L_a \right)^{-1} \sum_{a \neq b \in \mathcal{B}} L_{a,b} \left( \sum_{a \in \mathcal{B}} L_a \right)^{-1} + \left( \sum_{a \in \mathcal{B}} L_a \right)^{-1} \\ &= L_{12}^{-2} B^{-2} [\beta L_{12,13} + (B(B-1) - \beta) L_{12,34}] + L_{12}^{-1} B^{-1}, \end{aligned}$$

and  $V(B/A) = L_{12}^{-2} A^{-2} [\alpha(L_{12,13} - L_{12,34}) + A(A-1)L_{12,34}] + L_{12}^{-1} B^{-1}$ , where above  $\alpha = 6 \binom{d}{3}$  and  $\beta$  are the numbers of couples of pairs that share an index, among all possible couples of pairs and among all couples of pairs in  $\mathcal{B}$ , respectively. When  $A \rightarrow \infty$ , note that  $V(B/A) \sim L_{12}^{-2} L_{12,34} + L_{12}^{-1} B^{-1}$ .

It can be checked numerically from the formulas at the end of Section 4.1 that  $L_{12,13} -$

$L_{12,34} \geq 0$  and hence the best possible subset selection method is obtained when  $\beta = 0$ , leading to  $W(B) = L_{12}^{-2}L_{12,34} + (1 - L_{12}^{-1}L_{12,34})L_{12}^{-1}B^{-1}$ . The worst possible subset selection method is obtained when  $\beta = B(B - 1)$ , leading to  $W(B) = L_{12}^{-2}L_{12,13} + (1 - L_{12}^{-1}L_{12,13})L_{12}^{-1}B^{-1}$ . Note that setting  $\beta = 0$  or  $\beta = B(B - 1)$  may or may not be possible, depending on the choice of  $\mathcal{B}$ .

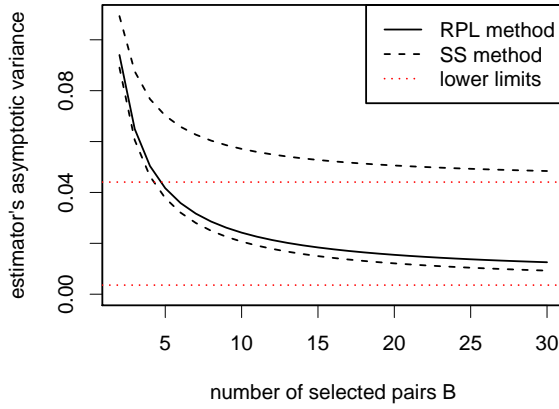


Figure 1: Asymptotic variances  $V(\pi) = V(B/A)$  and  $W(B)$  for the randomized pairwise likelihood (RPL) method and the best possible subset selection (SS) method, respectively, in the exchangeable standard Gaussian model with  $d = 50$  and  $\theta_0 = 0.7$ . The lower limits correspond to the limit values of  $W(B)$ , as  $B$  tends to infinity, for the cases  $\beta = B(B - 1)$  (above) and  $\beta = 0$  (below). The latter limit is also that of  $V(B/A)$  as  $A$  and  $B$  tend to infinity.

Figure 1 displays the values of  $V(B/A)$  and  $W(B)$  for  $B = 2, \dots, 30$  in the case  $d = 50$ . The curve for  $V(B/A)$  closely follows that for the best possible subset selection method.

### 4.3 An excursion to the infinite dimensional case

In this section we shall assume that the dimension increases as the sample size goes to infinity, that is, we shall let  $d = d_n$  go to infinity as  $n$  goes to infinity. In the special case of the exchangeable Gaussian model, there is only one parameter  $\theta_0$  to estimate whatever  $n$ , in spite of the growth in dimensions of the underlying Gaussian distribution. It turns out that the consistency of the MRPLE remains valid, regardless of the rate at which  $d_n$  increases.

**Proposition 6.** *If  $d_n \rightarrow \infty$ ,  $\pi_n \rightarrow 0$  and  $n\pi_n \rightarrow \infty$  then  $|\theta_n^{MRPL} - \theta_0| \xrightarrow{P} 0$  as  $n \rightarrow \infty$ .*

In this example, considering more and more variables means adding more and more pairs to the pairwise likelihood, and each of those pairs brings some information about  $\theta_0$ . It is therefore not surprising that the rate of  $d_n$  has no influence on the estimator's consistency; there is no condition of the kind " $d_n^2/n \rightarrow 0$ " or " $d_n/n \rightarrow 0$ " such as those found, for instance, in Portnoy (1988). Plausibly some condition of this kind would be needed for asymptotic normality, and even for consistency in more general correlation structures where the number of unknown parameters would grow with the dimension. We leave this problem for future work.

## 5 Copula models for multivariate count data

The problem of defining relevant models in high dimensions for discrete data has been addressed by various approaches (Chatelain et al., 2009; Karlis and Meligkotsidou, 2005; Berkhout and Plug, 2004; Karlis and Meligkotsidou, 2007; Chiquet et al., 2018, 2019).

An interesting approach uses copulas (Nelsen, 2006), allowing one to easily control the model marginals (Zhao and Joe, 2005; Nikoloulopoulos, 2013). However, inference

raises computational problems, which may be mitigated by using the randomized pairwise likelihood. Let  $m_1, \dots, m_{d+1}$  be natural integers with sum equal to  $q$ . Let  $\{F_i(\cdot; \mu_i), \mu_i \in \Theta_i \subset \mathbf{R}^{m_i}\}, i = 1, \dots, d$ , be families of univariate distribution functions. For every  $\mu_i \in \Theta_i$ , the distribution function  $F_i(\cdot; \mu_i)$  is also denoted by  $F_{\mu_i}$ . Let  $\{C(\bullet; \rho), \rho \in \Theta^{\text{cop}} \subset \mathbf{R}^{m_{d+1}}\}$  be a family of copulas defined on  $[0, 1]^d$ . For each  $\theta := (\mu_1, \dots, \mu_d, \rho) \in \Theta := \Theta_1 \times \dots \times \Theta_d \times \Theta^{\text{cop}}$ , the function defined by  $F(x_1, \dots, x_d; \theta) = C(F_{\mu_1}(x_1), \dots, F_{\mu_d}(x_d); \rho), x_1, \dots, x_d \in \mathbf{R}$ , is a well-defined distribution function on  $\mathbf{R}^d$  with marginals  $F_{\mu_1}, \dots, F_{\mu_d}$ . It is easy to show that if the univariate distribution function families and the copula family are identifiable then the resulting family of multivariate distribution functions is identifiable, too. From Sklar's theorem (Sklar, 1959), the copula is unique if the marginal distribution functions are continuous. In the discrete case, the copula is not unique in general but it still permits the construction of valid parametric statistical models. The difference with the continuous case is that the copula parameter alone does not characterize the dependence between the random variables at play (Genest and Nešlehová, 2007). Nevertheless, well-defined copula-based models are well-defined statistical models, to which the usual inference methods apply.

When the data are discrete, the probability mass function associated with the model  $C(F_{\mu_1}(x_1), \dots, F_{\mu_d}(x_d); \rho)$  is given by  $\sum_{(v_1, \dots, v_d)} \text{sgn}(v_1, \dots, v_d) C(F_{\mu_1}(v_1), \dots, F_{\mu_d}(v_d); \rho)$ , where the sum is over all  $(v_1, \dots, v_d) \in \{x_1 - 1, x_1\} \times \dots \times \{x_d - 1, x_d\}$ , and  $\text{sgn}(v_1, \dots, v_d) = 1$  if there is an even number of components  $v_j$  satisfying  $v_j = x_j - 1$ , and  $\text{sgn}(v_1, \dots, v_d) = -1$  if there is an odd number of components  $v_j$  satisfying  $v_j = x_j - 1$  (Panagiotelis et al., 2012). This sum, which has  $2^d$  terms, becomes intractable as the dimension increases. To perform the inference, it is computationally advantageous to use the pairwise likelihood. The functions  $\ell_a$  appearing in the pairwise likelihood formula (1) are expressed in terms of the

copula and the marginals: for every  $a = (i, j)$ ,

$$\begin{aligned} \ell_a(x_i, x_j; \mu_i, \mu_j, \rho) = & \log \left[ C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); \rho) - C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j - 1); \rho) \right. \\ & \left. - C_a(F_{\mu_i}(x_i - 1), F_{\mu_j}(x_j); \rho) + C_a(F_{\mu_i}(x_i - 1), F_{\mu_j}(x_j - 1); \rho) \right], \end{aligned}$$

where  $C_a(u_i, u_j; \rho) := C(1, \dots, u_i, \dots, u_j, \dots, 1; \rho)$  (all arguments have been replaced by ones but at the  $i$ th and  $j$ th positions) is the bivariate copula corresponding to the pair  $a$ , so that  $C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); \rho) = F_a(x_i, x_j; \theta)$ , where here  $F_a(\cdot, \cdot; \theta)$  denotes the bivariate distribution function corresponding to the pair  $a$ . Randomization of the pairwise likelihood pushes further the computational gain because not all of the  $nd(d-1)/2$  bivariate probability mass functions need to be evaluated and because tractable approximate confidence intervals can be calculated when  $\pi_n$  is small.

Recall that Assumption 2 is critical to the success of pairwise likelihood methods. It is satisfied if the conditions in Proposition 7 hold.

**Proposition 7.** *Suppose that the univariate distribution function families  $\{F_i(\cdot; \mu_i), \mu_i \in \Theta_i \subset \mathbf{R}^{m_i}\}$ ,  $i = 1, \dots, d$ , are identifiable. If, for every  $a \in \mathcal{A}$ , there is a function  $w_a$  on  $\Theta^{cop}$  into some Euclidean space and a family of bivariate copulas  $\{\tilde{C}_a(\cdot, \cdot; \varrho), \varrho \in \text{range } w_a\}$  such that*

(i) *the family  $\{\tilde{C}_a(\cdot, \cdot; \varrho), \varrho \in \text{range } w_a\}$  is identifiable*

(ii) *the copulas  $\tilde{C}_a(\cdot, \cdot; w_a(\rho)) = C_a(\cdot, \cdot; \rho)$  coincide for all  $\rho \in \Theta^{cop}$*

(iii) *the mapping  $W(\rho) := (w_a(\rho))_{a \in \mathcal{A}}$  is one-to-one*

*then Assumption 2 holds.*

The conditions in Proposition 7 are verifiable for at least some classes of models. For models based on the Gaussian copula, that is,

$$C(u_1, \dots, u_d; \rho) = \Phi_d(\Phi_1^{-1}(u_1), \dots, \Phi_1^{-1}(u_d); R(\rho)), \quad u_1, \dots, u_d \in (0, 1), \quad (6)$$

where  $\Phi_d(\bullet; R(\rho))$  is the distribution function of a standard  $d$ -variate Gaussian distribution with correlation matrix  $R(\rho)$ , it all depends on the structure of the correlation matrix. Simple suitable structures are given in Examples 1 and 2. More complex and suitable structures can be built.

**Example 1.** Let  $C$  be the Gaussian copula (6) with correlation matrix  $R(\rho)$ , where  $R(\rho)$  has 1s on its diagonal and  $\rho$  elsewhere,  $\rho \in (-1/(d-1), 1) =: \Theta^{cop}$ . Put  $w_a(\rho) = \rho$  so that  $\text{range } w_a = (-1/(d-1), 1)$ . The mapping  $W$  is one-to-one. Set  $\tilde{C}_a(\cdot, \cdot; \varrho)$  to be a bivariate Gaussian copula with correlation  $\varrho \in (-1/(d-1), 1)$ . Then clearly the family  $\{\tilde{C}_a(\cdot, \cdot; \varrho), \varrho \in (-1/(d-1), 1)\}$  is identifiable and the copulas  $\tilde{C}_a(\cdot, \cdot; w_a(\rho))$  and  $C_a(\cdot, \cdot; \rho)$  coincide for all  $\rho \in \Theta^{cop}$ . (Remember that  $C_a$  is the marginal of  $C$  corresponding to the pair  $a$ .)

**Example 2.** Let  $C$  be the Gaussian copula (6) with correlation matrix  $R(\rho)$ , where  $R(\rho)$  has 1s on its diagonal and  $\rho_{ij}$  at its  $i$ th row and  $j$ th column, with  $\rho = (\rho_{12}, \dots, \rho_{d-1,d}) \in (-1, 1)^{d(d-1)/2}$  such that  $R(\rho)$  is nonnegative definite. Let  $\Theta^{cop}$  be this space. If  $a = \{i, j\}$  then put  $w_a(\rho) = \rho_{ij}$  so that  $\text{range } w_a \subset (-1, 1)$ . Let  $\tilde{C}_a(\cdot, \cdot; \varrho)$  be the bivariate Gaussian copula with correlation  $\varrho \in (-1, 1)$ . The family  $\{\tilde{C}_a(\cdot, \cdot; \varrho)\}$  indexed by  $(-1, 1)$  is identifiable and hence so is this family restricted to  $\text{range } w_a$ . Moreover,  $\tilde{C}_a(\cdot, \cdot; \rho_{ij}) = C_a(\cdot, \cdot; \rho_{ij})$  for all  $\rho_{ij} \in \text{range } w_a$ . The mapping  $W$  is one-to-one.



## 6 Numerical illustrations

In both Section 6.1 and Section 6.2, 500 synthetic datasets of size  $n$  and dimension  $d$  are generated from a Gaussian copula and unit Poisson marginals. We used the `copula` R package (Yan, 2007). In Section 6.1,  $n \in \{100, 500, 1000\}$  and  $d = 30$ . In Section 6.2,  $n \in \{500, 1000, 5000\}$ ,  $d = 3$  and the copula parameters are given by  $\rho_{12} = 0.3$ ,  $\rho_{13} = 0.2$  and  $\rho_{23} = 0$ . A complementary set of simulations for the Gaussian exchangeable case can be found in Section A in the Supplementary Material.

### 6.1 Effect of the sampling parameter on the estimator’s performance and computational gains

We investigated the trade-off between efficiency and computational time for the randomized pairwise likelihood approach with  $d = 30$ . We considered two different cases: (1) a blockwise exchangeable correlation structure (for three blocks of dimension 10), corresponding to a total of 6 distinct copula parameters; and (2) a factorized correlation structure, where the element at the  $i$ th row and  $j$ th column of the copula correlation matrix  $R(\rho)$  is given by  $R(\rho)_{ij} = \rho_i \rho_j$ ,  $\rho = (\rho_1, \dots, \rho_d)$ , corresponding to a total of 30 distinct copula parameters. For the blockwise exchangeable case, true copula parameters were set to  $\rho = (0.75, 0.5, 0.25, 0.75, 0.5, 0.75)$  in lexicographical order. For the one-factor case,  $\rho$  was set to 30 equally spaced values between 0.1 and 0.9. Mean parameters were initialized using marginal means. Copula parameters for the blockwise exchangeable and one-factor simulations were respectively initialized using blockwise-averaged Pearson correlations or by optimizing the factorized correlations  $\rho_i \rho_j$  by minimizing the Euclidean distance to the Pearson correlation matrix. The randomized pairwise likelihood was applied with

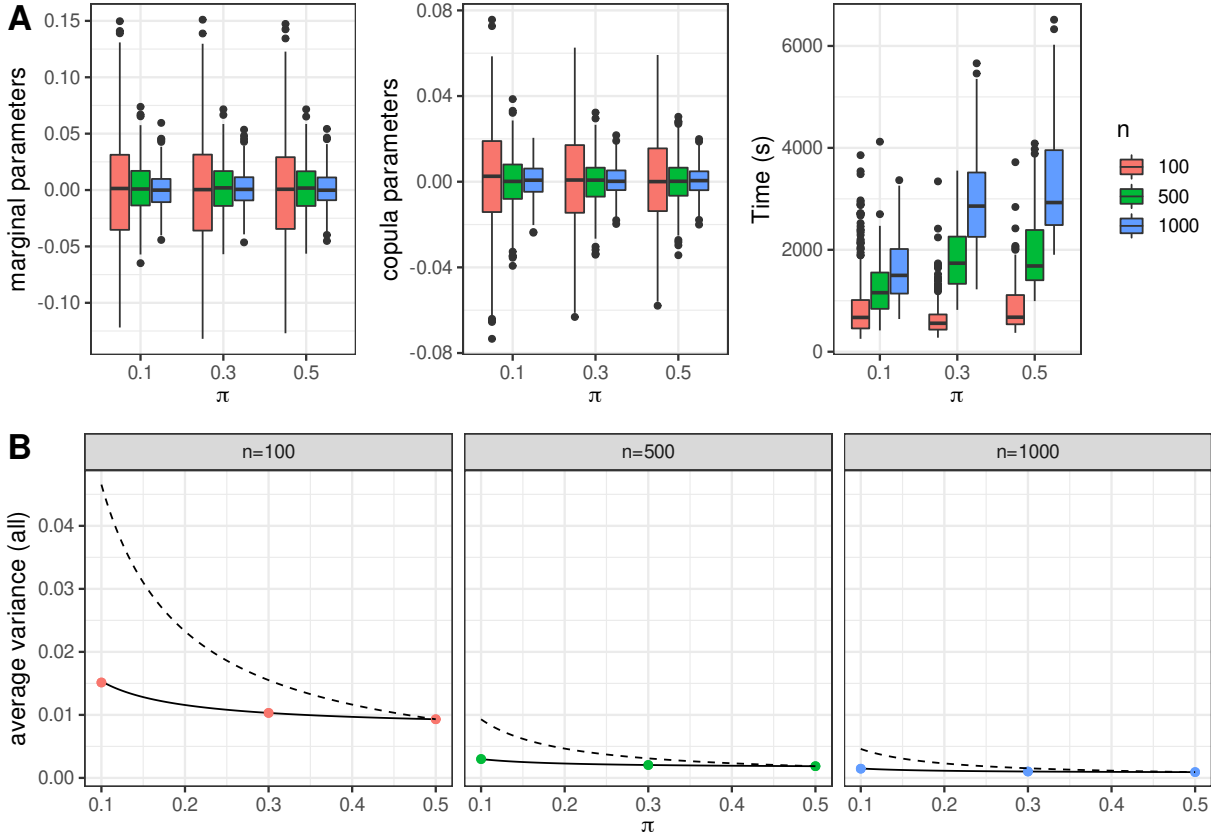


Figure 2: Performance of the randomized pairwise likelihood in the one-factor multivariate Poisson simulations with  $d = 30$  over 500 replications. (A) boxplot of the averaged centered estimates for the marginal parameters (left) and the copula parameters (middle), and the corresponding computational times in seconds (right). (B) Averaged variance estimates across parameters (points) for different values of  $\pi$ . The solid line connecting the points corresponds to the theoretical prediction for  $\pi = 0.1, 0.3$  knowing the variance at  $\pi = 0.5$ . The dotted line corresponds to the theoretical prediction under the assumption of a homogeneous inflation factor, knowing the variance at  $\pi = 0.5$ .

$$\pi \in \{0.1, 0.3, 0.5\}.$$

Results for efficiency and computational time of the randomized pairwise likelihood in the one-factor setting is shown in Figure 2. We remark that estimates are unbiased for all

values of  $n$  and  $\pi$ . The increase in variance as  $\pi$  decreases is not visible in the boxplots, but it does exist and agrees with theoretical predictions, see Figure 2B. It is accompanied by a decrease of computational time. We also computed the average absolute relative error for the mean parameters and the factorized correlations in Supplementary Figure S6: it increases as  $\pi$  and  $n$  decreases, although we note a greater impact in the effect of increasing  $n$  as compared to increasing  $\pi$ . Results for the blockwise exchangeable setting are shown in Supplementary Figure S7.

## 6.2 Coverage for the confidence intervals

We next sought to evaluate the asymptotic coverage of the confidence intervals constructed for the MRPLE for multivariate count data. To apply the randomized pairwise likelihood estimation procedure, we first initialized parameter values using the marginal means of each variable and the Pearson correlation of each pair of variables. Finally, we maximized the randomized pairwise likelihood with sampling parameter  $\pi \in [0.01, 0.60]$ .

Confidence intervals of level 95% based on the approximation  $S^{-1}/(n\pi)$  suggested by Theorem 3 were calculated for each parameter and each dataset. (An estimate of  $S$  was obtained from the data by the method-of-moments.) Coverage of the confidence intervals was computed as the proportion of replications for which the true parameter values were within the 95% confidence intervals. Results, corresponding to the mean coverage for the marginal and copula parameters, are presented in Figure 3.

In Figure 3(a), the coverage gets closer to its 95% target as  $\pi$  decreases from 0.60, agreeing with asymptotic theory. Then, for  $n = 500$  and  $n = 1000$ , the coverage drops at  $\pi = 0.01$ . For such a small  $\pi$ , the product  $n\pi$ , equal to 5 and 10 respectively, is too small for any inference to be reliable. For  $n = 5000$ , corresponding to  $n\pi = 50$ , there is

no drop at  $\pi = 0.01$ . We observe that there is a fairly large region where the coverage remains stable, suggesting that randomized pairwise likelihood methods might be robust with respect to the choice of  $\pi$ .

In Figure 3(b), a different pattern appears. Coverage performance is best for large values of  $\pi$ , seemingly contradicting the theory saying that for  $S^{-1}/(n\pi)$  to be a good approximation of the MRPLE’s variance  $S^{-1}CS^{-1}/n + S^{-1}/(n\pi)$ ,  $\pi$  must be small enough. However, remember that this condition is sufficient, not necessary. A possible explanation to the seeming contradiction is that the term  $S^{-1}CS^{-1}$  be negligible with respect to  $S^{-1}$ . In this case  $S^{-1}/(n\pi)$  is always a good approximation, whatever the value of  $\pi$ . In the cases  $n = 500$  and  $n = 1000$ , the coverage performance degradation as  $\pi$  approaches zero happens sooner compared with the marginal parameters, plausibly because copula parameters are more difficult to estimate than marginal parameters, in general. To keep performance above some given level as  $\pi$  gets smaller and smaller, the sample size  $n$  must increase to keep “the effective sample size”  $n\pi$  large enough. We remark that in Figure 3, all the coverages that exceed 92% correspond to  $\pi \leq 0.25$  and  $n\pi \geq 125$ . Thus the region  $(\pi, n\pi) \in (0, 0.25] \times [125, \infty)$  may be a reasonable rule of thumb to choose  $\pi$  in practice.

## 7 Application on transcriptomic data

We illustrate the application of the randomized pairwise likelihood procedure on multivariate count data from a study on the remodeling of the transcriptome over the life cycle of *Varroa destructor*, a parasitic mite that represents a significant threat to the western honeybee. Full details about the experimental design and pre-processing of RNA sequencing (RNA-seq) data may be found in Mondet et al. (2018). Our goal is to evaluate overall

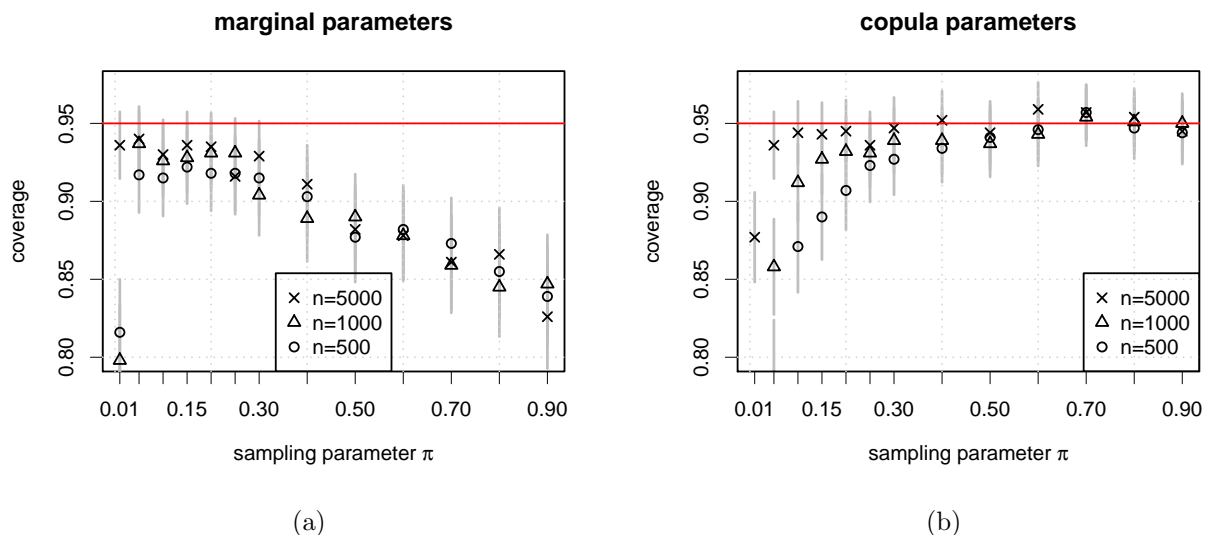


Figure 3: Coverage results for the marginal (a) and the copula (b) parameters.

transcriptome-wide correlations among different *Varroa* life stages from a single colony (R204) based on RNA-seq read counts for  $n=22,372$  contigs in  $d=10$  life cycle groups. In RNA-seq data, counts of expression are strongly positively associated with both the sequencing effort of each RNA sample (Robinson and Oshlack, 2010) and gene length; an offset accounting for these two factors are included in a Poisson generalized linear model (GLM) defined for the marginal distributions of each sample. To model the dependencies among life stages, these Poisson marginals were coupled with an unstructured Gaussian copula. Poisson GLM intercepts and Gaussian copula correlations were respectively initialized using marginal estimates and pairwise Pearson correlations, and the Nelder-Mead algorithm was used for optimization.

The randomized pairwise likelihood method was applied with  $\pi = 0.01$ , corresponding to  $n\pi = 224$  and standard errors of order less than  $10^{-4}$ , which was found to be sufficiently

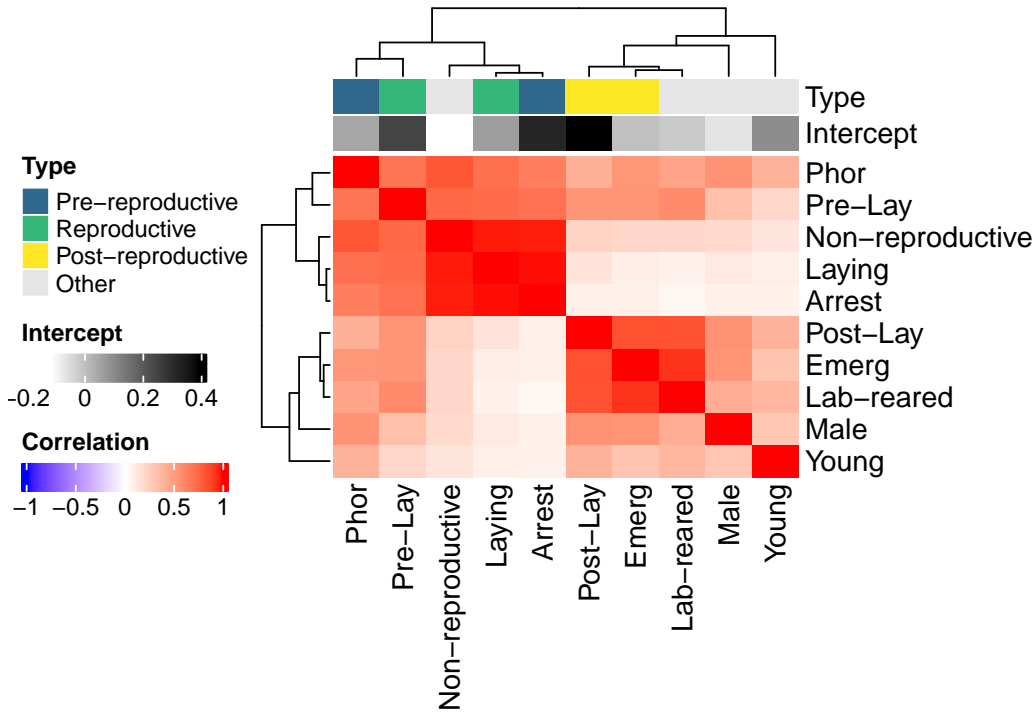


Figure 4: Clustered heatmap of the estimated copula parameters and per-sample intercepts (log-scale) for the *Varroa* life cycle transcriptome data, using the randomized pairwise likelihood ( $\pi = 0.01$ ) approach. Categorizations of life cycle groups according to reproductive status are included as a column annotation.

small with respect to the parameter estimate magnitudes. Standard errors for all the parameter estimates are given in Table S2. The value of  $\pi$  was chosen in accordance with Remark 3 and the region identified in the simulations of Section 6.2. Similar results, not included here, were observed for both  $\pi = 0.05$  and  $\pi = 1$ . A significant gain in computational time was observed: the maximization of the randomized pairwise likelihood with  $\pi = 0.01$  took 25 minutes, compared with 8.5 hours for the standard pairwise likelihood.

Figure 4 provides a visualization of the estimated copula parameters between life cycle

groups and marginal Poisson GLM intercepts based on the full dataset. We note a strong separation between the pre-reproductive/reproductive (phoretic, arresting, pre-laying, laying) versus post-reproductive (post-laying, emerging) phases. Non-reproductive females are clustered with reproductive females, supporting the hypothesis that mechanisms underlying reproductive failure occur before oogenesis in *Varroa* (Mondet et al., 2018). Lab-reared mites clustered with post-reproductive colony-collected females, suggesting that laboratory conditions do not provoke significant changes in the *Varroa* transcriptome. Two stages in particular exhibit distinct transcriptomic profiles as compared to the others: males (for which the largest estimated copula correlation of 0.56 is with post-lay females), and young mites, which are known to be characterized by a markedly immature physiology. Finally, the intercepts estimated for each marginal Poisson GLM provide intuition about the global over- or under-expression observed in each sample; the transcriptome appears to be most up-regulated in the transitions to (arrest and pre-lay) and from (post-lay) the reproductive stages.

In practice, transcriptome-wide analyses of RNA-seq data typically rely on the use of variance stabilizing transformations (e.g., log) before using exploratory methods such as principal components analysis, hierarchical clustering, or pairwise Pearson correlations; in this application, we have instead explicitly modelled the multivariate count nature of these transcriptome data via Poisson GLMs with appropriate offsets and a Gaussian copula to model the dependency structure among life stages.

## 8 Conclusions

The computational burden of pairwise likelihood methods can be reduced by randomization. Not only is the objective function easier to compute, but it also leads to easier computation of the confidence intervals, provided that the sampling parameter  $\pi$  is small enough and we have enough data. The proposed method is applicable in general but we focused on copula-based models for count data, in which the inference is challenging as soon as  $d$  is moderately large. We believe that the proposed method opens the door to designing affordable inference procedures in these models, and hence facilitating their use. To this end, we have implemented the randomized pairwise likelihood method for copula models of multivariate count data in the `rpl` R package, available at <https://github.com/andreamrau/rpl>. Randomized pairwise likelihood methods can also benefit other types of models, such as latent variable models, as alternatives to variational methods.

There is a downside to randomization, however. Since less data is used, the estimator's asymptotic variance increases. In some contexts the standard errors may still be small enough (as in Section 7), but in others they may not. In the latter case, an avenue for future research consists of optimizing several randomized pairwise likelihood in parallel and averaging the results. We expect the final estimator to be more efficient, see also Hector and Song (2020). This avenue is currently under investigation. Also the problem of obtaining theoretical results as  $d \rightarrow \infty$  is important, since in practice one benefits the most of randomization in high dimensional cases. This problem is difficult, but we obtained a first result in the exchangeable Gaussian model. We will pursue our efforts.

In the future, beyond the aforementioned points one could consider other sampling schemes to exploit known information about the data (such as temporal or spatial auto-correlation) or impose structural or sparsity constraints. For example, one could define a



threshold on the number of pairs sampled per observation or impose restrictions on the parameters—for instance, common correlations for some pairs. In addition, one could also consider alternative estimation strategies such as maximization by parts to split the full maximization problem into smaller ones.

## SUPPLEMENTARY MATERIAL

**Supplement to “A randomized pairwise likelihood method for complex statistical inferences”:** pdf file containing an additional simulation study, proofs and supplementary figures.

## Acknowledgements

We thank Mahendra Mariadassou for detailed comments on a first version of this manuscript and the Migale team of MaIAGE, INRAE, for providing us with a computing cluster. This work was supported by the INRAE DIGIT-BIO metaprogramme grant DINAMIC.

## References

- Bai, Y., J. Kang, and P. X.-K. Song (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics* 70(3), 661–670.
- Berkhout, P. and E. Plug (2004). A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica* 58(3), 349–364.
- Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518), 859–877.

- Chatelain, F., S. Lambert-Lacroix, and J.-Y. Tournet (2009). Pairwise likelihood estimation for multivariate mixed Poisson models generated by gamma intensities. *Statistics and Computing* 19(3), 283–301.
- Chiquet, J., M. Mariadassou, and S. Robin (2018). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics* 12(4), 2674–2698.
- Chiquet, J., S. Robin, and M. Mariadassou (2019). Variational inference for sparse network reconstruction from count data. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 1162–1171.
- Cox, D. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91(3), 729–737.
- Dillon, J. V. and G. Lebanon (2010). Stochastic composite likelihood. *Journal of Machine Learning Research* 11, 2597–2633.
- Ferrari, D., G. Qian, and T. Hunter (2016). Parsimonious and efficient likelihood composition by Gibbs sampling. *Journal of Computational and Graphical Statistics* 25(3), 935–953.
- Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA* 37(2), 475–515.
- Hector, E. C. and P. X.-K. Song (2020). A distributed and integrated method of moments for high-dimensional correlated data analysis. *Journal of the American Statistical Association* 116(534), 805–818.

- Huang, Z. and D. Ferrari (2021). Fast construction of optimal composite likelihoods. Preprint (arXiv:2106.05219).
- Joe, H. and Y. Lee (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis* 100(4), 670–685.
- Karlis, D. and L. Meligkotsidou (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing* 15(4), 255–265.
- Karlis, D. and L. Meligkotsidou (2007). Finite multivariate Poisson mixtures with applications. *Journal of Statistical Planning and Inference* 137, 1942–1960.
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80(1), 221–239.
- Lindsay, B. G., G. Y. Yi, and J. Sun (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* 21(1), 71–105.
- Mondet, F., A. Rau, C. Klopp, M. Rohmer, D. Severac, Y. Le Conte, and C. Alaux (2018). Transcriptome profiling of the honeybee parasite *varroa destructor* provides new biological insights into the mite adult life cycle. *BMC Genomics* 19(328).
- Nelsen, R. (2006). *An introduction to copulas*. Springer.
- Nikoloulopoulos, A. K. (2013). Copula-based models for multivariate discrete response data. In *Copulae in Mathematical and Quantitative Finance*, pp. 231–249. Springer.
- Panagiotelis, A., C. Czado, and H. Joe (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association* 107(499), 1063–1072.

- Papageorgiou, I. and I. Moustaki (2019). Sampling of pairs in pairwise likelihood estimation for latent variable models with categorical observed variables. *Statistics and Computing* 29, 351–365.
- Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics* 16(1), 356–366.
- Robinson, M. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11(R25).
- Sklar, A. (1959). Fonctions de répartition à  $n$  dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* 8, 229–231.
- Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica* 21(1), 5–42.
- Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika* 92(3), 519–528.
- Wang, X. and Y. Wu (2014). Theoretical properties of composite likelihoods. *Open Journal of Statistics* 4, 188–197.
- Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software* 21(4), 1–21.
- Zhao, Y. and H. Joe (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics* 33(3), 335–356.