# A randomized pairwise likelihood method for complex statistical inferences

Gildas Mazo, Dimitris Karlis, Andrea Rau

# A randomized pairwise likelihood method for complex statistical inferences

Gildas Mazo[1⋆], Dimitris Karlis[2], Andrea Rau[3,4]

[1]MaIAGE, INRAE, Université Paris Saclay, 78350 Jouy-en-Josas, France
[2]Athens University of Economics and Business
[3]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France
[4]BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne, 80200 Estrées-Mons, France
⋆Corresponding author: gildas.mazo@inrae.fr

**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 breaks the correlation structure between the individual bivariate likelihoods, allowing much more computationally inexpensive confidence intervals to be constructed. The proposed approach is illustrated in tandem with copula-based models for multivariate continuous and count data in simulations, and in real data from microbiome and transcriptome applications.

**Keywords:** *pairwise likelihood; composite likelihood; randomization; confidence intervals; mutivariate 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 the

high-throughput count data collected in molecular biology [28] and microbial ecology [17] applications. However, regardless of the multivariate model used for such data, the corresponding likelihood is often complex, costly to evaluate, or even intractable. To overcome this issue, a solution consists of maximizing a sum of lower-dimensional likelihoods, called a composite likelihood, instead of the full likelihood [25]. Often, the sum over all pairs of bivariate marginals is used, in which case the composite likelihood thus formed is called the pairwise likelihood. The advantage is computational, since it obviates the need to compute the full likelihood. In a large enough class of 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 [43]; we note that variational methods do not have this guarantee in general [2]. 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.

Pairwise likelihood methods have been successfully used in many applications, including correlated binary data [24], time series models [9], spatial models [42], mixed models for longitudinal profiles [12], extreme-value models [33] and image models [32]. Many variants of the composite likelihood method have been proposed in the literature to accommodate specific models, data or tasks. As one example, variable selection was performed in [15] in the context of multivariate mixed models. Also, several authors have proposed ways to improve the efficiency of composite likelihood methods, primarily by adding weights to the component likelihoods [20, 23, 45]. It appears, however, that finding and estimating the optimal weights in general is a very difficult problem which may not have a solution [26]. The composite likelihood approach was originally described in [25] and further developed during the last decade, see e.g. the review in [43]. In the following, we shall focus on the pairwise likelihood, the most popular version of composite likelihoods.

In the high dimensional context, the loss of efficiency may be less of a concern than the increase in computational complexity. If $d$ is the number of variables, then the number of pairs is of order $d^2$, which is large enough to make the application of the pairwise likelihood method cumbersome. The computation of confidence intervals is even more challenging: one needs to compute a double sum over pairs of pairs of order $d^4$ and all of the four-dimensional marginals. This is not only time-consuming, but also makes numerical instabilities more likely.

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 [14, 35], although selecting a good subset is a difficult problem. Some heuristics were proposed in [35], but no theoretical justification was provided and the asymptotic properties of the estimators are unknown. In [14], 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 marginal likelihoods was proposed in [10]. 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 $W_{nia}$, $i = 1, \ldots, n$, $a \in \{\{1, 2\}, \ldots, \{d-1, d\}\}$, with parameter $\pi_n$; all summands for which $W_{nia} = 0$ are discarded. A fundamental point here is that we allow the Bernoulli parameter $\pi_n$ to decrease with $n$—we shall come back to this later. The Bernoulli 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 drastic reduction of the computational cost. By letting $\pi_n \to 0$, we asymptotically break the correlation structure between the individual bivariate likelihoods thanks to the two-way random sampling mechanism. This has the important consequence that the term of computational complexity $d^4$ is removed in the asymptotic variance-covariance matrix of the resulting estimator. In practice, this means that we are able to compute approximate confidence intervals at a computational complexity cost of only $d^2$, involving only bivariate marginals.

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 investigates the randomized pairwise likelihood method, provides asymptotic results, both in the case where $\pi_n$ is fixed and $\pi_n \to 0$, and explains why the latter allows for a "cheap" approximation of confidence intervals. Section 4 analyses the specific case of Gaussian models, where explicit calculations are feasible, thus allowing a better understanding of the behavior of the proposed method. Section 5 reviews the state of the art for multivariate count data with a focus on copula-based models and shows how the randomized pairwise likelihood can contribute to it. Identifiability results are given for two correlation structures of the Gaussian copula. Section 6 reports simulation experiments to assess the behavior of the approach for multivariate continuous and count data, and Section 7 illustrates how the approach can be applied to a set of microbiome and transcriptome data with multivariate count data models based on Poisson marginals and Gaussian copulas. Concluding remarks may be found in Section 8.

# 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 [25, 43, 44]. 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 $X_i := (X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$, 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 $\{f(\bullet;\theta),\, \theta = (\theta_1, \ldots, \theta_q) \in \Theta\}$ 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 a subset of the set of all pairs of variables. Its cardinal is at most $d(d-1)/2$. 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 they are components. If $a = \{j, j'\}$ is a pair then $(X_{ij}, X_{ij'})$ is also denoted by $X_i^{(a)}$.

The pairwise log-likelihood function is given by

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

The population version of the pairwise log-likelihood function is $\sum_a L_a(\theta)$, where $L_a(\theta)$ stands for $\mathrm{E}\, \ell_a(X_1^{(a)};\theta)$. As usual, the goal is to estimate the maximizer of the population pairwise log-likelihood by maximizing the pairwise likelihood function. From the viewpoint of M-estimation theory, the population pairwise likelihood is the objective criterion function, the maximizer of which being 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. The authors in [46] 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 $\hat{\theta}_n^{\mathrm{MPL}}$ of $\Theta$ that satisfies $L_n^{\mathrm{PL}}(\hat{\theta}_n^{\mathrm{MPL}}) \geq L_n^{\mathrm{PL}}(\theta)$ 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*

$$\sup_{\theta\in\Theta} \left| \frac{\partial^3 \ell_a(X_1^{(a)};\theta)}{\partial\theta_{i_1}\partial\theta_{i_2}\partial\theta_{i_3}} \right| \leq \Psi_a(X_1^{(a)}),$$

*for all $1 \leq i_1 \leq i_2 \leq i_3 \leq q$. Finally, if $\mathfrak{m}_a$ stands for the base measure of which $f_a(\cdot,\cdot;\theta)$ is the density then $\int f_a(\cdot,\cdot;\theta)\,\mathrm{d}\mathfrak{m}$ and $\int(\partial/\partial\theta_{i_1})f_a(\cdot,\cdot;\theta)\,\mathrm{d}\mathfrak{m}$ 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 $\sum_a L_a(\theta)$. A sufficient condition for Assumption 2 to hold will be given in Section 5 in the context of copula-based models. In Section 4, Assumption 2 is checked directly.

**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 1 and Assumption 2 together imply that MPLEs are asymptotically normal. More precisely, we have that $\sqrt{n}(\hat{\theta}_n^{\mathrm{MPL}} - \theta_0)$ converges in distribution to a Gaussian random vector with mean zero and variance-covariance matrix

$$\left(\sum_{a\in\mathcal{A}} \mathrm{E}\,\dot{\ell}_a\dot{\ell}_a^\top\right)^{-1} \left(\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{A}} \mathrm{E}\,\dot{\ell}_a\dot{\ell}_b^\top\right) \left(\sum_{a\in\mathcal{A}} \mathrm{E}\,\dot{\ell}_a\dot{\ell}_a^\top\right)^{-1}, \tag{2}$$

where $\mathrm{E}\,\dot{\ell}_a \dot{\ell}_b^\top$ is a shorthand for $\mathrm{E}\,\dot{\ell}_a(X_1^{(a)};\theta_0)\dot{\ell}_b(X_1^{(b)};\theta_0)^\top$. This result is standard and known since at least [25] but, as it turns out, it is difficult to find in the literature precise conditions under which this result is true.

Assumption 2 is critical to ensure consistency of pairwise likelihood methods. Thus it is important to give verifiable conditions under which it holds.

**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 $\{\tilde{f}_a(\cdot,\cdot;\vartheta_a),\ \vartheta_a \in \operatorname{range} v_a\}$ such that*

*(i) the family $\{\tilde{f}_a(\cdot,\cdot;\vartheta_a),\ \vartheta_a \in \operatorname{range} v_a\}$ 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.*

Examples that satisfy the conditions of Proposition 1 will be given in Section 5 in the context of copula-based models.

To improve efficiency, weights could be added to the pairwise log-likelihood [25, 20, 26], leading to the maximization of

$$L_n^{\mathrm{WPL}}(\theta) = \frac{1}{n}\sum_{a\in\mathcal{A}} w_a \sum_{i=1}^{n} \ell_a(X_i^{(a)};\theta), \tag{3}$$

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". But this is not important, because if the weights are all positive then the conditions (i), (ii) and (iii) of Proposition 1 still suffice to check the modified assumption. (See the proof of Proposition 1.)

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 middle term in (2). This can be computationally challenging, as we shall see next. In the more realistic multivariate case, according to [25], a solution may not exist, and if it did 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 middle term in (2) 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 [35] but no theory is provided. In [14], 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 preliminary asymptotics

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

$$L_n^{\mathrm{RPL}}(\theta) = \frac{1}{n\pi_n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} W_{ni}^{(a)} \ell_a(X_i^{(a)}; \theta), \tag{4}$$

where, for each $n$, $W_{ni}^{(a)}$, $i = 1, \ldots, n$, $a \in \mathcal{A}$, are independent Bernoulli random variables with parameter $0 < \pi_n \leq 1$. They are assumed to be independent of $X_1, \ldots, X_n$. The unknown parameter $\theta_0$ is estimated by maximizing the function in (4). 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 (4) 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 (3) and (4) 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 Assumption 1 and Assumption 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*

$$\left(\sum_a \mathrm{E}\,\dot{\ell}_a\dot{\ell}_a^\top\right)^{-1}\left(\frac{\pi\,\mathrm{E}\sum_{a,b}\dot{\ell}_a\dot{\ell}_b^\top + (1-\pi)\sum_a \mathrm{E}\,\dot{\ell}_a\dot{\ell}_a^\top}{\pi}\right)\left(\sum_a \mathrm{E}\,\dot{\ell}_a\dot{\ell}_a^\top\right)^{-1}. \quad (5)$$

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

MRPLEs are asymptotically normal with an asymptotic variance-covariance matrix that depends on $\pi$. The numerator of the middle term in (5) can be rewritten as $\pi\,\mathrm{E}\sum_{a\neq b}\dot{\ell}_a\dot{\ell}_b^\top + \sum_a \mathrm{E}\,\dot{\ell}_a\dot{\ell}_a^\top$, where we see that the correlations between the scores appears with a factor $\pi$. An explanation for this is that the correlation structure is broken by the randomization introduced in (4). This is discussed further below. Note that Theorem 1 implies that, in probability, $\hat{\theta}_n^{\mathrm{MRPL}} \to \theta_0$ as $n \to \infty$. Choosing $\pi = 1$ allows us to recover the results of Section 2.

## 3.2 Advanced asymptotics to build cheap confidence intervals

Suppose that we want to build confidence intervals for $\theta_0$. Theorem 1 suggests that

$$\mathrm{Var}\,\hat{\theta}_n^{\mathrm{MRPL}} \approx \frac{S^{-1}(C-S)S^{-1}}{n} + \frac{S^{-1}}{n\pi},$$

where here $S = \sum_a \mathrm{E}\,\dot{\ell}_a \dot{\ell}_a^\top$ and $C = \sum_a \sum_b \mathrm{E}\,\dot{\ell}_a \dot{\ell}_b^\top$. The problem, as mentioned in Section 2.2, is that $C$ is difficult to compute: it requires up to $O(d^4)$ evaluations of a four-dimensional integral.

If $n\pi$ is much smaller than $n$, we expect $\mathrm{Var}\,\hat{\theta}_n^{\mathrm{MRPL}} \approx S^{-1}/(n\pi)$. This would be highly advantageous, because $S$ requires at most $O(d^2)$ computations of a 2-dimensional integral; compared with $O(d^4)$ computations of a 4-dimensional integral, the cost would thus be greatly reduced. However, if $n\pi$ is too small (with respect to $n$), we would use too little of the data and the MRPLE would be a poor estimate of $\theta_0$. The question is: how small can $n\pi$ be with respect to $n$, and notably, how small can $\pi$ be? To answer this question, we let $\pi = \pi_n \to 0$ as $n \to \infty$ and see whether we can still get the asymptotic normality of the estimators, and, if so, at what rate.

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

Theorem 2 is not a surprise. It says that the actual number of terms needed in the randomized pairwise likelihood must go to infinity, while at the same time it is allowed to be negligible with respect to the sample size. Notice that for a consistency result, many of the statements in Assumption 1 are not necessary.

**Theorem 3.** *Suppose that Assumption 1 and Assumption 2 hold. Let $\hat{\theta}_n^{MRPL}$ be 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 $\Lambda$. If $\pi_n \to 0$ such that, for all $\kappa > 0$ and all $a \in \mathcal{A}$,*

$$\frac{1}{\pi_n} \mathrm{E}\,\Phi_a(X_1^{(a)};\theta_0)^4 \exp\left(\frac{-n\pi_n\kappa}{\sum_{a\in\mathcal{A}} \Phi_a(X_1^{(a)};\theta_0)^2}\right) \to 0, \tag{6}$$

*where*

$$\Phi_a(X_1^{(a)};\theta) := \max_{i_1,i_2} \max\left(\left|\frac{\partial \ell_a(X_1^{(a)};\theta)}{\partial\theta_{i_1}}\right|, \left|\frac{\partial^2 \ell_a(X_1^{(a)};\theta)}{\partial\theta_{i_1}\partial\theta_{i_2}}\right|, \Psi_a(X_1^{(a)})\right)$$

*then, as $n \to \infty$,*

$$\sqrt{n\pi_n}\left(\hat{\theta}_n^{MRPL} - \theta_0\right) \xrightarrow{d} N\left(0, \left(\sum_{a\in\mathcal{A}} \mathrm{E}\,\dot{\ell}_a \dot{\ell}_a^\top\right)^{-1}\right). \tag{7}$$

Theorem 3 suggests exactly what we were looking for. Namely,

$$\mathrm{Var}\,\hat{\theta}_n^{\mathrm{MRPL}} \approx \frac{S^{-1}}{n\pi_n}, \tag{8}$$

whenever $n\pi_n$ is small with respect to $n$; "small" being captured by the condition (6). Notice that (6) implies $n\pi_n \to \infty$ and hence Theorem 2 and Theorem 3 are consistent with each other.

What has happened? Why is the numerator in (8) the cheap $S^{-1}$ and why has the arduous $S^{-1}CS^{-1}$ disappeared? It turns out that the randomization mechanism destroys the correlation structure between the scores $\dot{\ell}_a(X_1^{(a)};\theta_0)$, $a \in \mathcal{A}$. One way to see this is to rewrite $S^{-1}CS^{-1} = S^{-1} + S^{-1}DS^{-1}$, where here $D = \sum_{a \neq b} \mathrm{E}\, \dot{\ell}_a \dot{\ell}_b^\top$ and to notice that both sides of the equation are equal whenever $D = 0$.

Translating the condition (6) 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 2.** *Suppose that, in Assumption 1, the first and second derivatives and the functions $\Psi_a$ are bounded in absolute value by some constant. If $\pi_n \to 0$ such that $n\pi_n^2 \to \infty$ then (6) is satisfied.*

In the context of Proposition 2, if $\pi_n = n^{-\alpha}$, $0 < \alpha < 1/2$, then $n^{(1-\alpha)/2}(\hat{\theta}_n^{\mathrm{MRPL}} - \theta_0)$ goes to a Gaussian limit. The Bernoulli parameter $\pi_n$ can decrease almost as fast as $1/\sqrt{n}$. Another example that satisfies (6) is given in Section 4.

# 4 Example of the exchangeable standard Gaussian model

The exchangeable standard Gaussian model [8] 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 given by

$$f(x;\theta) = (2\varpi)^{-d/2} |\Sigma_\theta|^{-1/2} \exp\left(-\frac{1}{2} x^\top \Sigma_\theta^{-1} x\right), \tag{9}$$

$x \in \mathbf{R}^d$, $\theta \in (-1/(d-1) + \epsilon, 1 - \epsilon) =: \Theta$, $\epsilon > 0$, and

$$\Sigma_\theta = \begin{pmatrix} 1 & \cdots & \theta \\ & \ddots & \\ \theta & \cdots & 1 \end{pmatrix}.$$

(Above $\Sigma_\theta$ has 1 on its diagonal and $\theta$ elsewhere and $\varpi$ is such that $\sqrt{2\varpi} = \int e^{-x^2/2}\,\mathrm{d}x$.) 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 3.** *If* $\{f(\bullet;\theta),\, \theta \in \Theta\}$ *is the Gaussian model* (9) *then Assumption 1 and Assumption 2 hold.*

Proposition 3 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 $\hat{\theta}_n^{\mathrm{MRPL}}(\alpha)$ 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 $\{\hat{\theta}_n^{\mathrm{MRPL}}(\alpha),\, 0 < \alpha < 1\}$ has been defined and we may wonder whether all members of this class are asymptotically normal.

**Proposition 4.** *If* $\{f(\bullet;\theta),\, \theta \in \Theta\}$ *is the Gaussian model* (9) *and* $\pi_n = n^{-\alpha}$, $0 < \alpha \leq 1/4$*, then (6) is satisfied.*

Proposition 4 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* (9) *and* $\hat{\theta}_n^{MRPL}(\alpha)$ *is a MRPLE with* $0 < \alpha \leq 1/4$ *then*

$$n^{(1-\alpha)/2}(\hat{\theta}_n^{MRPL}(\alpha) - \theta_0) \to N\left(0, \frac{2}{d(d-1)\,\mathrm{E}\,\dot{\ell}_{12}^2}\right), \quad n \to \infty,$$

*where*

$$\mathrm{E}\,\dot{\ell}_{12}^2 = \mathrm{E}\,\frac{\partial \ell_{12}(X_{11}, X_{12}, \theta)}{\partial \theta}\bigg|_{\theta=\theta_0}^2 = \frac{\theta_0^6 - \theta_0^4 - \theta_0^2 + 1}{(1-\theta_0^2)^4}.$$

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 2. In Proposition 2 the Bernoulli parameter was allowed to go to zero faster because the assumed model had lighter (in fact, bounded) tails than the Gaussian model.

For the sake of completeness, we give the formulas for the cross-correlations:

$$\begin{aligned}(1-\theta_0^2)^4\,\mathrm{E}\,\dot{\ell}_{12}\dot{\ell}_{13} = {} & \theta_0^2(1-\theta_0^2)^2 - 4\theta_0^2(1-\theta_0^2) \\ & + 2\theta_0^2(1+\theta_0^2)(1-\theta_0^2) + 6\theta_0^2(1+\theta_0^2) - 2\theta_0^2(1+\theta_0^2)(4+2\theta_0) \\ & + \theta_0(1+\theta_0^2)^2(1+2\theta_0)\end{aligned}$$

and

$$(1-\theta_0^2)^4(\mathrm{E}\,\dot{\ell}_{12}(\dot{\ell}_{13} - \dot{\ell}_{34})) = (1+\theta_0^2)\theta_0(1-\theta_0)(1+\theta_0^2-4\theta_0) + 2\theta_0^2(1-\theta_0^2).$$

### 4.2 Insight into the impact of randomization on estimator precision

Although this is not true in general, removing the correlation structure can yield an improvement in the precision of the estimator. For instance, in the model (9) with $d = 6$ and $\theta_0 = 0.7$, we have compared the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{\theta}_n^{\mathrm{MRPL}} - \theta_0)$ to that of $\sqrt{n}(\hat{\theta}^{\mathcal{A}'} - \theta_0)$, where $\hat{\theta}^{\mathcal{A}'}$ is the standard pairwise likelihood estimator based on some subset of pairs $\mathcal{A}'$. The estimator $\hat{\theta}_n^{\mathrm{MRPL}}$ is based on the set of all pairs. The asymptotic variance-covariance matrix of the former is given by (5) and the the asymptotic variance-covariance matrix of the later is given by (2) with $\mathcal{A}$ replaced by $\mathcal{A}'$. To make the methods comparable, we set $\pi_n = 2|\mathcal{A}'|/(d(d-1))$, where $|\mathcal{A}'|$ is the size of $\mathcal{A}'$. Both methods require a computational budget equal to $|\mathcal{A}'|n = \pi_n nd(d-1)/2$. For each value of $|\mathcal{A}'|$, the subset of pairs was chosen randomly. The results are shown in Figure 1. In this case the budget divided by the sample size is $|\mathcal{A}'|$. When the budget is large with respect to the sample size, that is, when $\mathcal{A}'$ contains many pairs, or put differently still, when $\pi_n$ is large, both methods perform similarly. However, when fewer and fewer pairs are selected (i.e., moving from right to left in Figure 1), coinciding with a smaller and smaller $\pi_n$, the randomized pairwise likelihood method performs much better than the subset method.

This can be explained as follows. The variance of the estimator for the subset method is about

$$\operatorname{Var} \hat{\theta}^{\mathcal{A}'} \approx \frac{1}{n|\mathcal{A}'| \operatorname{E} \dot{\ell}_{12}^2} + \frac{\sum_{a,b \in \mathcal{A}',\, a \neq b} \operatorname{E} \dot{\ell}_a \dot{\ell}_b}{n|\mathcal{A}'|^2 (\operatorname{E} \dot{\ell}_{12}^2)^2},$$

and the variance of the MRPLE with $\pi_n = 2|\mathcal{A}'|/(d(d-1))$, when $|\mathcal{A}'|$ is small, is about $1/(n|\mathcal{A}'| \operatorname{E} \dot{\ell}_{12}^2)$; the difference, when $|\mathcal{A}'|$ is small, is about

$$\operatorname{Var} \hat{\theta}^{\mathcal{A}'} - \operatorname{Var} \hat{\theta}_n^{\mathrm{MRPL}} \approx \frac{\sum_{a,b \in \mathcal{A}',\, a \neq b} \operatorname{E} \dot{\ell}_a \dot{\ell}_b}{n|\mathcal{A}'|^2 (\operatorname{E} \dot{\ell}_{12}^2)^2}.$$

The only remaining term involves the cross-correlations between the scores. If the scores are positively correlated, then the randomized pairwise likelihood method will be better than the subset method. This in fact represents the most plausible situation in model (9); Figure 2 shows both (7) and (2) and we see that more often than not, (7) is smaller than (2).

## 5 Application to multivariate count models based on copulas

When working with multivariate count data, it is not straightforward to define appropriate models in high dimensions. Several research directions exist, including the generalization of lower-dimensional models (e.g., the bivariate Poisson

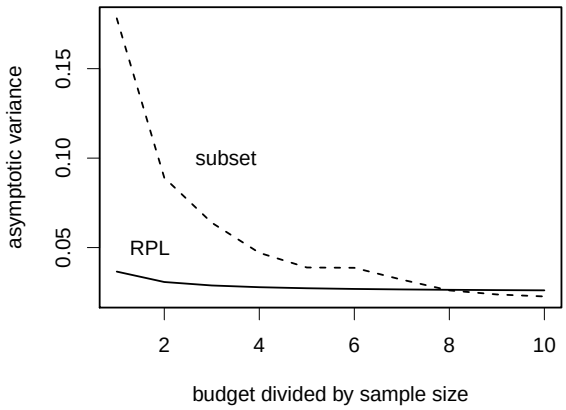


Figure 1: Asymptotic variances of $\sqrt{n}(\hat{\theta}_n^{\mathrm{MRPL}} - \theta_0)$ and $\sqrt{n}(\hat{\theta}_n^{\mathcal{A}'} - \theta_0)$ for different budgets.

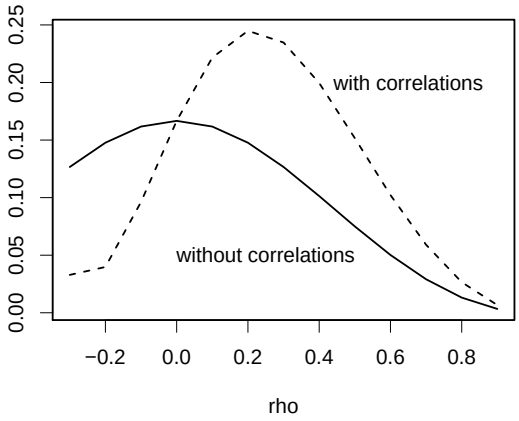


Figure 2: Values of the asymptotic variance of (2) ("with correlations") and (7) ("without correlations") in the model (9) with $d = 6$ and $\theta_0 =: \rho = 0.7$.

distribution) to higher dimensions (e.g., the multivariate Poisson distribution). In such cases, it is often necessary to specify a model that may be complicated and intractable in practice to obtain the full correlation structure of the data; see the details in [21]. Other proposals make use of models based on conditional distributions [1] or finite mixtures [22]. However, these strategies are generally greatly complicated by the discrete nature of the data as compared to similar models for continuous data, where the multivariate normal distribution is a cornerstone allowing for both great flexibility and feasible calculations; for a recent review see [16]. Another classic approach is the use of latent (continuous) variable models to describe interdependencies between the observed discrete variables. However, the joint distribution of the observable variables is often intractable, and inference typically relies on complex versions of the EM algorithm or variational methods [5, 6]. In the latter case, there are no theoretical guarantees on the estimators in general [2]. In addition, it may be difficult to control even simple aspects of the model, such as the marginal distributions.

An alternative interesting approach uses copulas [19, 29], which are quite common for continuous data [18] but not widely used for discrete data [see, e.g. 30, for a discussion of the challenges in applying copulas to multivariate counts]. Copula theory makes it quite simple to build multivariate models with the ability to control the marginals. Thus, for instance, one may replace without effort a Poisson marginal by a negative binomial marginal to account for overdispersion. However, for discrete data the difficulty lies not in the construction of the models, but rather in the inference. The computation of the likelihood of copula-based models for discrete data suffers from a combinatorial explosion [30]. Moreover, computational problems often arise from the need to invert large matrices and/or to approximate intractable integrals or sums.

We now consider models defined through copulas and use the randomized pairwise likelihood to facilitate their application for multivariate count data. A copula is a function which can "couple" the marginals to model the dependence structure [29]. A copula is a multivariate distribution function with uniform marginals. The importance of copulas in statistical modelling stems from Sklar's theorem [see, 29, §2.3], which shows that a copula corresponds to every multivariate distribution and, more importantly, provides a general mechanism to construct new multivariate models in a straightforward manner. More precisely, let

$$\{F_i(\cdot;\mu_i),\ \mu_i \in \Theta_i \subset \mathbf{R}^{m_i}\}, \tag{10}$$

$i = 1, \ldots, d$, 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

$$\{C(\bullet;\rho),\ \rho \in \Theta^{\text{cop}} \subset \mathbf{R}^{m_{d+1}}\} \tag{11}$$

be a family of copulas defined on $[0,1]^d$. For each $\theta := (\mu_1, \ldots, \mu_d, \rho) \in \Theta := \Theta_1 \times \cdots \times \Theta_d \times \Theta^{\text{cop}}$, the function defined by

$$F(x_1, \ldots, x_d; \theta) = C(F_{\mu_1}(x_1), \ldots, F_{\mu_d}(x_d); \rho), \tag{12}$$

$x_1, \ldots, x_d \in \mathbf{R}$, is a well-defined distribution function on $\mathbf{R}^d$ with marginals $F_{\mu_1}, \ldots, F_{\mu_d}$. For consistency with Section 2, we can assume $m_1 + \cdots + m_{d+1} = q$. It is easy to show that if the families (10) and the copula family (11) are identifiable then the family of multivariate distribution functions defined by (12) is identifiable, too. Note that if the marginal distribution functions are continuous then the copula is unique. 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 [13]. Nevertheless, since any well-defined copula-based model is a particular instance of a statistical model, the tools and the methods of the latter can be applied to the former.

For count data, a common starting point is to use the Poisson distribution for the marginals:

$$F_{\mu_j}(x_j) = \sum_{m=0}^{x_j} \frac{\mu_j^m}{m!} e^{-\mu_j}, \quad x_j = 0, 1, \ldots,$$

$\mu_j > 0$, $j = 1, \ldots, d$. We can then couple the marginals to add a dependence structure. In practice there are few copulas that can consider a full structure for $d$-dimensional data, allowing for flexible modelling of such data. A common choice is the Gaussian copula, given by

$$C(u_1, \ldots, u_d; \rho) = \Phi_d(\Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_d); R(\rho)), \quad u_1, \ldots, u_d \in (0, 1), \quad (13)$$

where $\Phi_d(\bullet; R(\rho))$ is the distribution function of a standard $d$-variate Gaussian distribution with correlation matrix $R(\rho)$.

Although copulas make model building straightforward, maximum likelihood inference lead to combinatorial difficulties. Indeed, the probability mass function associated with (12) is given by

$$\sum_{(v_1, \ldots, v_d)} \operatorname{sgn}(v_1, \ldots, v_d) C(F_{\mu_1}(v_1), \ldots, F_{\mu_d}(v_d); \rho), \quad (14)$$

where the sum is over all $(v_1, \ldots, v_d) \in \{x_1 - 1, x_1\} \times \ldots \times \{x_d - 1, x_d\}$ and $\operatorname{sgn}(v_1, \ldots, v_d) = 1$ if $v_j = x_j - 1$ is even , and $\operatorname{sgn}(v_1, \ldots, v_d) = -1$ if $v_j = x_j - 1$ is odd [34]. This sum has $2^d$ terms, which can quickly lead to a prohibitive computational cost.

In this context, it is advantageous to consider pairwise likelihood based methods to perform the inference. Instead of having to deal with the model up to $d$ dimensions, which can lead to an intractable likelihood as extensive summations are needed, it suffices to handle bivariate margins alone, which are much easier to obtain and whose probability mass functions require only summations of small dimension. Denote by $F_a(\cdot, \cdot; \theta)$ the bivariate distribution function corresponding to the pair $a = \{i, j\}$; that is, if $(X_{11}, \ldots, X_{1d}) \sim F(\bullet; \theta)$ then $(X_{1i}, X_{1j}) \sim F_a(\cdot, \cdot; \theta)$. Then $F_a(x_i, x_j; \theta) = C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); \rho)$, where

$C_a(u_i, u_j; \rho) := C(1, \ldots, u_i, \ldots, u_j, \ldots, 1; \rho)$ is the bivariate copula corresponding to the pair $a$ (all arguments have been replaced by ones but at the $i$th and $j$th positions). The bivariate density associated with $F_a(\cdot, \cdot; \theta)$ is then given by

$$\begin{aligned} f_a(x_i, x_j, \mu_i, \mu_j; \rho) =& C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); \rho) - C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j - 1); \rho) \\ & - C_a(F_{\mu_i}(x_i - 1), F_{\mu_j}(x_j); \rho) + C_a(F_{\mu_i}(x_i - 1), F_{\mu_j}(x_j - 1); \rho). \end{aligned}$$

Through such an approach, we thus avoid the need to fully specify the likelihood. Similar approaches have been proposed in the past for discrete data [31]. 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 confidence intervals can be obtained in the case where $\pi_n$ is small. Pairwise likelihood methods can estimate all correlation parameters, while at the same time we do not need to implement probability mass functions of dimension higher than 2.

Recall that Assumption 2 is critical to the success of pairwise likelihood methods. We specialize Proposition 1 to the case of copula-based models.

**Proposition 5.** *Suppose that the families* (10) *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* $\{\tilde{C}_a(\cdot, \cdot; \varrho),\ \varrho \in \operatorname{range} w_a\}$ *such that*

*(i) the family* $\{\tilde{C}_a(\cdot, \cdot; \varrho),\ \varrho \in \operatorname{range} w_a\}$ *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 5 are verifiable for at least some classes of models. For models of the form (12) and (13), it all depends on the structure of the correlation matrix. Simple suitable structures are given in Example 1 and Example 2. More complex and suitable structures can be built.

**Example 1.** *Let* $C$ *be the Gaussian copula (13) with correlation matrix*

$$R(\rho) = \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{bmatrix},$$

*for* $\rho \in (-1/(d-1), 1) =: \Theta^{cop}$*. Put* $w_a(\rho) = \rho$ *so that* $\operatorname{range} w_a = (-1/(d-1), 1)$*. 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 (13) with correlation matrix*

$$R(\rho) = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1d} \\ \rho_{21} & 1 & \dots & \rho_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \rho_{d2} & \dots & 1 \end{bmatrix},$$

*for all $\rho = (\rho_{12}, \dots, \rho_{d-1,d}) \in (-1,1)^{d(d-1)/2}$ 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* $\operatorname{range} w_a \subset (-1,1)$. *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* $\operatorname{range} w_a$. *Moreover, $\tilde{C}_a(\cdot,\cdot;\rho_{ij}) = C_a(\cdot,\cdot;\rho_{ij})$ for all $\rho_{ij} \in \operatorname{range} w_a$. The mapping $W$ is one-to-one.*

# 6 Simulations

To investigate the performance of the proposed randomized pairwise likelihood approach for multivariate count data, we performed three independent sets of simulation experiments. In the first two, based on a Gaussian distribution as in [8], we aim to establish the general asymptotic efficiency and coverage of our proposed two-way sampling approach for continuous multivariate data. In the third, we turn our attention to the case of discrete data specifically, using the copula framework described in Section 5, to provide insight into the trade-off between efficiency and computational time for multivariate count data of moderate dimension.

## 6.1 Asymptotic efficiency

First, we simulate a set of $d$-dimensional vectors $Y_i$, $i = 1, \dots, n$ from a symmetric multivariate Gaussian distribution with mean vector $\mu$ and covariance matrix $\Sigma$, where all means $\mu$ are considered to be known and set to 0, and all variances and correlations are fixed to 1 and $\rho$, respectively. In this case, the only parameter to be estimated is thus $\rho$; in different simulation settings, the true value of $\rho$ was set to be equal to one of $\{-0.1, 0, 0.1, 0.2, ..., 0.9\}$. We consider $n = 100, 1000$, and 5000 observations, and the dimension was set to $d = 4$.

To evaluate the efficiency of the randomized pairwise likelihood, we consider sub-sampling parameters of $\pi = 0.5$ and $\pi = 0.2$ as compared to the results from the full maximum likelihood, pairwise likelihood using all pairs of variables and all observations, and the randomized pairwise likelihood for each considered value of $\pi$; simulations were repeated 50,000 times. Efficiency was calculated as the ratio of the variance of parameter estimates across simulated datasets in the pairwise likelihood and randomized pairwise likelihood methods with respect to the full maximum likelihood approach. For all values of $\rho$ considered, all methods considered successfully recover the true value of $\rho$, although as expected,

the variance of estimators increases from the full maximum likelihood to the pairwise likelihood, and further increases in the randomized pairwise likelihood as the sampling parameter $\pi$ decreases (Supplementary Figures S1-S3). In comparing the efficiency of estimators in the pairwise approaches with that of the full maximum likelihood, we remark that the efficiency of the pairwise likelihood is as reported in [8] for $d = 4$; in addition, as expected, the loss of efficiency for the randomized pairwise likelihood is consistent with the theoretical results with respect to the sampling fraction for each value of $\pi$.

## 6.2 Asymptotic coverage

In order to examine the asymptotic properties described, we also performed simulations to evaluate the coverage probabilities for the asymptotic confidence intervals. We still use the compound symmetry model with known means and variances and we estimate the common correlation parameter $\rho$ using randomized pairwise likelihood. Based on Theorem 3 and the derivations of Proposition 3, when $n$ is large, we have that, approximately, $\sqrt{n\pi}(\hat{\rho}-\rho) \sim N(0, V(\hat{\rho}))$, where $\hat{\rho}$ is the randomized pairwise likelihood estimate, $d$ is the dimension and

$$V(\hat{\rho}) = \frac{(1-\hat{\rho}^2)^4}{\frac{d(d-1)}{2}(\hat{\rho}^6 - \hat{\rho}^4 - \hat{\rho}^2 + 1)}.$$

One can create an asymptotically $100(1-\alpha)\%$ confidence interval as

$$\hat{\rho} \pm Z_{1-\alpha/2}\sqrt{\frac{V(\hat{\rho})}{n\pi}}$$

where $Z_a$ is the $a-$quantile of the standard normal distribution.

We simulated 50,000 samples of dimension $d = 4$ for values of $\rho \in \{-0.1, 0.2, \ldots, 0.9\}$, $n \in \{500, 1000, 5000, 10000\}$ and corresponding values of $\pi$ to yield subsample sizes of 100 and 200. For each sample we created the asymptotic confidence interval described above, and we estimated as coverage probability the proportion of times the true value was inside the interval (using $\alpha = 0.05$). The results are depicted in Figure 3. Additional results for fixed values of $\pi \in \{0.01, 0.05, 0.2, 0.5\}$ are shown in Supplementary Figure S4. We can see that, as theoretical results suggest, when the sample size increases, the asymptotic coverage gets closer to the nominal level verifying the potential of the asymptotic results for inference. This also highlights the potential of randomized pairwise likelihood for inference.

## 6.3 Multivariate count data

We next turn our attention to the specific case of multivariate count data. Using the random number generator of an elliptical copula implemented in the `rCopula` function of the `copula` R package [47], we first generate $n$ $d$-dimensional random variates from a Gaussian copula with an exchangeable dispersion matrix parameterized by $\rho$. Subsequently, by pairing these random

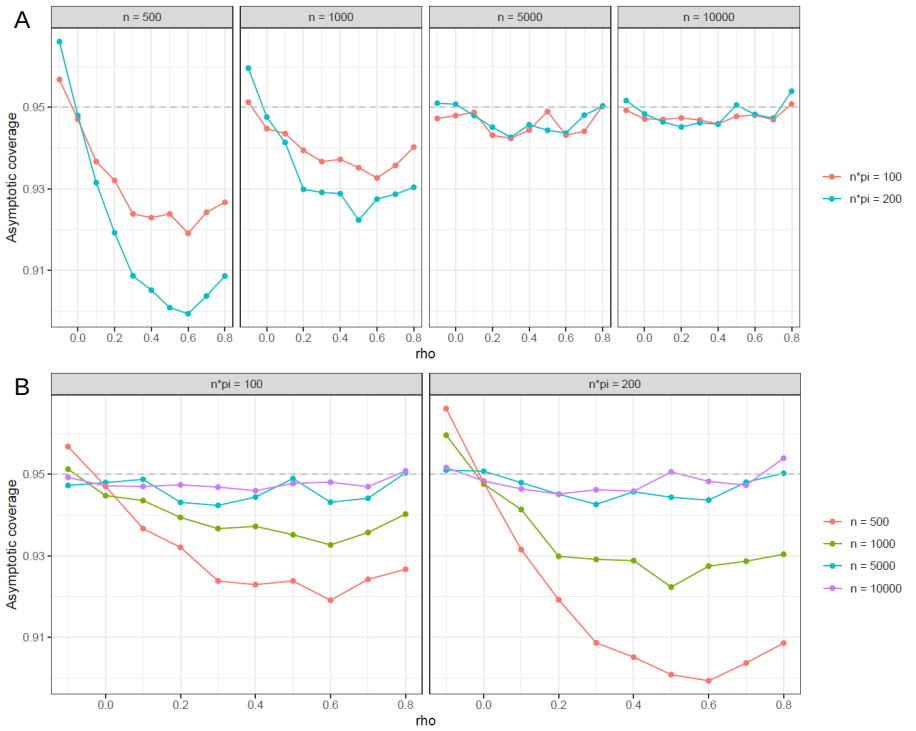


Figure 3: Asymptotic coverage for the compound symmetry example, with $\alpha = 5\%$, averaged over 50,000 replications. The values represent the proportion of times the asymptotic interval contains the true value used to simulate the data. (A) $\rho$ versus asymptotic coverage by sample size $n$; (B) $\rho$ versus asymptotic coverage by subsample size $n \times \pi$.

variates with Poisson marginal distributions parameterized by $\lambda = 1$, we then generate a multivariate Poisson variable for individual $i$ using the Poisson probability quantile function. We consider a variety of different simulation settings: $n \in \{200, 500, 1000, 5000\}$ observations with dimension $d = 4, \ldots, 8$, correlation parameter $\rho \in \{0.25, 0.75\}$, and sampling parameter $\pi \in \{0.1, 0.3, 0.5, 0.7, 1\}$; for each combination of parameters, simulations were repeated 300 times. Note that the setting where $\pi = 1$ corresponds to the classical pairwise likelihood approach; in all other cases, pairs of variables and observations are subsampled according to Bernoulli probability $\pi$. Finally, unlike the previous simulation experiment, here we consider that both the $d$-dimensional mean vector $\lambda$ and the $(d \times (d-1))$-dimensional vector $\rho$ (corresponding to the off-diagonal elements of the copula dispersion matrix, which is assumed to be unstructured) are unknown and must be estimated.

Note that for the Gaussian copula the joint marginal distributions are char-

acterized by the same copula and hence the pairwise likelihood approach uses the correct bivariate marginals. This in turn implies that we estimate the parameters used to generate the data. To apply the randomized pairwise likelihood estimation procedure, we first initialize parameter values for $\lambda$ and $\rho$ using the marginal means of each variable and the Pearson correlation of each pair of variables, respectively. Finally, we maximize the randomized pairwise likelihood using the optimization algorithm of [4] (`"L-BFGS-B"` method in the general-purpose optimization R function `optim`); the maximum number of iterations for the optimization algorithm is capped at 30,000.

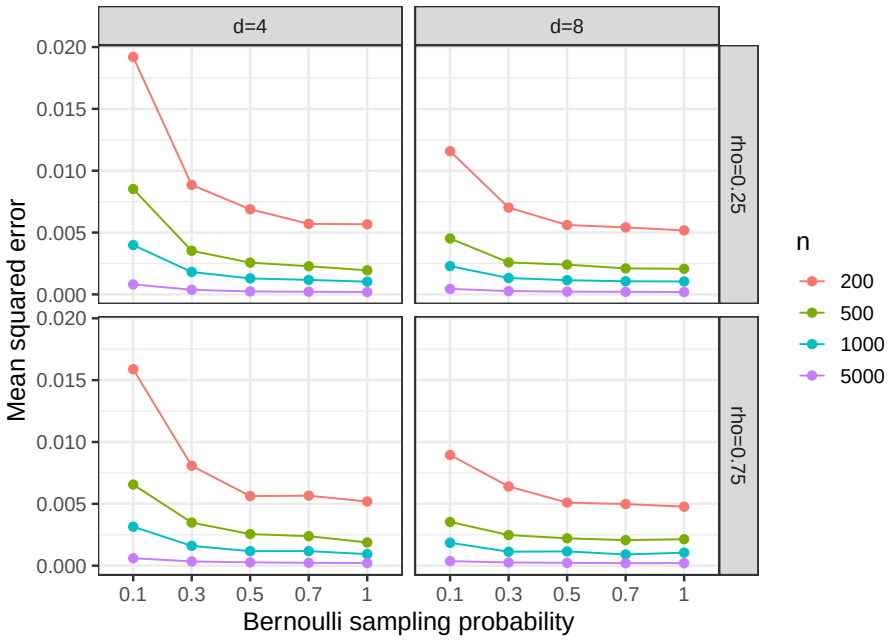


Figure 4: Average mean squared error (across 300 replications) using the randomized pairwise likelihood approach with $\pi = \{0.1, 0.3, 0.5, 0.7, 1\}$ for simulated data with dimension $d = 4$ or 8, copula dispersion $\rho = 0.25$ or 0.75, and number of observations ranging from $n = 200$ to 5000.

We first investigate the trade-off between efficiency and computational time, with respect to the chosen value for the Bernoulli sampling probability $\pi$, for the randomized pairwise likelihood approach. In Figure 4, for cases of moderately ($\rho = 0.25$) and highly structured data ($\rho = 0.75$) in data of dimension $d = 4$ or 8, we present the mean squared error (MSE) of parameter estimates for the randomized pairwise likelihood with varying Bernoulli sampling rates. The results for dimensions $d = 5, 6, 7$ are shown in Figure S5. We remark that the randomized pairwise likelihood yields similar MSEs as compared to the full maximum likelihood ($\pi = 1$) when a sufficiently large Bernoulli sampling rate

is used (e.g., $\pi = 0.5$); in addition, this appears to hold true for varying sample sizes ($n = 200$ to 5000). Perhaps unsurprisingly, differences in efficiency for the randomized pairwise likelihood compared to the full maximum likelihood tend to be smaller for highly structured data ($\rho = 0.75$), as the subsampling of observations and variable pairs removes largely redundant information without negatively impacting parameter estimation. In addition, the efficiency of parameter estimation does not appear to suffer with even smaller rates ($\pi = 0.1$); particularly when data have a sufficiently large number of observations (e.g., $n \geq 500$); this is notable, as the efficiency is not unduly impacted despite a significant decrease in the amount of data used for estimation.

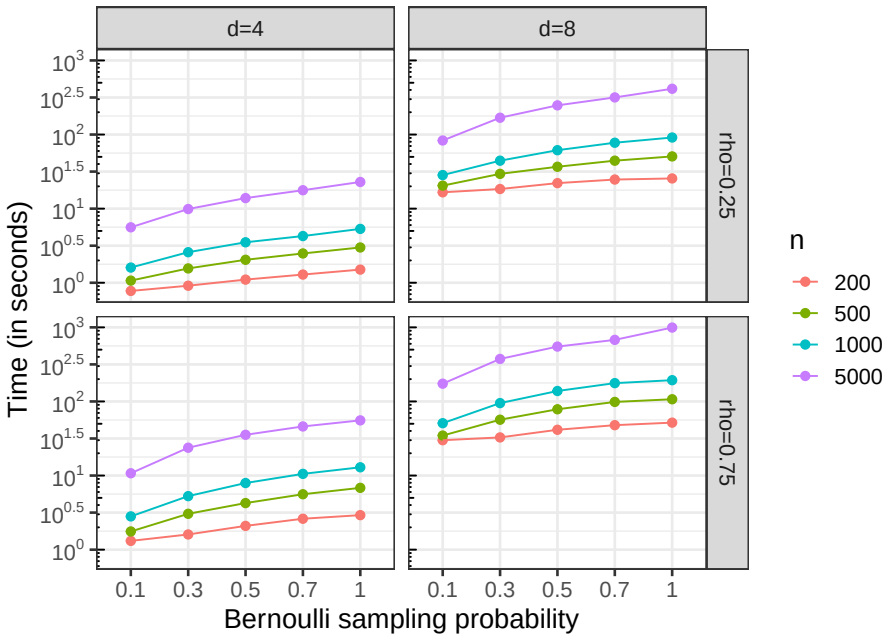


Figure 5: Average computational time in seconds (across 300 replications) using the randomized pairwise likelihood approach with $\pi = \{0.1, 0.3, 0.5, 0.7, 1\}$ for simulated data with dimension $d = 4$ or 8, copula dispersion $\rho = 0.25$ or 0.75, and number of observations ranging from $n = 200$ to 5000.

The reduction in data used for estimation due to the pairwise sampling scheme brings with it considerable gains in computational time (Figure 5), particularly for larger sample sizes; for example, in the case of $n = 5000$ observations and $\rho = 0.75$, a fivefold gain in average computational time can be achieved for all dimensions considered ($d = 4, \ldots, 8$) by using the randomized pairwise likelihood with $\pi = 0.1$ instead of the full maximum likelihood.

# 7 Application on microbiome and transcriptome data

In the following section, we illustrate the application of our proposed randomized pairwise likelihood procedure for multivariate count data on two biological applications, focusing in particular on the gain in computational time achieved through the proposed two-way sampling strategy.

## 7.1 Fungal interaction networks in oak powdery mildew

Microorganisms are known to form a variety of complex ecological relationships with one another, including phenomena such as mutualism, parasitism, and competition, and the inference of these interaction networks from microbial abundance data is a primary question of interest [11]. In a recent study, [17] sought to identify microbial interactions between the causal agent of a fungus called oak powdery mildew, *Erysiphe alphitoides*, and other foliar microorganisms of the pedunculate oak (*Quercus robur* L.) In their experiment, DNA was extracted from a total of approximately 40 leaves collected from each of three trees, and 454 pyrosequencing was performed for fungal and bacterial assemblages using marker-gene (ITS1 and 16S, respectively) based metabarcoding. A variety of other covariates were collected for each sample, including leaf position and infection level of the leaf. Full details of the experimental design may be found in [17]. Following pre-processing, the data consist of counts of microbial species (operational taxonomic units; OTU) in $n = 116$ samples for a total of 114 species (*E. alphitoides*, 47 fungal OTUs, and 66 bacterial OTUs).

As in the original study, our goal is to identify the interactions between *E. alphitoides* and the other microbial species; in this work, we focus our attention on the 13 fungal OTUs identified as having putative interactions with *E. alphitoides* by [17], and we use the randomized pairwise likelihood with an unstructured Gaussian copula paired with Poisson marginals. To account for the recorded environmental covariates, we first fit a marginal Poisson generalized linear model to each OTU abundance using leaf distance to base, leaf distance to trunk, leaf distance to ground, and orientation as predictors. The marginal expected values (on the response scale) from the GLM were subsequently plugged in for the $d$-dimensional mean vector $\lambda$. Initial values for the $d \times (d-1)$-dimensional dispersion vector $\rho$, corresponding to the off-diagonal elements of the unstructured copula dispersion matrix, were set to be equal to the pairwise Pearson correlations among OTUs, and the randomized pairwise likelihood was maximized using the `"L-BFGS-B"` method constrained by $[-1, 1]$.

In Table 1, we report the estimated parameters of the unstructured Gaussian copula correlation matrix between each considered fungal OTU and *E. alphitoides* using the pairwise likelihood and randomized pairwise likelihood strategies. Estimated values for these parameters are similar whether the randomized pairwise likelihood for all observations and variable pairs ($\pi = 1$) or the randomized pairwise likelihood ($\pi = 1/\log n$) are used, but computational time is nearly

Table 1: Fungal OTUs and their estimated interaction with *E. alphitoides* by pairing Poisson marginals with an unstructured Gaussian copula and using the pairwise likelihood (PL; $\pi = 1$) or randomized pairwise likelihood (RPL; $\pi = 1/\log n$).

| OTU | Putative species/genus | PL | RPL |
|---|---|---|---|
| 1 | *Naevala minutissima* | -0.22 | -0.19 |
| 2 | *Mycosphaerella punctiformis* | -0.19 | -0.16 |
| 9 | *Cladosporium cladosporioides* | 0.00 | 0.01 |
| 10 | — | -0.06 | -0.04 |
| 15 | *Monochaetia kansensis* | -0.05 | -0.03 |
| 19 | — | -0.13 | -0.11 |
| 20 | *Lalaria inositophila* | -0.03 | -0.01 |
| 23 | *Sporobolomyces roseus* | 0.29 | 0.27 |
| 25 | — | -0.02 | 0.01 |
| 26 | *Taphrina carpini* | 0.56 | 0.57 |
| 28 | *Sporobolomyces gracilis* | 0.12 | 0.14 |
| 1278 | *Mycosphaerella punctiformis* | -0.20 | -0.21 |
| 1567 | — | -0.04 | -0.06 |
| Computational run time (minutes) | | 9.68 | 5.41 |

halved by using randomized sampling. The signs and values of the estimated correlations are largely in agreement with the interactions reported in [17], with the exception of OTUs 28, 1567, and 20. In particular, OTU 26 (*Taphrina carpin*), which was identified as having the largest positive interaction effect with *E. alphitoides* by [17] using a Bayesian network inference approach, similarly has the largest estimated interaction using pairwise likelihoods. Finally, as in the original study, we note that a majority of the pairwise interactions among fungal OTUs appear to be positive for this community, suggesting the strong role of mutualism and commensalism in oak leaves.

We remark that the inference of microbial networks is currently an active area of research [see, for example, 6, 27, 7], with many proposed approaches focusing on the identification of direct versus indirect associations, simultaneous estimation of covariate effects and interactions, the presence of overdispersed and zero-inflated counts, and the compositional nature of the microbial abundance data. Although it is beyond the scope of this work to extensively evaluate these approaches, our results suggest the potential benefit of incorporating a pairwise sampling scheme into these approaches.

## 7.2 Global transcriptome correlations across life cycles in a honeybee parasite

The parasitic mite *Varroa destructor* is widely considered to represent a significant threat to the western honeybee *Apis mellifera*, but progress in developing solutions to control it have been slowed by a lack of knowledge of its biology. To

address this gap, [28] generated a transcriptomic catalogue over ten points in the full life cycle in *Varroa* mites, including seven stages in reproducing females (young, phoretic, arresting, pre-laying, laying, post-laying, and emerging mites), non-reproducing female mites, males, and artificially reared phoretic mites, with a total of 4 replicated pools of 10 mites for each group. Using high-throughput sequencing of total RNAs, counts of mapped sequencing reads were obtained for 41,801 contigs; additional details about the experimental design and data pre-processing may be found in [28].

Our goal here is to perform an exploratory analysis to evaluate the overall transcriptome-wide correlation among different life stages in *Varroa*. We focus on the gene expression data obtained from a single colony (R204) across the 10 different life cycle groups, yielding count data for $n$=31,267 contigs and $d$=10 life cycle groups. For RNA-seq data, counts of expression in a given sample are strongly associated with the sequencing effort of each sample (known as the library size) and the length of the gene; larger library sizes and longer genes tend to have higher counts. To adjust for these two biases, for each life cycle group we first fit a marginal Poisson generalized linear model with a per-contig offset term corresponding to the log of the normalized (using the Trimmed Mean of M values approach; [38]) total expression. As for the microbiome example, GLM marginal expected values on the response scale were plugged in for the mean vector $\lambda$, the dispersion vector $\rho$ was initialized using pairwise Pearson correlations among life cycle groups, and `"L-BFGS-B"` was used to optimize the randomized pairwise likelihood.

Similarly to the microbiome example above, the use of the randomized pairwise likelihood (with $\pi = 1/\log n$) in the place of the pairwise likelihood represented a significant gain in computational time, respectively corresponding to 37.1 minutes as compared to 96.7 minutes. In this example, however, a more modest agreement was found between the pairwise likelihood and randomized pairwise likelihood strategies (Spearman correlation $\rho = 0.21$), perhaps due to the wider range of counts observed in the RNA-seq data (median = 94, maximum = 11,178,256) compared to the microbiome data (median = 14, maximum = 2027). Figure 6 provides a visualization of the pairwise dispersion between groups estimated by the randomized pairwise likelihood. Similar to the results of [28], we note a distinction between reproductive (young, arresting, pre-laying, laying, phoretic) and post- or non-reproductive stages (emerging, post-lay, non-reproductive female); interestingly, male and lab-reared *Varroas* appear to cluster with the latter group as well. In practice, exploratory analyses of transcriptome sequencing data typically rely on the use of variance stabilizing transformations to facilitate the application of methods such as principal components analysis; in this illustration, we have instead explicitly modelled the count nature of these expression data via a Poisson distribution and used Gaussian copulas to model the dependency structure among groups.

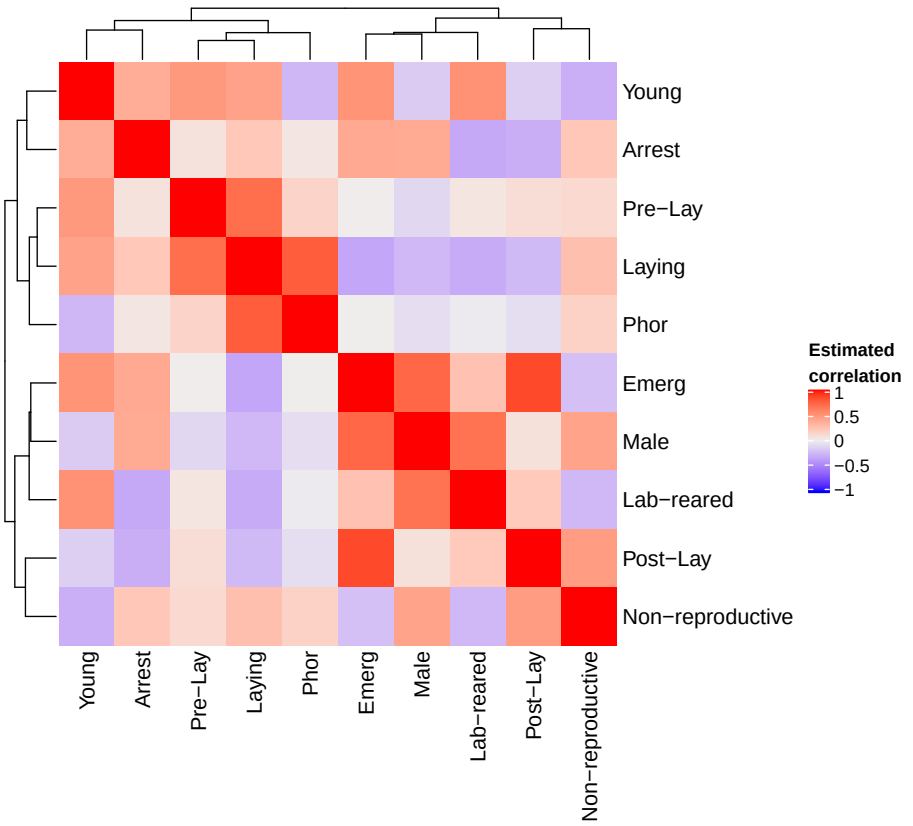


Figure 6: Clustered heatmap of the estimated pairwise correlation parameters, using the randomized pairwise likelihood ($\pi = 1/\log n$) between *Varroa* life cycle groups for the [28] transcriptome data.

# 8 Conclusions

In this work, we considered a randomized pairwise likelihood to reduce the computational burden associated with the standard pairwise likelihood. We showed that when $\pi_n \to 0$, the two-way random sampling mechanism permits the computation of approximate confidence intervals of computational complexity of order $d^2$, where only bivariate marginals are involved, to be compared to the original problem of computational complexity of order $d^4$, where up to four-dimensional marginals were involved. Moreover, the number of summands that comprise the randomized pairwise likelihood has been lowered by a factor $\pi_n$.

The proposed method is applicable in general but we had a particular focus on copula-based models for count data, where inference is particularly challenging and remains largely an open problem once the number of variables is more than a few. We believe that the proposed method opens the door to design-

ing affordable inference procedures in these model,s and hence facilitating their use. In the two data applications we presented, we used a two-step approach to first estimate the marginal parameters and then the correlation parameters, but it is also possible to simultaneously estimate the dependence parameters and marginal parameters. Note that the randomized pairwise likelihood method can also benefit other types of models, such as latent variable models, especially as alternatives to variational methods for which the asymptotic properties of the estimators remain unknown in general.

In the future, other sampling schemes could be implemented to exploit information of the data 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. We also wonder whether the maximum randomized pairwise likelihood estimator could serve as a starting point to the procedure of [14]. The rate of the maximum randomized pairwise likelihood estimator is not $\sqrt{n}$ as soon as $\pi_n \to 0$, but it can be close. This raises the theoretical question of whether it is possible to get asymptotic results when $d \to \infty$. This makes sense in the high dimensional context, but the general problem appears to be very challenging. Finally, although not discussed in detail here, the maximization of the randomized pairwise likelihood may be made easier by considering other estimation strategies. For instance, maximization by parts approaches, which split the full maximization problem into smaller ones, can likely lead to further considerable gains in computational time.


## Acknowledgements

We thank Mahendra Mariadassou for detailed comments on a first version of this manuscript and for providing the microbiome data from [17].


## A Proofs of the theorems

In the proofs, it will be convenient to consider the bivariate functions $f_a(X_i^{(a)};\theta)$ as functions taking as an argument the whole vector $X_i$ so that $f_a(X_i^{(a)};\theta)$ will be denoted by $f_a(X_i;\theta)$. To take advantage of empirical process techniques, we shall build empirical processes related to our problem.

Let $\mathcal{G}_a$, $a = 1, 2, \ldots, A$, be classes of functions $g_a : \mathbf{R}^d \to \mathbf{R}^L$ satisfying $\mathrm{E}\, g_a(X_1)^2 < \infty$ componentwise. Let $\mathcal{M}(\mathcal{G}_1, \ldots, \mathcal{G}_A)$ be the set of functions $m$ of the form $m(x, w) = \sum_{a=1}^{A} w_a g_a(x)$, $x \in \mathbf{R}^d$, $w = (w_1, \ldots, w_A) \in [0, \infty)^A$, $g_a \in \mathcal{G}_a$, $a = 1, \ldots, A$. Let $X_i$, $i = 1, \ldots, n$, be i.i.d. random vectors in $\mathbf{R}^d$ with law $P$. For each $n$, let $W_{ni}^{(a)}$, $i = 1, \ldots, n$, $a = 1, \ldots, A$, be i.i.d. Bernoulli random variables with parameter $0 < \pi_n \leq 1$. For each $n$, $X_1, \ldots, X_n$ and $W_{n1}^{(1)}, W_{n1}^{(2)}, \ldots, W_{nn}^{(A)}$ are independent. For $i = 1, \ldots, n$, let $W_{ni}$ be the vector with components $W_{ni}^{(a)}$, $a = 1, \ldots, A$. For a probability measure $P$ and a function $f$, $Pf$ denotes $\int f \, \mathrm{d}P$. Let $P_{nn}$ be the average of Dirac measures at

the points $(X_i, W_{ni}/\pi_n)$, $i = 1, \ldots, n$; thus if $m \in \mathcal{M}(\mathcal{G}_1, \ldots, \mathcal{G}_A)$ then

$$P_{nn}m = \int m \, \mathrm{d}P_{nn} = \frac{1}{n} \sum_{i=1}^{n} m\left(X_i, \frac{W_{ni}}{\pi_n}\right) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{A} \frac{W_{ni}^{(a)}}{\pi_n} g_a(X_i).$$

Let $P_n^*$ be the probability distribution of $(X_1, W_{n1}/\pi_n)$; thus

$$P_n^* m = \mathrm{E}\ m\left(X_1, \frac{W_{n1}}{\pi_n}\right) = \sum_{a=1}^{A} \mathrm{E}\, \frac{W_{n1}^{(a)}}{\pi_n} g_a(X_1) = \sum_{a=1}^{A} \mathrm{E}\, g_a(X_1) = Pm(\cdot, 1).$$

Notice that it does not depend on $n$. Denote by $G_{nn}^*$ the signed measure $\sqrt{n\pi_n}(P_{nn} - P_n^*)$. We shall use the concept of a bracketing number [39, 41, 36]. If $\mathcal{G}$ is a class of real-valued functions on some Euclidean space equipped with a probability measure $P$ and $\delta$ is a positive real number, then the bracketing number of $\mathcal{G}$, denoted by $N(\delta, \mathcal{G}, P)$, is the smallest number $N$ of brackets $[g_j^L, g_j^U]$, $j = 1, \ldots, N$, such that (i) $Pg_j^{\mathrm{U}} - Pg_j^{\mathrm{L}} \leq \delta$, $j = 1, \ldots, N$, and (ii) for all $g$ in $\mathcal{G}$, there is $j \in \{1, \ldots, N\}$ such that $g_j^{\mathrm{L}} \leq g \leq g_j^{\mathrm{U}}$. Recall that two asymptotic frameworks are considered: $\pi_n = \pi$ is constant and $\pi_n \to 0$ as $n \to \infty$.

The following lemmas establish a uniform law of large numbers and a central limit theorem expressed in terms of the new empirical processes. These results are the building blocks on the top of which the proofs of the theorems rest. Measurability issues are ignored. See [41, 40] for a way of addressing this.

**Lemma A.1.** *Let $m \in \mathcal{M}(\mathcal{G}_1, \ldots, \mathcal{G}_A)$ with $L = 1$. If $\pi_n > 0$ is constant or if $\pi_n \to 0$ such that $n\pi_n \to \infty$ then $|P_{nn}m - P_n^*m| \xrightarrow{P} 0$ as $n \to \infty$.*

**Lemma A.2.** *Let $m \in \mathcal{M}(\mathcal{G}_1, \ldots, \mathcal{G}_A)$ with $L = 1$. Assume furthermore that $N(\delta, \mathcal{G}_a, P) < \infty$ for all $\delta > 0$ and all $a = 1, \ldots, A$. If $\pi_n > 0$ is constant or if $\pi_n \to 0$ such that $n\pi_n \to \infty$ then*

$$\sup_{m \in \mathcal{M}(\mathcal{G}_1, \ldots, \mathcal{G}_A)} |P_{nn}m - P_n^*m| \xrightarrow{P} 0, \quad n \to \infty.$$

**Lemma A.3.** *Let $m \in \mathcal{M}(\mathcal{G}_1, \ldots, \mathcal{G}_A)$. If $\pi_n = \pi$ is constant then $G_{nn}^*m$ converges in distribution to a centered Gaussian vector with variance-covariance matrix*

$$(1 - \pi) \left( \sum_{a=1}^{A} \mathrm{E}\, g_a(X_1) g_a(X_1)^\top \right) + \pi \left( \sum_{a=1}^{A} \sum_{b=1}^{A} \mathrm{E}\, g_a(X_1) g_b(X_1)^\top - \mathrm{E}\, g_a(X_1)\, \mathrm{E}\, g_b(X_1)^\top \right).$$

*If $\pi_n \to 0$ such that*

$$\mathrm{E}\, g_{al}(X_1)^4 \exp\left( -\frac{n\pi_n \kappa}{\sum_{a=1}^{A} g_{al'}(X_1)^2} \right) = o(\pi_n) \tag{15}$$

*for all $\kappa > 0$ and all $l, l' = 1, \ldots, L$, then $G_{nn}^* m$ converges in distribution to a centered Gaussian random vector with variance-covariance matrix*

$$\sum_{a=1}^{A} \mathrm{E}\, g_a(X_1) g_a(X_1)^\top. \tag{16}$$

## Proof of Theorem 1

One can follow almost word for word the proofs of Theorem 2 and Theorem 3. The appropriate changes are easily made: it suffices to switch to the appropriate asymptotic frameworks in Lemma A.2 and Lemma A.3.

## Proof of Theorem 2

Since $\hat{\theta}_n^{\mathrm{MRPL}}$ is a MRPLE, there is a compact subset $\Lambda \subset \Theta$ that contains $\theta_0$ such that $L_n^{\mathrm{RPL}}(\hat{\theta}_n^{\mathrm{MRPL}}) \geq L_n^{\mathrm{RPL}}(\theta)$ for all $\theta \in \Lambda$. Denote $L^{\mathrm{PL}}(\theta) = \sum_a L_a(\theta)$, $\theta \in \Theta$. Then $L^{\mathrm{PL}}$ is uniquely maximized at $\theta_0 \in \Lambda$ and $\mathrm{E}\, L_n^{\mathrm{RPL}}(\theta) = L^{\mathrm{PL}}(\theta)$, $\theta \in \Theta$. Since $\theta_0 \in \Lambda$, certainly

$$L_n^{\mathrm{RPL}}(\hat{\theta}_n^{\mathrm{MRPL}}) \geq \sup_{\theta \in \Lambda} L_n^{\mathrm{RPL}}(\theta) \geq L_n^{\mathrm{RPL}}(\theta_0).$$

Theorem 5.7 in [40] asserts that if the conditions

(i) $\forall \varepsilon > 0, \quad \sup_{\theta \in \Lambda : |\theta - \theta_0| \geq \varepsilon} L^{\mathrm{PL}}(\theta) < L^{\mathrm{PL}}(\theta_0)$

(ii) $\sup_{\theta \in \Lambda} |L_n^{\mathrm{RPL}}(\theta) - L^{\mathrm{PL}}(\theta)| \xrightarrow{P} 0$

hold, then $\hat{\theta}^{\mathrm{MRPL}} \xrightarrow{P} \theta_0$ as $n \to \infty$.

Let us check (i). Since $f(\cdot, \theta_0)$ belongs to $L_2(\mathbf{R}^d)$, it follows that $L^{\mathrm{PL}}(\theta_0) < \infty$. By Assumption 1, the function $L^{\mathrm{PL}} : \Lambda \longrightarrow [-\infty, \infty)$ is continuous on $\Lambda$. Since the set $\{\theta \in \Lambda : |\theta - \theta_0| \geq \varepsilon\}$ is compact, the supremum of $L^{\mathrm{PL}}$ is reached. But this supremum must be less than $L^{\mathrm{PL}}(\theta_0)$, because, by Assumption 2, the point $\theta_0$ is the unique maximizer. Condition (i) is fulfilled.

Let us check (ii). Using the notation introduced at the beginning of this section, we can write

$$\begin{aligned}
&\sup_{\theta \in \Lambda} |L_n^{\mathrm{RPL}}(\theta) - L^{\mathrm{PL}}(\theta)| \\
&= \sup_{\theta \in \Lambda} \left| \sum_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{W_{ni}^{(a)}}{\pi_n} \log f_a(X_i; \theta) - \mathrm{E} \log f_a(X_1; \theta) \right) \right| \\
&\leq \sup_{m \in \mathcal{M}(\mathcal{G}_a, a \in \mathcal{A})} |P_{nn} m - P_n^* m|,
\end{aligned}$$

where $\mathcal{G}_a = \{\log f_a(\cdot; \theta),\, \theta \in \Lambda\}$, $a \in \mathcal{A}$. By Lemma A.2, the condition (ii) will hold if we can show that the bracketing numbers $N(\delta, \mathcal{G}_a, P)$, $\delta > 0$, are finite.

But it is well known that classes indexed by a compact subset of an Euclidean space have finite bracketing numbers; see for instance Lemma 3.10 in [39] for a proof. Hence condition (ii) is fulfilled as well.

## Proof of Theorem 3

Recall the notation introduced at the beginning of this section and let $m(x, w, \theta) = \sum_{a \in \mathcal{A}} w_a \ell_a(x; \theta)$. As in the proof of Theorem 2 let $L^{\mathrm{PL}}(\theta) = \sum_a L_a(\theta)$. Denote the gradient of $m$ with respect to $\theta$ by $\nabla m$. Denote the Hessian matrix of $L^{\mathrm{PL}}$ at $\theta_0$ by $\nabla^2 L^{\mathrm{PL}}(\theta_0)$. If we can show

$$\sqrt{n\pi_n}(\hat{\theta}^{\mathrm{MRPL}} - \theta_0) = -\left[\nabla^2 L^{\mathrm{PL}}(\theta_0)\right]^{-1} G^*_{nn} \nabla m(\cdot, \cdot, \theta_0) + o_P(1), \tag{17}$$

then Lemma A.3 will imply that $\sqrt{n\pi_n}(\hat{\theta}^{\mathrm{MRPL}} - \theta_0)$ converges in distribution to a centered Gaussian random vector with variance-covariance matrix

$$\left[\nabla^2 L^{\mathrm{PL}}(\theta_0)\right]^{-1} \left[ (1-\pi) \sum_a \mathrm{E}\, \dot{\ell}_a \dot{\ell}_a^\top + \pi \left( \sum_{a,b} \mathrm{E}\, \dot{\ell}_a \dot{\ell}_b^\top - \mathrm{E}\, \dot{\ell}_a\, \mathrm{E}\, \dot{\ell}_b^\top \right) \right] \left[\nabla^2 L^{\mathrm{PL}}(\theta_0)\right]^{-1},$$

if $\pi_n$ is a constant, and $\sum_a \mathrm{E}\, \dot{\ell}_a \dot{\ell}_a^\top$ if $\pi_n \to 0$. The asymptotic variance-covariance matrices above are those announced by Theorem 1 and Theorem 3, respectively, because Assumption 1 implies $\mathrm{E}\, \dot{\ell}_a = 0$ and $\nabla^2 L^{\mathrm{PL}}(\theta_0) = -\sum_a \mathrm{E}\, \dot{\ell}_a \dot{\ell}_a^\top$.

So we need to show (17). The map $L^{\mathrm{PL}}$ is two times continuously differentiable at $\theta_0$ with gradient $\nabla L^{\mathrm{PL}}(\theta_0) = P \nabla m(\cdot, 1, \theta_0)$ and negative definite Hessian matrix $\nabla^2 L^{\mathrm{PL}}(\theta_0) = P \nabla^2 m(\cdot, 1, \theta_0)$. Let $\mathring{\Lambda}$ be the interior of $\Lambda$, that is, its biggest open subset. For every $n$,

$$L_n^{\mathrm{RPL}}(\hat{\theta}_n^{\mathrm{MRPL}}) \geq \sup_{\theta \in \mathring{\Lambda}} L_n^{\mathrm{RPL}}(\theta)$$

and $\hat{\theta}_n^{\mathrm{MRPL}}$ is consistent for $\theta_0$ by Theorem 2. Thereore equation (17) follows from Theorem 3.2.16 of [41, p. 300], which itself is a generalization of an idea of [36, 37], provided that

$$\begin{aligned} \sqrt{n\pi_n} \left( \left[ L_n^{\mathrm{RPL}}(\theta_0 + \tilde{h}_n) - L^{\mathrm{PL}}(\theta_0 + \tilde{h}_n) \right] - \left[ L_n^{\mathrm{RPL}}(\theta_0) - L^{\mathrm{PL}}(\theta_0) \right] \right) \\ = \tilde{h}_n^\top G^*_{nn} \nabla m(\cdot, \cdot, \theta_0) + o_P\left( \|\tilde{h}_n\| + \sqrt{n\pi_n} \|\tilde{h}_n\|^2 + \frac{1}{\sqrt{n\pi_n}} \right), \end{aligned}$$

for all random sequences $\tilde{h}_n = o_P(1)$. Denoting

$$\nabla_{i_1} m(\cdot, \cdot, \theta) = \frac{\partial m(\cdot, \cdot, \theta)}{\partial \theta_{i_1}}, \quad \nabla^2_{i_1 i_2} m(\cdot, \cdot, \theta) = \frac{\partial^2 m(\cdot, \cdot, \theta)}{\partial \theta_{i_1} \partial \theta_{i_2}}, \quad \text{etc,}$$

and using the notation introduced at the beginning of this section, one can see

that this condition boils down to

$$\frac{1}{2}\sum_{i_1,i_2}\tilde{h}_{i_1}\tilde{h}_{i_2}G^*_{nn}\nabla^2_{i_1i_2}m(\cdot,\cdot,\theta_0)+\frac{1}{6}\sum_{i_1,i_2,i_3}\tilde{h}_{i_1}\tilde{h}_{i_2}\tilde{h}_{i_3}G^*_{nn}\nabla^3_{i_1i_2i_3}m(\cdot,\cdot,\hat{h})$$
$$=o_P\left(\|\tilde{h}\|+\sqrt{n\pi_n}\|\tilde{h}\|^2+\frac{1}{\sqrt{n\pi_n}}\right), \quad (18)$$

where $\hat{h}$ is a point between $\theta_0$ and $\theta_0+\tilde{h}$. Above we have dropped the subscripts $n$ of $\tilde{h}$ and $\hat{h}$. In view of Assumption 1 and (6), Lemma A.3 implies $G^*_{nn}\nabla^2_{i_1i_2}m(\cdot,\cdot,\theta_0)=O_P(1)$ whether $\pi_n$ is a constant or $\pi_n\to 0$. Remember that the third derivatives are bounded by the functions $\Psi_a$, put $\Psi(x,w):=\sum_{a\in\mathcal{A}}w_a\Psi_a(x)$ so that $|\nabla^3_{i_1i_2i_3}m(x,w,\hat{h})|\le\Psi(x,w)$, which entails

$$|G^*_{nn}\nabla^3_{i_1i_2i_3}m(\cdot,\cdot,\hat{h})|\le G^*_{nn}\Psi+2\sqrt{n\pi_n}P\Psi(\cdot,1)=O_P(\sqrt{n\pi_n}),$$

because $G^*_{nn}\Psi=O_P(1)$ by Lemma A.3. Thus, in both cases $\pi_n\to 0$ and $\pi_n$ constant, the left hand side in (18) is $O_P\left(\|\tilde{h}\|^2\left(1+\|\tilde{h}\|\sqrt{n\pi_n}\right)\right)$. The proof is complete.

# B Proofs of the Lemmas in Section A

## Proof of Lemma A.1

We have

$$|P_{nn}m-P^*_nm|=\left|\frac{1}{n}\sum_{i=1}^n\sum_{a=1}^A\left(\frac{W^{(a)}_{ni}}{\pi_n}g_a(X_i)-\mathrm{E}\,g_a(X_1)\right)\right|.$$

Let $\epsilon>0$. Since $\sum_a(\pi_n^{-1}W^{(a)}_{ni}g_a(X_i)-\mathrm{E}\,g_a(X_1))$, $i=1,\dots,n$, are i.i.d., Chebychev's inequality yields

$$P\left(\left|\frac{1}{n}\sum_i\sum_a\left(\frac{W^{(a)}_{ni}}{\pi_n}g_a(X_i)-\mathrm{E}\,g_a(X_1)\right)\right|>\epsilon\right)$$
$$\le\frac{\mathrm{Var}\sum_a\left(W^{(a)}_{n1}g_a(X_1)/\pi_n-\mathrm{E}\,g_a(X_1)\right)}{n\epsilon^2}$$
$$=\frac{(1-\pi_n)\sum_a\mathrm{E}\,g_a(X_1)^2}{n\pi_n\epsilon^2}+\frac{\mathrm{E}\left(\sum_a g_a(X_1)-\mathrm{E}\,g_a(X_1)\right)^2}{n\epsilon^2}\to 0$$

whether $\pi_n$ is constant or $\pi_n\to 0$ because $n\pi_n\to\infty$ either way.

## Proof of Lemma A.2

We have

$$\sup_{m\in\mathcal{M}(\mathcal{G}_1,\dots,\mathcal{G}_A)}|P_{nn}m-P^*_nm|=\sup_{g_1\in\mathcal{G}_1,\dots,g_A\in\mathcal{G}_A}\left|\frac{1}{n}\sum_{i=1}^n\sum_{a=1}^A\frac{W^{(a)}_{ni}}{\pi_n}g_a(X_i)-\mathrm{E}\,g_a(X_1)\right|.$$

Let $\delta > 0$. Denote $N_a = N(\delta, \mathcal{G}_a, P)$. For every $a = 1, \ldots, A$, there are brackets $[g_{a,j}^{\mathrm{L}}, g_{a,j}^{\mathrm{U}}]$, $j = 1, \ldots, N_a$, such that (i) $\int g_{a,j}^{\mathrm{U}} - g_{a,j}^{\mathrm{L}} \, \mathrm{d}P < \delta$ for all $j \in \{1, \ldots, N_a\}$ and (ii) for every $g_a \in \mathcal{G}_a$, there is $j(a) \in \{1, \ldots, N_a\}$ such that $g_{a,j(a)}^{\mathrm{L}} \leq g_a \leq g_{a,j(a)}^{\mathrm{U}}$. This implies

$$
\begin{aligned}
&- A\delta + \frac{1}{n} \sum_{i,a} \left( \frac{W_{ni}^{(a)}}{\pi_n} g_{a,j(a)}^{\mathrm{L}}(X_i) - \mathrm{E}\, g_{a,j(a)}^{\mathrm{L}}(X_1) \right) \\
&\leq \frac{1}{n} \sum_{i,a} \left( \frac{W_{ni}^{(a)}}{\pi_n} g_a(X_i) - \mathrm{E}\, g_a(X_1) \right) \\
&\leq \frac{1}{n} \sum_{i,a} \left( \frac{W_{ni}^{(a)}}{\pi_n} g_{a,j(a)}^{\mathrm{U}}(X_i) - \mathrm{E}\, g_{a,j(a)}^{\mathrm{U}}(X_1) \right) + A\delta
\end{aligned}
$$

and hence

$$
\begin{aligned}
&\sup_{g_1 \in \mathcal{G}_1, \ldots, g_A \in \mathcal{G}_A} \left| \frac{1}{n} \sum_{i,a} \left( \frac{W_{ni}^{(a)}}{\pi_n} g_a(X_i) - \mathrm{E}\, g_a(X_1) \right) \right| \\
&\leq \max \left\{ \left| \frac{1}{n} \sum_{i,a} \left( \frac{W_{ni}^{(a)}}{\pi_n} g_{a,j(a)}^{\mathrm{S}}(X_i) - \mathrm{E}\, g_{a,j(a)}^{\mathrm{S}}(X_1) \right) \right|, \mathrm{S} \in \{\mathrm{L}, \mathrm{R}\} \right\} + A\delta.
\end{aligned}
$$

Regardless of the behavior of the sequence $\pi_n$, the first term in the right-hand side goes to zero in probability by Lemma A.1. Since $\delta$ was arbitrary, the proof is complete.

## Proof of Lemma A.3

*Case $\pi_n = \pi$ constant.* We have

$$
G_{nn}^* m = \frac{\sqrt{\pi}}{\sqrt{n}} \sum_{i=1}^{n} Y_i,
$$

where

$$
Y_i = \sum_{a=1}^{A} \left( \frac{W_{ni}^{(a)}}{\pi} g_a(X_i) - \mathrm{E}\, g_a(X_1) \right), \quad i = 1, \ldots, n,
$$

are independent, identically distributed and centered random vectors. Therefore, by the central limit theorem, $G_{nn}^* m$ goes to a centered Gaussian random vector with variance-covariance matrix $(1 - \pi)\, \mathrm{E} \sum_a g_a(X_1) g_a(X_1)^\top + \pi \sum_{a,b} (\mathrm{E}\, g_a g_b^\top - \mathrm{E}\, g_a\, \mathrm{E}\, g_b^\top)$.

*Case* $\pi_n \to 0$. We have

$$\begin{aligned} G_{nn}^* m =& \frac{1}{\sqrt{n\pi_n}} \sum_{i=1}^{n} \left( \sum_{a=1}^{A} W_{ni}^{(a)} g_a(X_i) - \pi_n \,\mathrm{E}\, g_a(X_1) \right) \\ =& \frac{1}{\sqrt{n\pi_n}} \sum_{i,a} (W_{ni}^{(a)} - \pi_n) g_a(X_i) + \sqrt{n\pi_n} \left( \frac{1}{n} \sum_{i,a} g_a(X_i) - \mathrm{E}\, g_a(X_1) \right), \end{aligned}$$

where the second term is of order $\sqrt{\pi_n} O_P(1)$ and hence vanishes in probability as $n \to \infty$. It remains to show that the first term goes to a Gaussian distribution. By Lindeberg-Feller's central limit theorem (see e.g. [40, p. 20]), this is true under two conditions:

(C1) $\displaystyle\sum_i \mathrm{Var} \left[ \frac{1}{\sqrt{n\pi_n}} \sum_a (W_{ni}^{(a)} - \pi_n) g_a(X_i) \right] \to \Sigma,$

(C2) For all $\epsilon > 0$,

$$\sum_i \mathrm{E} \left[ \left\| \frac{1}{\sqrt{n\pi_n}} \sum_a (W_{ni}^{(a)} - \pi_n) g_a(X_i) \right\|^2 \mathbf{1} \left\{ \left\| \frac{1}{\sqrt{n\pi_n}} \sum_a (W_{ni}^{(a)} - \pi_n) g_a(X_i) \right\| > \epsilon \right\} \right] \to 0.$$

Since the random vectors $\sum_a (W_{ni}^{(a)} - \pi_n) g_a(X_i)$, $a = 1, \ldots, A$, are independent and identically distributed, the condition (C1) boils down to

$$\frac{1}{\pi_n} \mathrm{Var} \left( \sum_a (W_{n1}^{(a)} - \pi_n) g_a(X_1) \right) \to \Sigma.$$

Thanks to the independence between $\{W_{n1}^{(a)},\, a = 1, \ldots, A\}$ and $X_1$, the $l$th row and $l'$th column of the variance-covariance matrix

$$\begin{aligned} &\mathrm{Var} \left( \sum_a (W_{n1}^{(a)} - \pi_n) g_a(X_1) \right) \\ &= \mathrm{E}\, \mathrm{E} \left( \left[ \sum_a (W_{n1}^{(a)} - \pi_n) g_a(X_1) \right] \left[ \sum_a (W_{n1}^{(a)} - \pi_n) g_a(X_1) \right]^\top \Big| X_1 \right) \end{aligned}$$

is given by

$$\begin{aligned} &\mathrm{E} \sum_{a,a'} g_{al}(X_1) g_{a'l'}(X_1)\, \mathrm{E}(W_{n1}^{(a)} - \pi_n)(W_{n1}^{(a')} - \pi_n) \\ &= \mathrm{E}\, \pi_n(1 - \pi_n) \sum_a g_{al}(X_1) g_{al'}(X_1). \end{aligned}$$

Thus, the left-hand side in the condition (C1) is $(1-\pi_n)\,\mathrm{E}\sum_a g_a(X_1)g_a(X_1)^\top$ and we have shown that it goes to $\Sigma = \mathrm{E}\sum_a g_a(X_1)g_a(X_1)^\top$.

Let us now show that the condition (C2) holds. Choosing the Euclidean norm, the condition boils down to

$$\mathrm{E}\left[\mathrm{E}\left(\left\|\sum_{a=1}^{A}\frac{W_{n1}^{(a)}-\pi_n}{\sqrt{\pi_n}}g_a(X_1)\right\|^2 B_n\middle|X_1\right)\right]\to 0,$$

where $B_n = \mathbf{1}\left\{\left\|\sum_a (W_{n1}^{(a)}-\pi_n)g_a(X_1)\right\| > \epsilon\sqrt{n\pi_n}\right\}$. The inner expectation is bounded by

$$2^{A-1}\sum_{a=1}^{A}\sum_{l=1}^{L}\mathrm{E}\left(\left(\frac{W_{n1}^{(a)}-\pi_n}{\sqrt{\pi_n}}\right)^2 g_{al}(X_1)^2 B_n\middle|X_1\right)$$

By Cauchy-Schwartz's inequality and the independence between $X_1$ and $W_{n1}^{(a)}$, the expectation above is less than

$$\sqrt{\mathrm{E}\left(\frac{W_{n1}^{(a)}-\pi_n}{\sqrt{\pi_n}}\right)^4}\sqrt{g_{al}(X_1)^4\,\mathrm{E}(B_n|X_1)}.$$

Straightforward calculations show that the first factor is equivalent to $1/\sqrt{\pi_n}$. Let us bound the second one. We have

$$\begin{aligned}
\mathrm{E}(B_n|X_1) &= P\left(\left\|\sum_a (W_{n1}^{(a)}-\pi_n)g_a(X_1)\right\|_2 > \epsilon\sqrt{n\pi_n}\middle|X_1\right)\\
&\le P\left(\left\|\sum_a (W_{n1}^{(a)}-\pi_n)g_a(X_1)\right\|_\infty > \frac{\epsilon\sqrt{n\pi_n}}{\sqrt{L}}\middle|X_1\right)\\
&\le \sum_{l=1}^{L} P\left(\left|\sum_a (W_{n1}^{(a)}-\pi_n)g_{al}(X_1)\right| > \frac{\epsilon\sqrt{n\pi_n}}{\sqrt{L}}\middle|X_1\right)\\
&\le \sum_{l=1}^{L} 2\exp\left(-\frac{2n\pi_n\epsilon^2}{L\sum_a 4(1-\pi_n)^2|g_{al}(X_1)|^2}\right).
\end{aligned}$$

The last inequality is an application of Hoeffding's inequality, see e.g [39, p. 33]. Gluing the pieces together, the left-hand side in condition (C2) is bounded above by

$$2^{A-1/2}\sum_{a=1}^{A}\sum_{l=1}^{L}\sqrt{\sum_{l'=1}^{L}\mathrm{E}\,\frac{g_{al}(X_1)^4}{\pi_n}\exp\left(-\frac{2n\pi_n\epsilon^2}{L\sum_{a'=1}^{A}4(1-\pi_n)^2|g_{a'l'}(X_1)|^2}\right)}.$$

The condition in Lemma A.3 implies that the expectation above goes to zero. The proof is complete.

# C Proofs of the Propositions

## Proof of Proposition 1

We begin with a lemma.

**Lemma C.1.** *Let $w_a > 0$ for all $a \in \mathcal{A}$. If the two statements*

*(i) $\theta_0$ is a maximizer of $L_a$ for every $a \in \mathcal{A}$*

*(ii) $\theta \neq \theta'$ implies that there exists a pair $a$ such that $L_a(\theta) \neq L_a(\theta')$*

*are true then the maximizer of $\theta \mapsto \sum_a w_a L_a(\theta)$ is unique.*

*Proof.* If $\theta_0'$ were another maximizer of $\sum_a w_a L_a$ then there is $a \in \mathcal{A}$ such that $w_a L_a(\theta_0') < w_a L_a(\theta_0)$. But then $\sum_a w_a L_a(\theta_0') < \sum_a w_a L_a(\theta_0)$, which is a contradiction. □

It is straightforward to show that Lemma C.1 (i) is true. It remains to ensure that Lemma C.1 (ii) is true as well. Take $a = \{i, j\} \in \mathcal{A}$, choose $\theta, \theta' \in \Theta$ and assume $L_a(\theta) = L_a(\theta')$. By (ii) of the Proposition, $\mathrm{E} \log \tilde{f}_a(X_{1i}, X_{1j}; v_a(\theta)) = \mathrm{E} \log \tilde{f}_a(X_{1i}, X_{1j}; v_a(\theta'))$ and hence, by (i), $v_a(\theta) = v_a(\theta')$. Since the pair $a$ was arbitrary, (iii) implies $\theta = \theta'$. The proof is complete.

## Proof of Proposition 2

In this case the functions $\Phi_a$ in Theorem 3 are bounded by a constant, say $C$. Let $A$ be the cardinal of $\mathcal{A}$. The left hand side of (6) is bounded by

$$\frac{1}{\pi_n} C^4 \exp\left(\frac{-n\pi_n\kappa}{AC^2}\right),$$

which goes to zero because $\pi_n^{-1} e^{-\pi_n^{-1}} \to 0$ and $\exp([AC^2 - n\pi_n^2\kappa]/[AC^2\pi_n]) \leq 1$ as soon as $n\pi_n^2\kappa \geq AC^2$.

## Proof of Proposition 3

*Assumption 1:* Clearly, for all $x \in \mathbf{R}^2$,

$$\max\left(\left|\frac{\partial \ell_a(x;\theta)}{\partial \theta}\right|, \left|\frac{\partial^2 \ell_a(x;\theta)}{\partial \theta^2}\right|, \left|\frac{\partial^3 \ell_a(x;\theta)}{\partial \theta^3}\right|, \right) \leq \varphi(\theta)(1 + \|x\|^2),$$

for some positive and continuous function $\varphi$ defined on $(-1/(d-1) + \epsilon, 1 - \epsilon)$. This set can be extended to the compact set $[-1/(d-1) + \epsilon/2, 1 - \epsilon/2]$ and hence

$$\mathrm{E}\, \Phi_a(X_1; \theta_0)^2 \leq C(1 + \|x\|^2)^2 \tag{19}$$

for some constant $C$. (Remember that $\theta_0$ is the true parameter.) Since $\mathrm{E}(1 + \|X_1\|^2)^2 < \infty$, the first statements in Assumption 1 have been checked. Also, it

is clear that the derivatives can be passed under the integral sign. Assumption 1 has been checked.

*Assumption 2:* We have

$$L_a(\theta) = -\frac{\log(1-\theta^2)}{2} - \frac{1}{1-\theta^2} + \frac{\theta\theta_0}{1-\theta^2} + \text{ constant}$$

and hence $\partial L_a(\theta)\partial\theta = 0$ iff $-\theta^3 + \theta_0\theta^2 - \theta + \theta_0 = 0$. This polynomial in $\theta$ has only one real root (the two other are complex) and hence the maximizer of $\sum_a L_a(\theta) = d(d-1)L_{12}(\theta)/2$ is unique.

## Proof of Proposition 4

In view of (19) and since the the left hand side in (6) is an increasing function of $\Phi_a$, $a \in \mathcal{A}$, it suffices to show that

$$\begin{aligned}
&\mathrm{E}\left(1 + \|X_1\|^2\right)^4 \exp\left(-\frac{n\pi_n\kappa}{(1+\|X_1\|^2)^2}\right) \\
&\propto \int_{\mathbf{R}^d} (1+\|x\|^2)^4 \exp\left(-\frac{n\pi_n\kappa}{(1+\|x\|^2)^2}\right) \exp\left(-\frac{1}{2}x^\top \Sigma_{\theta_0}^{-1} x\right) \mathrm{d}x \\
&\le \int_{\mathbf{R}^d} (1+\|x\|^2)^4 \exp\left(-\frac{n\pi_n\kappa}{(1+\|x\|^2)^2} - \frac{\|x\|^2}{4\lambda_{\max}}\right) \mathrm{d}x \\
&= \int_0^\infty (1+r^2)^4 r^{d-1} \exp\left(-\frac{n\pi_n\kappa}{(1+r^2)^2} - \frac{r^2}{4\lambda_{\max}}\right) \mathrm{d}r
\end{aligned}$$

is of order $o(\pi_n)$ for all $\kappa > 0$. The inequality above is true because $\Sigma_{\theta_0}^{-1} - 1/(4\lambda_{\max})I$ is positive definite. The last equality holds by a change of variables [3]. Since $(1+r^2)^4 r^{d-1}$ is a polynomial in $r$, the last integral is a sum of integrals of the form given in Lemma D.2 and hence, by Corollary D.1, it is of order $O(\exp(-[n\pi_n\kappa]^{1/3}/(8\lambda_{\max} \vee 1)))$ whenever $n\pi_n \to \infty$. Substituting $\pi_n = n^{-\alpha}$ with $0 < \alpha \le 1/4$ and letting $n$ go to infinity completes the proof.

## Proof of Proposition 5

It suffices to check (i), (ii) and (iii) in Proposition 1. Let $a = \{i, j\}$. Put $v_a(\theta) = v_a(\mu_i, \mu_j, \rho) = (\mu_i, \mu_j, w_a(\rho))$ so that range $v_a = \Theta_i \times \Theta_j \times \operatorname{range} w_a$. The condition (ii) in Proposition 1 is checked because $F_a(x_i, x_j; \theta) = C_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); \rho) = \tilde{C}_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); w_a(\rho)) =: F_a(x_i, x_j; v_a(\theta))$. These distribution functions define a family indexed by range $v_a$. This family is identifiable: if $(\mu_i, \mu_j, \varrho), (\mu_i', \mu_j', \varrho') \in \operatorname{range} v_a$ and $\tilde{C}_a(F_{\mu_i}(x_i), F_{\mu_j}(x_j); \varrho)) = \tilde{C}_a(F_{\mu_i'}(x_i), F_{\mu_j'}(x_j); \varrho')$ then letting $x_i \to \infty$ yields that $\mu_j = \mu_j'$ and by the same token $\mu_i = \mu_i'$ and hence $\varrho = \varrho'$. Thus the condition (i) in Proposition 1 is true. Finally, choose $\theta = (\mu_1, \ldots, \mu_d, \rho)$ and $\theta' = (\mu_1', \ldots, \mu_d', \rho')$ in $\Theta$. If $V(\theta) = V(\theta')$ then clearly $\mu_1 = \mu_1', \ldots, \mu_d = \mu_d'$ and $w_a(\rho) = w_a(\rho')$ for all $a \in \mathcal{A}$. But then $\rho = \rho'$ because the mapping $W$ is one-to-one. Thus the last condition (iii) in Proposition 1 is checked.

# D Bound on an integral

**Lemma D.1.** *If $f$ is a function defined by*

$$f(x) = \frac{-\alpha \log x}{x^2} + \frac{\lambda}{(\beta + \gamma x^4)x^2},$$

*$x > 0$, $\lambda > 0$, $\alpha > 0$, $\beta > 0$, $\gamma > 0$, then there is $x^\star \in (0, \infty)$ such that $f(x) \geq f(x^\star)$ for all $x$ and $-\alpha(1-2\log x^\star)(\beta+\gamma x^{\star 4})^2 - \lambda\gamma x^{\star 4} = 2\lambda\beta$. Moreover, $f(x^\star) \to 0$ as $\lambda \to \infty$.*

*Proof.* We have $f'(x) \geq 0$ iff

$$-\alpha(1 - 2\log x)(\beta + \gamma x^4)^2 - \lambda\gamma x^4 \geq 2\lambda\beta. \tag{20}$$

Note that if $x \leq e^{1/2}$ then $f'(x) \leq 0$. Otherwise, (20) is equivalent to

$$x^4(\varphi_1(x) + \varphi_2(x) - \lambda\gamma) + \varphi_3(x) \geq 2\lambda\beta, \tag{21}$$

where $\varphi_1(x) = -\alpha\gamma^2(1 - 2\log x)x^4$, $\varphi_2(x) = -2\alpha\beta\gamma(1 - 2\log x)$ and $\varphi_3(x) = -\alpha\beta^2(1 - 2\log x)$. The functions $\varphi_1$, $\varphi_2$ and $\varphi_3$ are increasing and nonnegative on $[e^{1/2}, \infty)$. Thus the function in the left-hand side of (21) is continuous and increasing and is equal to $-\lambda\gamma e^2$ at $e^{1/2}$. Therefore, it reaches $2\lambda\beta$ at a unique point $x^\star > e^{1/2}$; this point satisfies (21) and hence (20) with "=" instead of "≥". It follows that the function $f$ is decreasing on $(0, x^\star)$, reaches its global minimum at $x^\star$ and is increasing on $(x^\star, \infty)$. It remains to show that $f(x^\star) \to 0$ as $\lambda \to \infty$. We have

$$f(x^\star) = \frac{-\alpha \log x^\star}{x^{\star 2}} + \frac{\lambda}{(\beta + \gamma x^{\star 4})x^{\star 2}}$$

and from (21) we know that $x^\star \to \infty$. This implies that the limit is as required. □

**Lemma D.2.** *Let $\varpi$ be such that $\sqrt{2\varpi} = \int e^{-x^2/2}\,\mathrm{d}x$. If*

$$I(\lambda) = \int_0^\infty x^\alpha \exp\left[-\frac{\lambda}{(1+x^2)^2} - \frac{x^2}{2\sigma^2}\right] \mathrm{d}x,$$

*$\sigma > 0$, $\lambda > 0$, $\alpha > 0$, then for every $0 < \gamma \leq 1$, there are $\eta_0 > 0$ and $\lambda_0 > 0$ such that*

$$I(\lambda) \leq \left(\eta^\alpha \exp\left[-\frac{\lambda}{1+\gamma\eta^4}\right]\frac{\sigma\sqrt{2\varpi}}{2} + \exp\left[-\frac{\eta^2}{4\sigma^2}\right]\right)\exp\left[\frac{\lambda}{1+\gamma\eta^4} - \frac{\lambda}{(1+\eta^2)^2}\right]$$

*for all $\eta > \eta_0$ and $\lambda > \lambda_0$.*

*Proof.* Choose $0 < \gamma \leq 1$ and put $f(x) := (1+\gamma x^4)^{-1} - (1+x^2)^{-2}$. There is $\eta_0 > 0$ such that $f'(x) < 0$ for all $x > \eta_0$. Now choose $\eta > \eta_0$. Then

$$\begin{aligned} B &:= \int_\eta^\infty x^\alpha \exp\left[-\frac{\lambda}{(1+x^2)^2} - \frac{x^2}{2\sigma^2}\right] \mathrm{d}x \\ &\leq \exp\left[\lambda f(\eta)\right] \int_\eta^\infty x^\alpha \exp\left[-\frac{\lambda}{1+\gamma x^4} - \frac{x^2}{2\sigma^2}\right] \mathrm{d}x. \end{aligned}$$

Let $\nu > 0$. The integrand above is bounded by $\exp[-x^2/(2\nu^2)]$ iff $-2\alpha \log(x)/x^2 + 2\lambda/[(1+\gamma x^4)x^2] \geq 1/\nu^2 - 1/\sigma^2$. In the above inequality, the function in the left is bounded below by some constant that goes to zero as $\lambda$ goes to infinity. (See Lemma D.1.) Taking $\nu^2 = 2\sigma^2$ ensures that the inequality is true for all $x$ as soon as $\lambda$ is greater than some number $\lambda_0$. Therefore,

$$\begin{aligned} B &\leq \exp\left[\frac{\lambda}{1+\gamma\eta^4} - \frac{\lambda}{(1+\eta^2)^2}\right] \int_\eta^\infty \exp\left[-\frac{x^2}{4\sigma^2}\right] \mathrm{d}x \\ &\leq \exp\left[\frac{\lambda}{1+\gamma\eta^4} - \frac{\lambda}{(1+\eta^2)^2}\right] \exp\left[-\frac{\eta^2}{4\sigma^2}\right] \end{aligned}$$

for all $\eta > \eta_0$ and $\lambda > \lambda_0$. Finally,

$$\begin{aligned} A &:= \int_0^\eta x^\alpha \exp\left[-\frac{\lambda}{(1+x^2)^2} - \frac{x^2}{2\sigma^2}\right] \mathrm{d}x \\ &\leq \eta^\alpha \exp\left[-\frac{\lambda}{(1+\eta^2)^2}\right] \int_0^\eta \exp\left[-\frac{x^2}{2\sigma^2}\right] \mathrm{d}x \\ &\leq \eta^\alpha \exp\left[-\frac{\lambda}{(1+\eta^2)^2}\right] \frac{\sigma\sqrt{2\varpi}}{2} \end{aligned}$$

and, since $I(\lambda) = A + B$, the proof is complete. □

**Corollary D.1.** *The integral $I(\lambda)$ defined in Lemma D.2 satisfies*

$$I(\lambda) = O\left(\exp\left[-\frac{\lambda^{1/3}}{4\sigma^2 \vee 2}\right]\right), \quad \lambda \to \infty.$$

*Proof.* In Lemma D.2, we may take $\eta = \lambda^a$, $a > 0$, because both $\eta$ and $\lambda$ are allowed to go to infinity. If, furthermore, $a < 1/4$, then the first factor in the upper bound go to zero. If $\gamma = 1$ and $a \geq 1/6$ then the second factor goes to a nonnegative constant, say $K$. Now, with $\gamma = 1$ and $a = 1/6$,

$$\begin{aligned} &\left(\lambda^{\alpha/6} \exp\left[-\frac{\lambda}{1+\lambda^{2/3}}\right] \frac{\sigma\sqrt{2\varpi}}{2} + \exp\left[-\frac{\lambda^{1/3}}{4\sigma^2}\right]\right) \exp\left[\frac{\lambda^{1/3}}{4\sigma^2 \vee 2}\right] \\ &= \lambda^{\alpha/6} \exp\left[\frac{\lambda^{1/3}}{4\sigma^2 \vee 2} - \frac{\lambda^{1/3}}{\lambda^{-2/3}+1}\right] \frac{\sigma\sqrt{2\varpi}}{2} + \exp\left[\frac{\lambda^{1/3}}{4\sigma^2 \vee 2} - \frac{\lambda^{1/3}}{4\sigma^2}\right]. \end{aligned}$$

The limit is zero if $4\sigma^2 < 2$ and one if $4\sigma^2 \geq 2$. Therefore the limit of $I(\lambda)\exp[\lambda^{1/3}/(4\sigma^2 \vee 2)]$ is at most $K$. The proof is complete. □

# E Supplementary Figures

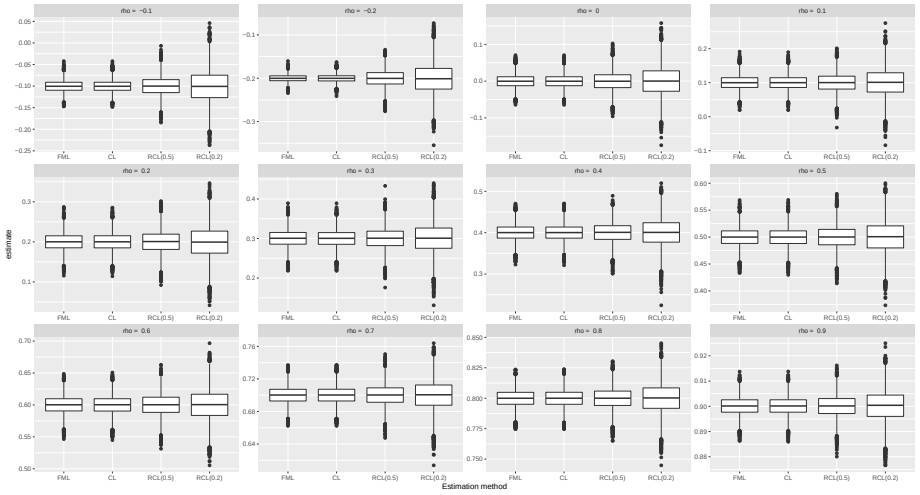


Figure S1: Boxplots of parameter estimates for $n = 500$ across 50,000 simulated datasets using the full maximum likelihood (FML), composite pairwise likelihood (CL), and randomized pairwise composite likelihood (RCL) approaches with $\pi = 0.5$ and 0.2 for $\rho = \{-0.1 \ldots, 0.9\}$.

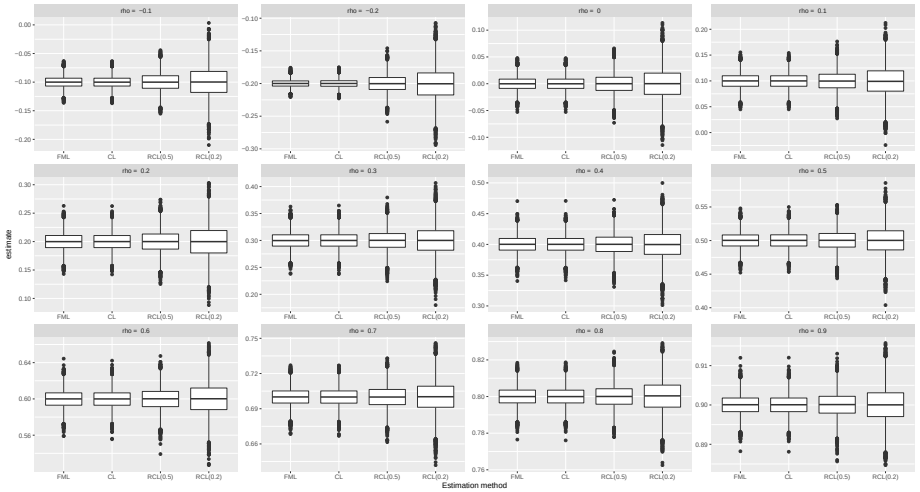


Figure S2: Boxplots of parameter estimates for $n = 1000$ across 50,000 simulated datasets using the full maximum likelihood (FML), composite pairwise likelihood (CL), and randomized pairwise composite likelihood (RCL) approaches with $\pi = 0.5$ and 0.2 for $\rho = \{-0.1 \ldots, 0.9\}$.

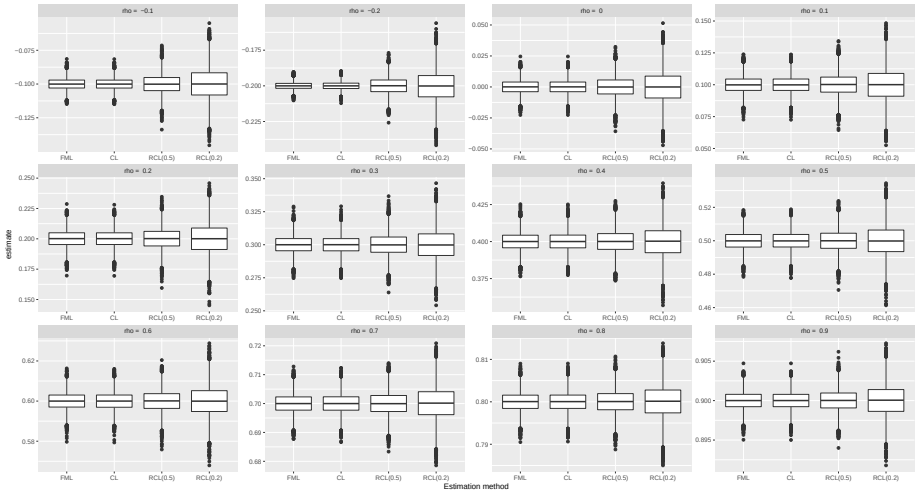


Figure S3: Boxplots of parameter estimates for $n = 5000$ across 50,000 simulated datasets using the full maximum likelihood (FML), composite pairwise likelihood (CL), and randomized pairwise composite likelihood (RCL) approaches with $\pi = 0.5$ and 0.2 for $\rho = \{-0.1 \ldots, 0.9\}$.

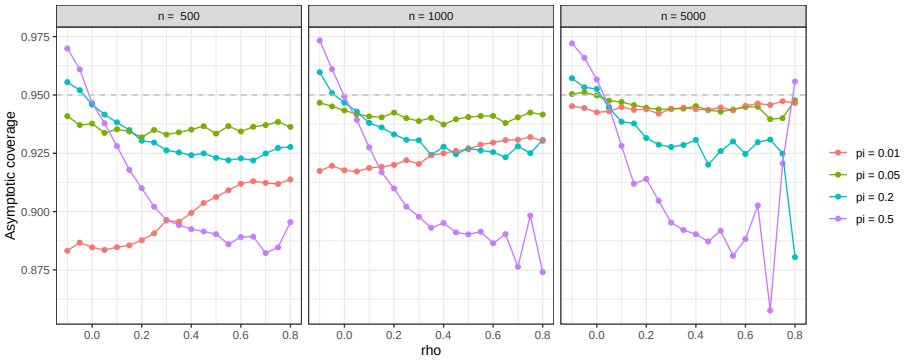


Figure S4: Asymptotic coverage for the compound symmetry example, with $\alpha = 5\%$, averaged over 50,000 replications. The values represent the proportion of times the asymptotic interval contains the true value used to simulate the data.

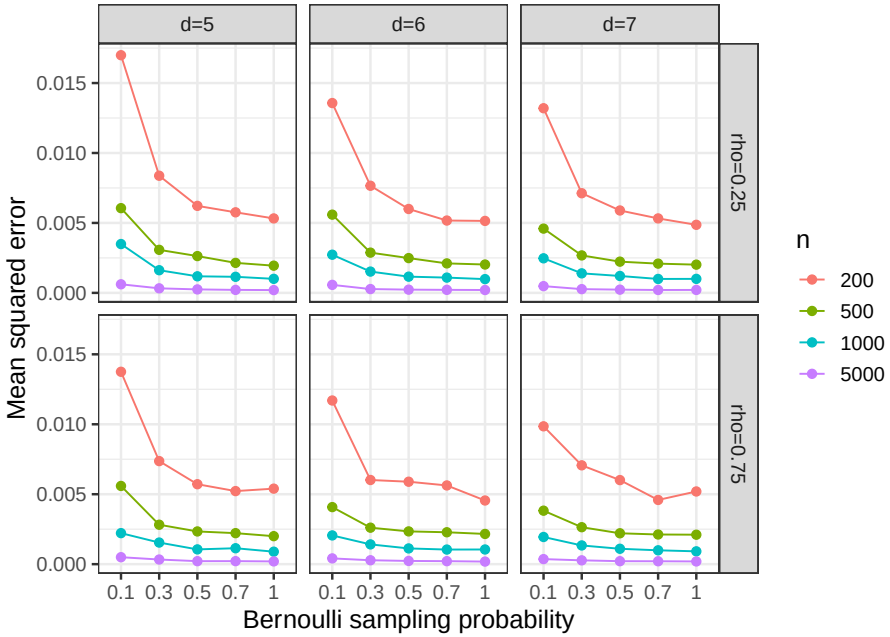


Figure S5: Mean squared error averaged across 300 replications using the randomized pairwise likelihood approach with $\pi = \{0.1, 0.3, 0.5, 0.7, 1\}$ for simulated data with dimension $d = 5, 6, 7$, copula dispersion $\rho = 0.25$ or 0.75, and number of observations ranging from $n = 200$ to 5000.

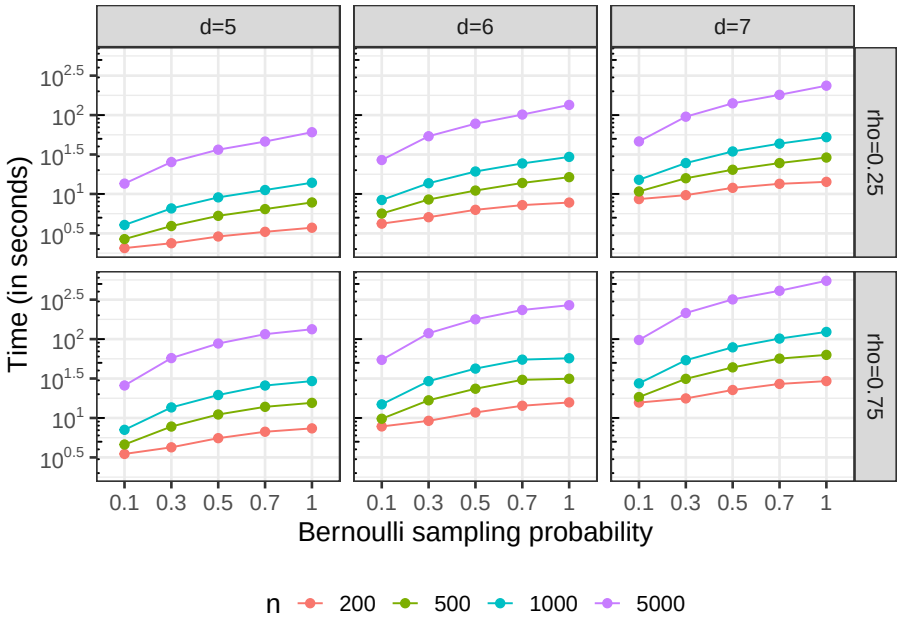


Figure S6: Average computational time in seconds (across 300 replications) using the randomized pairwise likelihood approach with $\pi = \{0.1, 0.3, 0.5, 0.7, 1\}$ for simulated data with dimension $d = 5, 6$ or 7, copula dispersion $\rho = 0.25$ or 0.75, and number of observations ranging from $n = 200$ to 5000.

# References


[1] P. Berkhout and E. Plug. A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica*, 58(3):349–364, 2004.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[3] L. E. Blumenson. A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly*, 67(1):63–66, 1960.

[4] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[5] J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

[6] J. Chiquet, S. Robin, and M. Mariadassou. Variational inference for sparse network reconstruction from count data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1162–1171, 2019.

[7] A. Cougoul, X. Bailly, and E. C. Wit. MAGMA: inference of sparse microbial association networks. *bioRxiv*, http://dx.doi.org/10.1101/538579, 2019.

[8] D. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

[9] R. A. Davis and C. Y. Yau. Comments on pairwise likelihood in time series models. *Statistica Sinica*, 21(1):255–277, 2011.

[10] J. V. Dillon and G. Lebanon. Stochastic composite likelihood. *Journal of Machine Learning Research*, 11(Oct):2597–2633, 2010.

[11] K. Faust and J. Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10:538–550, 2012.

[12] S. Fieuws and G. Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006.

[13] C. Genest and J. Nešlehová. A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515, 2007.

[14] Z. Huang and D. Ferrari. Fast construction of efficient composite likelihood equations. arXiv preprint arXiv:1709.03234, 2017.

[15] F. K. Hui, S. Müller, and A. Welsh. Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association*, 113(524):1759–1769, 2018.

[16] D. I. Inouye, E. Yang, G. I. Allen, and P. Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398, 2017.

[17] B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen *Eryshiphe alphitoides*. *Environmental Microbiology*, 72:870–880, 2016.

[18] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, London, 1997.

[19] H. Joe. *Dependence modeling with copulas*. CRC press, 2014.

[20] H. Joe and Y. Lee. On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100(4):670–685, 2009.

[21] D. Karlis and L. Meligkotsidou. Multivariate Poisson regression with covariance structure. *Statistics and Computing*, 15(4):255–265, 2005.

[22] D. Karlis and L. Meligkotsidou. Finite multivariate Poisson mixtures with applications. *Journal of Statistical Planning and Inference*, 137:1942–1960, 2007.

[23] A. Y. Kuk. A hybrid pairwise likelihood method. *Biometrika*, 94(4):939–952, 2007.

[24] A. Y. Kuk and D. J. Nott. A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, 47(4):329–335, 2000.

[25] B. G. Lindsay. Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239, 1988.

[26] B. G. Lindsay, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105, 2011.

[27] R. Momal, S. Robin, and C. Ambroise. Tree-based inference of species interaction network from abundance data, 2019.

[28] F. Mondet, A. Rau, C. Klopp, M. Rohmer, D. Severac, Y. Le Conte, and C. Alaux. Transcriptome profiling of the honeybee parasite *varroa destructor* provides new biological insights into the mite adult life cycle. *BMC Genomics*, 19(328), 2018.

[29] R. Nelsen. *An introduction to copulas*. Springer, 2006.

[30] A. K. Nikoloulopoulos. Copula-based models for multivariate discrete response data. In *Copulae in Mathematical and Quantitative Finance*, pages 231–249. Springer, 2013.

[31] A. K. Nikoloulopoulos, H. Joe, and N. R. Chaganty. Weighted scores method for regression models with dependent data. *Biostatistics*, 12(4):653–665, 2011.

[32] D. J. Nott and T. Rydén. Pairwise likelihood methods for inference in image models. *Biometrika*, 86(3):661–676, 1999.

[33] S. A. Padoan, M. Ribatet, and S. A. Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277, 2010.

[34] A. Panagiotelis, C. Czado, and H. Joe. Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072, 2012.

[35] I. Papageorgiou and I. Moustaki. Sampling of pairs in pairwise likelihood estimation for latent variable models with categorical observed variables. *Statistics and Computing*, 29:351–365, 2019.

[36] D. Pollard. *Convergence of stochastic processes*. Springer, 1984.

[37] D. Pollard. New ways to prove central limit theorems. *Econometric Theory*, 1(3):295–313, 1985.

[38] M. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(R25), 2010.

[39] S. A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

[40] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[41] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[42] C. Varin, G. Høst, and Ø. Skare. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational statistics & data analysis*, 49(4):1173–1191, 2005.

[43] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

[44] C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005.

[45] V. G. Vasdekis, D. Rizopoulos, and I. Moustaki. Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics*, 15(4):677–689, 2014.

[46] X. Wang and Y. Wu. Theoretical properties of composite likelihoods. *Open Journal of Statistics*, 4:188–197, 2014.

[47] J. Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007.