A randomized pairwise likelihood method for complex statistical inferences

Gildas Mazo, Dimitris Karlis, Andrea Rau

To cite this version:

Gildas Mazo, Dimitris Karlis, Andrea Rau. A randomized pairwise likelihood method for complex statistical inferences. 2022. hal-03126620v3

HAL Id: hal-03126620
https://hal.inrae.fr/hal-03126620v3

Preprint submitted on 29 Nov 2022 (v3), last revised 13 Sep 2023 (v5)

A randomized pairwise likelihood method for complex statistical inferences

Gildas Mazo[1], Dimitris Karlis[2], and 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

Abstract

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

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

Pairwise likelihood methods have been successfully used in many applications (Varin et al., 2011). Many variants have been derived to accommodate specific models, data or tasks, notably for multivariate binary data (le Cessie and Van Houwelingen, 1994; Kuk and Nott, 2000), as well as spatial and image data. An early approach for spatial models used conditionally specified likelihoods for spatial image data (Besag, 1975). More recently, pairwise likelihoods were used for binary or indicator data in space (Heagerty and Lele, 1998), for spatial-clustered data (Bai et al., 2014), and with random field models for image data (Nott and Rydén, 1999). Several authors have further proposed ways to improve the efficiency of composite likelihood methods (Ferrari et al., 2016), primarily by adding weights to the component likelihoods (see, e.g. Joe and Lee, 2009). It appears, however, that finding and estimating the optimal weights in general is a very difficult problem which may not have a solution (Lindsay et al., 2011). In the following, we shall focus on the pairwise likelihood, the most popular version of composite likelihoods.

In high dimensions, applying the pairwise likelihood method may be cumbersome. With d variables, the number of pairs is of order d^2. To get confidence intervals, one needs to compute a double sum over pairs of pairs of order d^4 and all of the four-dimensional marginals. All this can be time-consuming and burdensome. Although there is little literature on how to address these computational issues, several research directions have been proposed. For instance, instead of taking all of the pairs, one can consider a small subset (Huang and Ferrari, 2021; Papageorgiou and Moustaki, 2019), although selecting a good subset is a difficult problem. Some heuristics were proposed in Papageorgiou and Moustaki (2019), but no theoretical justification was provided and the asymptotic properties of the estimators are unknown. For spatial data, a simple approximation to the likelihood can be obtained by using only sufficiently close points (Vecchia, 1988), although such an approach can be improved by also considering some distant pairs (Stein et al., 2004) or using a spatial blocking strategy (Eidsvik et al., 2014). In the case of spatiotemporal data, Bai et al. (2012) proposed selecting pairs representing spatial, temporal and spatiotemporal effects, while Bevilacqua et al. (2012) proposed a weighted approach with a corresponding information criteria for model selection. In Huang and Ferrari (2021), pair selection was performed by regularization to identify informative pairs of variables. However, like all methods that select a subset of pairs, some of them are necessarily dropped. This implicitly assumes that all model parameters can be estimated from only a subset of pairs. In the context of conditional random fields, a stochastic combination of low-dimensional conditional likelihoods was proposed in Dillon and Lebanon (2010). This allows a reduction in the number of times the conditional log-densities of the model are evaluated, but it does not solve the problem for the construction of confidence intervals.

To alleviate the computational issues of the pairwise likelihood method, we consider a randomized pairwise likelihood approach. Only summands randomly sampled across observations and pairs are used for the estimation of the parameters. To implement this strategy, one draws, for each sample size n, i.i.d. Bernoulli weights $W_{ni}^{(a)}$, $i = 1, \ldots, n$, $a \in \{\{1, 2\}, \ldots, \{d - 1, d\}\}$, with parameter π_n; all summands for which $W_{ni}^{(a)} = 0$ are discarded. A fundamental point here is that we allow the sampling parameter π_n to decrease with n—we shall come back to this later. The sampling parameter controls the tradeoff between the computational complexity and the statistical efficiency. An intuitive way to see this is to notice that the average number of summands needed to compute the randomized pairwise likelihood is equal to $n\pi_n d(d - 1)/2$. However, there is an additional reason why π_n permits a reduction of the computational cost. By letting $\pi_n \to 0$, the bivariate log density gradients become asymptotically independent, leading to the disappearance of the term of computational complexity d^4 in the estimator's asymptotic variance. In practice, this suggests that one may be able to approximate confidence intervals at a much lower cost than in the standard pairwise likelihood method.

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

2 Maximum pairwise likelihood inference

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

2.1 Definition, assumptions and asymptotic properties

Let $X_i := (X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$, 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 Θ of $\mathbf{R}^q$. Let θ_0 denote the element of Θ such that $f_0(\bullet) = f(\bullet;\theta_0)$. Let $\mathcal{A}$ be the set of all pairs of variables. Its cardinal is $d(d-1)/2$. The pairs in $\mathcal{A}$ are ordered in the lexicographical order. Denote by $f_a(\cdot,\cdot;\theta)$ the marginal density corresponding to the pair a and write $\ell_a(\cdot,\cdot;\theta)$ for $\log f_a(\cdot,\cdot;\theta)$. Whenever it exists, denote by $\dot{\ell}_a(\cdot,\cdot;\theta)$ the gradient of $\ell_a(\cdot,\cdot;\theta)$ with respect to θ. Whenever a function is encountered with a bullet symbol, it means that the argument it replaces is a vector with three components or more. Otherwise, there are as many dot symbols as there are components. If $a = \{j, j'\}$ is a pair then $(X_{ij}, X_{ij'})$ 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 log-likelihood function. From the viewpoint of M-estimation theory, the population pairwise likelihood is the objective criterion function, the maximizer of which is the parameter of interest. In this case the objective criterion is the sum of "bivariate" Kullback-Leibler information criteria. This is the viewpoint we shall adopt throughout the paper. Wang and Wu (2014) provide a different view. According to them, maximizing the pairwise likelihood function can also be seen as maximizing the full Kullback-Leibler information under some information constraints.

We call the maximum pairwise likelihood estimator (MPLE) every element $\hat{\theta}_n^{\mathrm{MPL}}$ of Θ that satisfies $L_n^{\mathrm{PL}}(\hat{\theta}_n^{\mathrm{MPL}}) \geq L_n^{\mathrm{PL}}(\theta)$ for all θ in some compact subset of Θ. 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 θ is estimated contains θ_0.

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

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

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

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

It is easy to see that each L_a is maximized at θ_0 and hence so is the mapping $\sum_a L_a(\theta)$. Thus, we deduce from Assumption 2 that θ_0 is the only maximizer of $\sum_a L_a(\theta)$.

Remark 1. *Even if* θ_0 *is the only maximizer of* $\sum_a L_a(\theta)$*, it does not mean that* θ_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* ρ_0*, so that* $\theta_0 = (\mu_{01}, \mu_{02}, \mu_{03}, \rho_0)$*. Then not only is* L_{12} *maximized at* θ_0*, but also at* $(\mu_{01}, \mu_{02}, \mu, \rho_0)$ *for any* μ*.*

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

Proposition 1. *If, for every* $a \in \mathcal{A}$*, there is a function* v_a *on* Θ *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* θ

(iii) the mapping $V(\theta) := (v_a(\theta))_{a\in\mathcal{A}}$ *is one-to-one*

then Assumption 2 holds.

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

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

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

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

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

The problem of choosing the optimal weights is difficult. In the one-dimensional case, that is, when the parameter is a scalar, a formula for the optimal weights exists but it requires the computation of the between-scores correlation matrix C. This can be computationally challenging, as we shall see next. In the more realistic multivariate case, according to Lindsay (1988), a solution may not exist, and if it existed it would be difficult to compute.

2.2 Computational issues in higher dimensions

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

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

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

3 The randomized pairwise likelihood method

We introduce a new estimator of θ_0 based on a randomized version of the pairwise log-likelihood function and thus cheaper to compute. Interestingly, confidence intervals can be computed with no more than $O(d^2)$ computations.

3.1 Definition and first results

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

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

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 $\theta_0 = (\theta_{01}, \ldots, \theta_{0q})$ is estimated by maximizing the function in (3). In practice, one first draws the Bernoulli weights, which allows certain terms to be excluded from the pairwise log-likelihood function, and then maximizes the sum of the remaining terms. If $\pi_n = 1$ then $\Pr(W_{ni}^{(a)} = 1) = 1$ and hence the functions (3) and (1) coincide.

Definition 1. *Every element* $\hat{\theta}_n^{MRPL}$ *of* Θ *that satisfies* $L_n^{RPL}(\hat{\theta}_n^{MRPL}) \geq L_n^{RPL}(\theta)$ *for all* θ *in some compact subset of* Θ *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 θ_0. The parameter π_n controls the computational cost. For clarity, suppose that $\mathcal{A}$ is the set of all pairs. Since there are n observations and $d(d-1)/2$ pairs, the expected number of terms in the randomized pairwise log-likelihood function is $nd(d-1)\pi_n/2$. For instance, if $\pi_n = 1/6$, $d = 3$ and $n = 10000$ then one needs to sum 5000 terms on average to compute the randomized pairwise likelihood, and 30000 to compute the standard pairwise likelihood method.

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

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

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Assume that π_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 Λ is a compact subset of Θ and θ_0 is an interior point of Λ, then $\sqrt{n}\left(\hat{\theta}_n^{MRPL} - \theta_0\right)$ converges in distribution to a Gaussian random vector with mean zero and variance-covariance matrix $S^{-1}CS^{-1} + \pi^{-1}S^{-1}$.*

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

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

Remark 3. *If, in formula (3), the weights $W_{ni}^{(a)}$ were not chosen randomly but according to availability of the data—that is, zero if the data are missing and one otherwise—then it was shown in Molenberghs et al. (2011) that, under the "Missing Completely At Random" (aka*

MCAR) framework (Rubin, 1976), the gradient of the expectation of (3) would be equal to zero and hence consistent inference should result under appropriate conditions.

3.2 Statistical versus computational efficiency

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

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

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

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

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

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

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

3.3 Consequences of letting the sampling parameter vanish

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

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Let* $\hat{\theta}_n^{MRPL}$ *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 3. *Suppose that Assumptions 1 and 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* Λ *is a compact subset of* Θ *and* θ_0 *is an interior point of* Λ*. If* $\pi_n \to 0$ *such that, for all* $\kappa > 0$ *and all* $a \in \mathcal{A}$*,*

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

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

Subject to conditions on n and π_n (discussed in more detail next), Theorem 3 predicts that a reasonable approximation of the MRPLE's variance is given by $S^{-1}/(n\pi_n)$. In comparison with the previous formula $S^{-1}CS^{-1}/n + S^{-1}/(n\pi)$, the term $S^{-1}CS^{-1}/n$, which is the only term that involves the correlations between the scores, has disappeared. This can be exploited to build approximate confidence intervals without the need to estimate the onerous matrix C.

Let us come back to the conditions on n and π_n. First, notice that Theorems 2 and 3 are consistent with each other, because the condition (4) implies $n\pi_n \to \infty$. To benefit from the approximation suggested by Theorem 3, the sampling parameter π_n must be small, but not too small; the meaning of "not too small" is captured by the condition (4), which in particular implies that $n\pi_n$ must be large enough. The quantity $n\pi_n$ may be regarded as the "effective sample size". The choice of π_n is discussed in Section 6.2.

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

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

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

4 Standard Gaussian model examples

The standard Gaussian model facilitates our understanding of the randomized pairwise likelihood method because explicit calculations are feasible. The density of this model at $x \in \mathbf{R}^d$ is proportional to $f(x;\theta) \propto |\Sigma_\theta|^{-1/2} \exp\left(-\frac{1}{2}x^\top \Sigma_\theta^{-1} x\right)$, where $\theta = (\theta_1, \ldots, \theta_q) \in \Theta \subset (-1,1)^q$, $1 \leq q \leq d(d-1)/2$, is the parameter vector determining the correlation matrix Σ_θ, so that each of its entries is a function of θ, denoted by $v_a(\theta)$ (as in Proposition 1). In other words, $f_a(\cdot,\cdot;\theta)$ depends on θ only through $v_a(\theta)$. The case $q = 1$ with $v_a(\theta) = \theta_1$ for every $a \in \mathcal{A}$ corresponds to the *exchangeable* correlation structure (Cox and Reid, 2004). The case $q = d(d-1)/2$ with $\theta = (\theta_{(1,2)}, \ldots, \theta_{(d-1,d)})$ and $v_{(j,j')}(\theta) = \theta_{(j,j')}$ for all $j < j'$; $j, j' = 1, \ldots, d$, corresponds to the "free" correlation structure. We assume that Σ_θ is positive definite. For instance, if $q = 1$, the biggest subset of $(-1,1)^d$ on which Σ_θ is positive definite is $\Theta = (-1/(d-1), 1)$.

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 α because they are maximizers of the randomized pairwise likelihood, which depends on π_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 standard Gaussian model with an exchangeable correlation structure then Assumptions 1 and 2 hold.*

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

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

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

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

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

$\theta_0^2)+6\theta_0^2(1+\theta_0^2)-2\theta_0^2(1+\theta_0^2)(4+2\theta_0)+\theta_0(1+\theta_0^2)^2(1+2\theta_0)$ 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 Comparison to the subset selection method

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

Next, the two methods are compared for the exchangeable standard Gaussian model. For simplicity, put $L_{ij,kl}=\mathrm{E}\,\dot\ell_{ij}\dot\ell_{kl}$, $L_{ij}=\mathrm{E}\,\dot\ell_{ij}^2$, $|\mathcal{B}|=B\le A=|\mathcal{A}|$. To make the methods comparable, set $\pi=B/A$, so that, on average, both the randomized pairwise likelihood and the pairwise likelihood based on the set of pairs $\mathcal{B}$ have the same computational cost, measured by the number of times the density of a bivariate Gaussian distribution is evaluated. As in Section 3.2, let $V(\pi)=V(B/A)=L_{12}^{-2}A^{-2}[\alpha(L_{12,13}-L_{12,34})+A(A-1)L_{12,34}]+L_{12}^{-1}B^{-1}$ be the asymptotic variance of the MRPLE, where here $\alpha=6\binom{d}{3}$ is the number of couples of pairs that share an index, among all possible couples of pairs. When $A\to\infty$, note that $V(B/A)\sim L_{12}^{-2}L_{12,34}+L_{12}^{-1}B^{-1}$.

The performance of the subset selection method depends on the number of couples of pairs that share an index among all couples of pairs in $\mathcal{B}$. Denote this number by β. Denote by $W_\beta(B)=W_\beta(A\pi)$ the asymptotic variance of the estimator obtained from the subset selection method. According to Theorem 1, for $B \geq 2$, we have

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

It can be checked numerically from the formulas at the end of Section 4.1 that $L_{12,13} - L_{12,34} \geq 0$ and hence the best possible subset selection method is obtained when $\beta = 0$, leading to $W_0(B) = L_{12}^{-2} L_{12,34} + (1 - L_{12}^{-1} L_{12,34}) L_{12}^{-1} B^{-1}$. The worst possible subset selection method is obtained when $\beta = B(B-1)$, leading to $W_{B(B-1)}(B) = L_{12}^{-2} L_{12,13} + (1 - L_{12}^{-1} L_{12,13}) L_{12}^{-1} B^{-1}$. Note that setting $\beta = 0$ or $\beta = B(B-1)$ may or may not be possible, depending on the choice of $\mathcal{B}$.

Figure 1 displays the values of $V(B/A)$ and $W_\beta(B)$, $\beta = 0, B(B-1)$, for $B = 2, \ldots, 30$ in the case $d = 50$. The curve for $V(B/A)$ is contained in the strip delimited by the curves $W_0(B)$ (bottom, best possible subset selection method) and $W_{B(B-1)}$ (top, worst possible subset selection method). The curve for $V(B/A)$ closely follows that for the best possible subset selection method.

4.3 An excursion to the infinite dimensional case

Recall that the correlation matrix Σ_θ is determined by the parameter vector $\theta = (\theta_1, \ldots, \theta_q)$. In this section, the dimension increases as the sample size goes to infinity, that is, we let $d = d_n$ go to infinity as n goes to infinity. The number of parameters q is arbitrary but fixed. For every $a \in \mathcal{A}$, we assume that $v_a(\theta) = \theta_i$ for some $i = 1, \ldots, q$, or $v_a(\theta) = 0$. For each $i = 1, \ldots, q$, let N_i denote the number of entries in the upper triangular part of Σ_θ that are equal to θ_i. Denote by $N_{\min}$ and $N_{\max}$ the minimum and maximum of the numbers

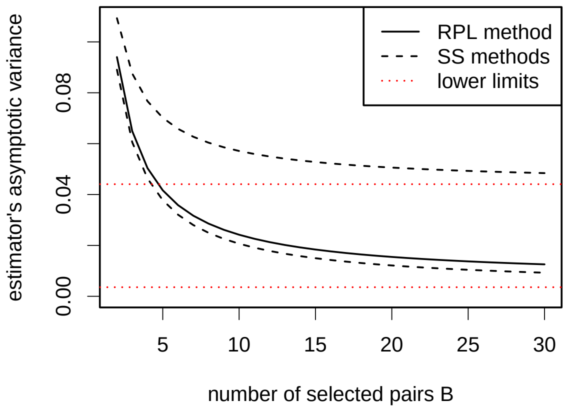

Figure 1: Asymptotic variances $V(\pi) = V(B/A)$, $W_0(B)$ and $W_{B(B-1)}(B)$ for the randomized pairwise likelihood (RPL) method and the subset selection (SS) methods in the cases $\beta = 0$ and $\beta = B(B-1)$, respectively, in the exchangeable standard Gaussian model with $d = 50$ and $\theta_0 = 0.7$. The lower limits correspond to the limit values of $W_\beta(B)$, as B tends to infinity, for the cases $\beta = B(B-1)$ (above) and $\beta = 0$ (below). The latter limit is also that of $V(B/A)$ as A and B tend to infinity.

N_i, respectively. Each N_i is allowed to grow to infinity. In other words, the correlation matrix is partitioned into $q+1$ homogeneous blocks allowed to grow to infinity. Let $\mathcal{A}^+$ denote the set of all those $a \in \mathcal{A}$ that satisfy $v_a(\theta) = \theta_i$ for some $i = 1, \dots, q$.

Proposition 6. *If $d_n \to \infty$, $\pi_n \to 0$, $N_{max}/(n\pi_n N_{min}^2) \to 0$, $N_{max}^2/(nN_{min}^2) \to 0$ and $|\mathcal{A}^+| = O(N_{min})$, then $\|\hat{\theta}_n^{MRPL} - \theta_0\| \xrightarrow{P} 0$ as $n \to \infty$.*

Proposition 6 says that if the largest block is not too large with respect to the smallest block, then consistency holds even if the dimension goes to infinity. For instance, if all the blocks have the same size, we have that $N_{\min} = N_{\max}$ and $|\mathcal{A}^+|$ are all of order d_n^2, and the conditions become $1/(n\pi_n d_n^2) \to 0$. Here, the dimension is a blessing, not a curse. In this example, considering more and more variables means adding more and more pairs to the pairwise likelihood, and each of those pairs brings some information about θ_0.

5 Copula models for multivariate count data

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

An interesting approach uses copulas (Nelsen, 2006), allowing one to easily control the model marginals (Zhao and Joe, 2005; Nikoloulopoulos, 2013). However, inference raises computational problems, which may be mitigated by using the randomized pairwise likelihood. Let $m_1, \dots, m_{d+1}$ be natural integers with sum equal to q. Let $\{F_i(\cdot; \mu_i), \, \mu_i \in \Theta_i \subset \mathbf{R}^{m_i}\}$, $i = 1, \dots, d$, be families of univariate distribution functions. For every $\mu_i \in \Theta_i$, the distribution function $F_i(\cdot; \mu_i)$ is also denoted by F_{μ_i}. Let $\{C(\bullet; \rho), \, \rho \in \Theta^{\text{cop}} \subset \mathbf{R}^{m_{d+1}}\}$ be a family of copulas defined on $[0,1]^d$. For each $\theta := (\mu_1, \dots, \mu_d, \rho) \in \Theta := \Theta_1 \times \cdots \times \Theta_d \times \Theta^{\text{cop}}$, the function defined by $F(x_1, \dots, x_d; \theta) = C(F_{\mu_1}(x_1), \dots, F_{\mu_d}(x_d); \rho)$, $x_1, \dots, x_d \in \mathbf{R}$, is a well-defined distribution function on $\mathbf{R}^d$ with marginals $F_{\mu_1}, \dots, F_{\mu_d}$. It is easy to show that if the univariate distribution function families and the copula family are identifiable then the resulting family of multivariate distribution functions is identifiable, too. From Sklar's theorem (Sklar, 1959), the copula is unique if the marginal distribution functions are continuous. In the discrete case, the copula is not unique in general but it still permits the construction of valid parametric statistical models. The difference with the continuous case is that the copula parameter alone does not characterize the dependence between the random variables at play (Genest and Nešlehová, 2007). Nevertheless, well-defined copula-based models are well-defined statistical models, to which the usual inference methods apply.

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

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

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

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

Proposition 7. *Suppose that the univariate distribution function families* $\{F_i(\cdot;\mu_i),\ \mu_i \in \Theta_i \subset \mathbf{R}^{m_i}\}$, $i = 1,\dots,d$, *are identifiable. If, for every* $a \in \mathcal{A}$, *there is a function* w_a *on* Θ^{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 7 are verifiable for at least some classes of models. For models based on the Gaussian copula, that is,

$$C(u_1,\dots,u_d;\rho) = \Phi_d(\Phi_1^{-1}(u_1),\dots,\Phi_1^{-1}(u_d);R(\rho)), \quad u_1,\dots,u_d \in (0,1), \tag{5}$$

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

Example 1. *Let* C *be the Gaussian copula (5) with correlation matrix* $R(\rho)$, *where* $R(\rho)$ *has* 1*s on its diagonal and* ρ *elsewhere,* $\rho \in (-1/(d-1),1) =: \Theta^{cop}$. *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 (5) with correlation matrix* $R(\rho)$*, where* $R(\rho)$ *has* 1*s on its diagonal and* ρ_{ij} *at its* i*th row and* j*th column, with* $\rho = (\rho_{12}, \ldots, \rho_{d-1,d}) \in (-1,1)^{d(d-1)/2}$ *such that* $R(\rho)$ *is nonnegative definite. Let* Θ^{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 Numerical illustrations

In both Section 6.1 and Section 6.2, 500 synthetic datasets of size n and dimension d are generated from a Gaussian copula and unit Poisson marginals. We used the `copula` R package (Yan, 2007). In Section 6.1, $n \in \{100, 500, 1000\}$ and $d = 30$. In Section 6.2, $n \in \{500, 1000, 5000\}$, $d = 3$ or $d = 10$. A complementary set of simulations for the Gaussian exchangeable case can be found in Section A in the Supplementary Material.

6.1 Effect of the sampling parameter on the estimator's perfomance and computational gains

We investigated the trade-off between efficiency and computational time for the randomized pairwise likelihood approach with $d = 30$. We considered two different cases: (1) a blockwise exchangeable correlation structure (for three blocks of dimension 10), corresponding to a total of 6 distinct copula parameters; and (2) a factorized correlation structure, where the element at the ith row and jth column of the copula correlation matrix $R(\rho)$ is given by $R(\rho)_{ij} = \rho_i \rho_j$, $\rho = (\rho_1, \ldots, \rho_d)$, corresponding to a total of 30 distinct copula parameters. For the blockwise exchangeable case, true copula parameters were set to $\rho = (0.75, 0.5, 0.25, 0.75, 0.5, 0.75)$ in lexicographical order. For the one-factor case, ρ was set to 30 equally spaced values between 0.1 and 0.9. Mean parameters were initialized using marginal means. Copula parameters for the blockwise exchangeable and one-factor simulations were respectively initialized using blockwise-averaged Pearson correlations or by optimizing the factorized correlations $\rho_i \rho_j$ by minimizing the Euclidean distance to the Pearson correlation matrix. The randomized pairwise likelihood was applied with $\pi \in \{0.1, 0.3, 0.5\}$.

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

6.2 Coverage for the confidence intervals

We next sought to evaluate the asymptotic coverage of the confidence intervals constructed for the MRPLE for multivariate count data. For the $d = 3$ case, unstructured copula parameters were given by $\rho_{12} = 0.3$, $\rho_{13} = 0.2$ and $\rho_{23} = 0$. For the $d = 10$ case, we

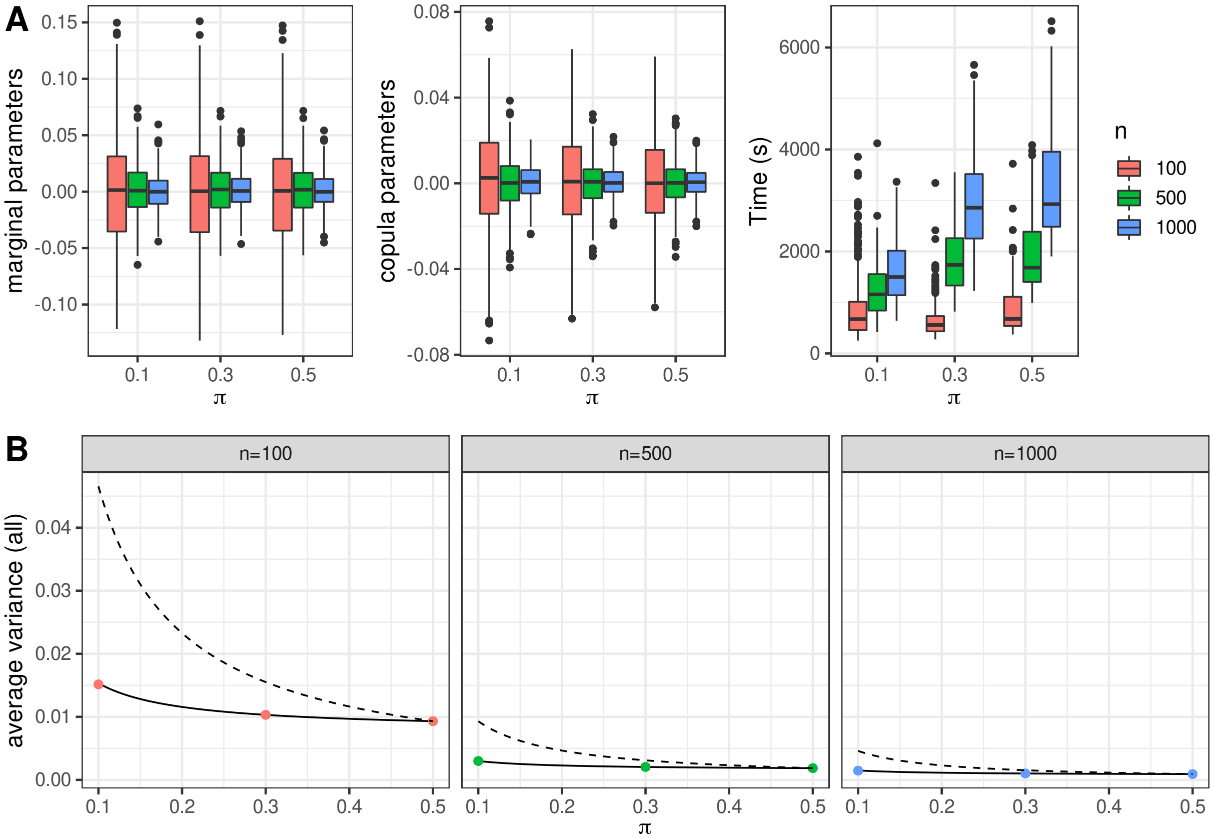

Figure 2: Performance of the randomized pairwise likelihood in the one-factor multivariate Poisson simulations with $d = 30$ over 500 replications. (A) boxplot of the averaged centered estimates for the marginal parameters (left) and the copula parameters (middle), and the corresponding computational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

used a factorized correlation structure with values set as in the previous section. To apply the randomized pairwise likelihood estimation procedure, we first initialized parameter values using the marginal means of each variable and the Pearson correlation of each pair of variables. Finally, we maximized the randomized pairwise likelihood with sampling parameter $\pi \in [0.01, 0.90]$. Confidence intervals of level 95% based on the approximation $S^{-1}/(n\pi)$ suggested by Theorem 3 were calculated for each parameter and each dataset. (Estimates of S were obtained as in Appendix D.) Coverage of the confidence intervals was computed as the proportion of replications for which the true parameter values were within the 95% confidence intervals. Results, corresponding to the mean coverage for the marginal and copula parameters, are presented in Figure 3.

In Figures 3A, C and D, the coverage gets closer to its 95% target as π decreases, agreeing with asymptotic theory. Then, for $n = 500$ and $n = 1000$, the coverage drops at $\pi = 0.01$. For such a small π, the product $n\pi$, equal to 5 and 10 respectively, is too small for any inference to be reliable. For $n = 5000$, corresponding to $n\pi = 50$, there is no drop at $\pi = 0.01$.

In Figure 3B, a different pattern appears. Coverage performance is best for moderate to large values of π, seemingly contradicting the theory saying that for $S^{-1}/(n\pi)$ to be a good approximation of the MRPLE's variance $S^{-1}CS^{-1}/n + S^{-1}/(n\pi)$, π must be small enough. A possible explanation for this seeming contradiction is that the term $S^{-1}CS^{-1}$ may be negligible with respect to S^{-1}. In this case $S^{-1}/(n\pi)$ is always a good approximation, whatever the value of π. The coverage performance degrades as π approaches zero, plausibly because "the effective sample size" $n\pi$ gets too small.

In all panels in Figure 3, which covers a variety of settings, we observe satisfactory results for π of about 0.1 and an effective sample size of about 100–200. Although the choice of π necessarily depends on the model under investigation, a pragmatic recommendation would be to choose (i) a value of π smaller than 0.1, such that (ii) $n\pi$ is greater than 100. In the case of very large sample sizes n, one could either satisfy (ii) with π much smaller than 0.1,

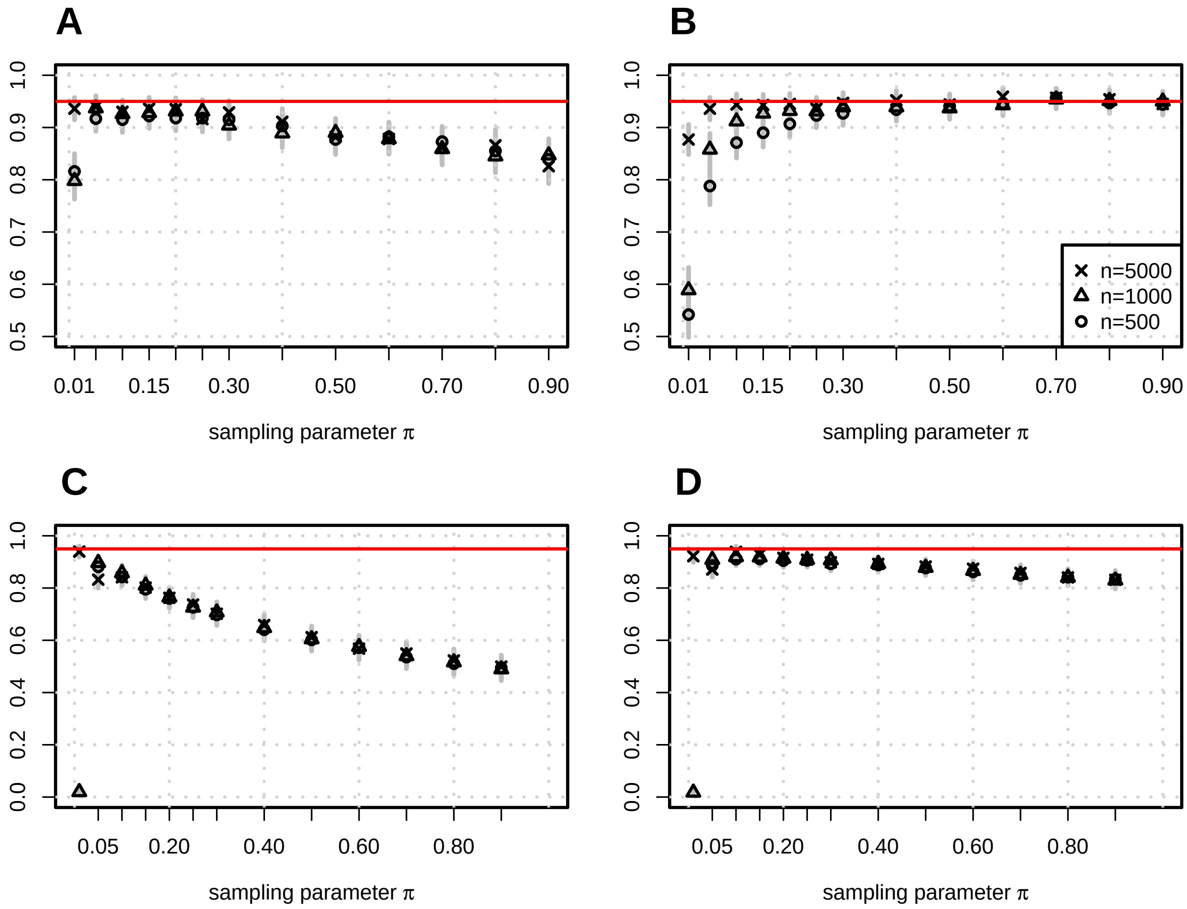

Figure 3: Coverage results for $d = 3$ marginal (A) and copula (B) parameters, and coverage results for $d = 10$ marginal (C) and copula (D) parameters.

or satisfy (i) with an effective sample size much larger than 100. Which alternative is to be preferred depends on the objective of the data analysis at hand.

7 Application on transcriptomic data

We illustrate the application of the randomized pairwise likelihood procedure on multivariate count data from a study on the remodeling of the transcriptome over the life cycle of *Varroa destructor*, a parastic mite that represents a significant threat to the western honeybee. Full details about the experimental design and pre-processing of RNA sequencing (RNA-seq) data may be found in Mondet et al. (2018). Our goal is to evaluate overall transcriptome-wide correlations among different *Varroa* life stages from a single colony (R204) based on RNA-seq read counts for n=22,372 contigs in d=10 life cycle groups. In RNA-seq data, counts of expression are strongly positively associated with both the sequencing effort of each RNA sample (Robinson and Oshlack, 2010) and gene length; an offset accounting for these two factors are included in a Poisson generalized linear model (GLM) defined for the marginal distributions of each sample. To model the dependencies among life stages, these Poisson marginals were coupled with an unstructured Gaussian copula. Poisson GLM intercepts and Gaussian copula correlations were respectively initialized using marginal estimates and pairwise Pearson correlations, and the Nelder-Mead algorithm was used for optimization.

The randomized pairwise likelihood method was applied with $\pi = 0.01$, corresponding to $n\pi = 224$ and standard errors of order less than 10^{-4}, which was found to be sufficiently small with respect to the parameter estimate magnitudes. (The matrix S was estimated from the data as in Appendix D.) Standard errors for all the parameter estimates are given in Table S2. The value of π was chosen in accordance with the guidelines given in Section 6.2. Similar results, not included here, were observed for both $\pi = 0.05$ and $\pi = 1$. A significant gain in computational time was observed: the maximization of the randomized pairwise likelihood with $\pi = 0.01$ took 25 minutes, compared with 8.5 hours

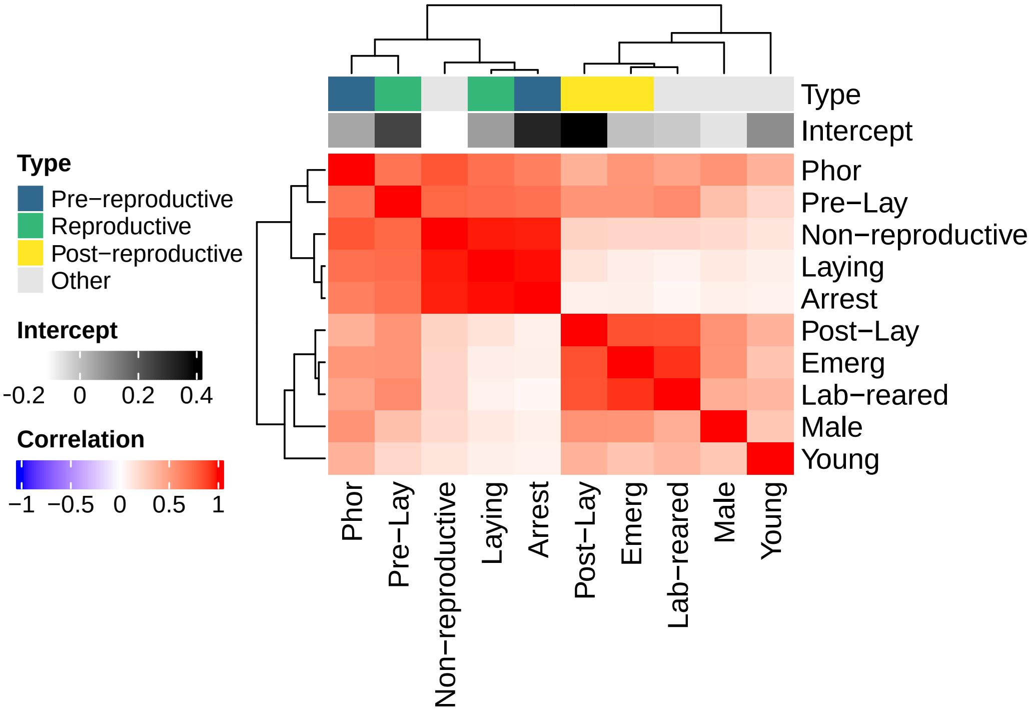

Figure 4: Clustered heatmap of the estimated copula parameters and per-sample intercepts (log-scale) for the *Varroa* life cycle transcriptome data, using the randomized pairwise likelihood ($\pi = 0.01$) approach. Categorizations of life cycle groups according to reproductive status are included as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

for the standard pairwise likelihood.

Figure 4 provides a visualization of the estimated copula parameters between life cycle groups and marginal Poisson GLM intercepts based on the full dataset. We note a strong separation between the pre-reproductive/reproductive (phoretic, arresting, pre-laying, laying) versus post-reproductive (post-laying, emerging) phases. Non-reproductive females are clustered with reproductive females, supporting the hypothesis that mechanisms underlying reproductive failure occur before oogenesis in *Varroa* (Mondet et al., 2018). Lab-reared mites clustered with post-reproductive colony-collected females, suggesting that laboratory conditions do not provoke significant changes in the *Varroa* transcriptome. Two stages in particular exhibit distinct transcriptomic profiles as compared to the others: males (for which the largest estimated copula correlation of 0.56 is with post-lay females), and young mites, which are known to be characterized by a markedly immature physiology. Finally, the intercepts estimated for each marginal Poisson GLM provide intuition about the global over- or under-expression observed in each sample; the transcriptome appears to be most up-regulated in the transitions to (arrest and pre-lay) and from (post-lay) the reprodutive stages.

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

8 Conclusions

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

There is a downside to randomization, however. Since less data is used, the estimator's asymptotic variance increases. In some contexts the standard errors may still be small enough (as in Section 7), but in others they may not. In the latter case, an avenue for future research consists of optimizing several randomized pairwise likelihood in parallel and averaging the results. We expect the final estimator to be more efficient, see also Hector and Song (2020). Also, it may well be that, in some particular models, increasing the number of dimensions actually reduces the variance, as it was shown in our preliminary asymptotic results as $d \to \infty$.

In the future, beyond the aforementioned points one could consider other sampling schemes to exploit known information about the data (such as temporal or spatial autocorrelation) or impose structural or sparsity constraints. For example, one could define a threshold on the number of pairs sampled per observation or impose restrictions on the parameters—for instance, common correlations for some pairs. In addition, one could also consider alternative estimation strategies such as maximization by parts to split the full maximization problem into smaller ones.

SUPPLEMENTARY MATERIAL

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

Acknowledgements

We thank the reviewers and editors for valuable comments that helped to improve this work. We thank Mahendra Mariadassou for detailed comments on a first version of this manuscript and the Migale team of MaIAGE, INRAE and Centre de Traitement de l'Information Génétique (CTIG) of the INRAE Animal Genetics department for providing us with computing clusters. This work was supported by the INRAE DIGIT-BIO metaprogramme grant DINAMIC.

References

Bai, Y., J. Kang, and P. X.-K. Song (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics 70*(3), 661–670.

Bai, Y., P. X.-K. Song, and T. Raghunathan (2012). Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(5), 799–824.

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

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician) 24*(3), 179–195.

Bevilacqua, M., C. Gaetan, J. Mateu, and E. Porcu (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association 107*(497), 268–280.

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

Chatelain, F., S. Lambert-Lacroix, and J.-Y. Tourneret (2009). Pairwise likelihood estimation for multivariate mixed Poisson models generated by gamma intensities. *Statistics and Computing 19*(3), 283–301.

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

Chiquet, J., S. Robin, and M. Mariadassou (2019). Variational inference for sparse network reconstruction from count data. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 1162–1171.

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

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

Eidsvik, J., B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics 23*(2), 295–315.

Ferrari, D., G. Qian, and T. Hunter (2016). Parsimonious and efficient likelihood composition by Gibbs sampling. *Journal of Computational and Graphical Statistics 25*(3), 935–953.

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

Heagerty, P. J. and S. R. Lele (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association 93*(443), 1099–1111.

Hector, E. C. and P. X.-K. Song (2020). A distributed and integrated method of moments for high-dimensional correlated data analysis. *Journal of the American Statistical Association 116*(534), 805–818.

Huang, Z. and D. Ferrari (2021). Fast construction of optimal composite likelihoods. Preprint (arXiv:2106.05219).

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

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

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

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

le Cessie, S. and J. Van Houwelingen (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 43*(1), 95–108.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics 80*(1), 221–239.

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

Molenberghs, G., M. G. Kenward, G. Verbeke, and T. Birhanu (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica 21*(1), 187–206. Publisher: Institute of Statistical Science, Academia Sinica.

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

Nelsen, R. (2006). *An introduction to copulas*. Springer.

Nikoloulopoulos, A. K. (2013). Copula-based models for multivariate discrete response data. In *Copulae in Mathematical and Quantitative Finance*, pp. 231–249. Springer.

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

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

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

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

Rubin, D. B. (1976, December). Inference and missing data. *Biometrika 63*(3), 581–592.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris 8*, 229–231.

Stein, M. L., Z. Chi, and L. J. Welty (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 275–296.

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

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

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological) 50*(2), 297–312.

Wang, X. and Y. Wu (2014). Theoretical properties of composite likelihoods. *Open Journal of Statistics 4*, 188–197.

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

Zhao, Y. and H. Joe (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics 33*(3), 335–356.

Supplement to "A randomized pairwise likelihood method for complex statistical inferences"

Gildas Mazo
MaIAGE, INRAE, Université Paris Saclay,
78350 Jouy-en-Josas, France
and
Dimitris Karlis
Athens University of Economics and Business
and
Andrea Rau
Université Paris-Saclay, INRAE, AgroParisTech, GABI,
78350 Jouy-en-Josas, France;
BioEcoAgro Joint Research Unit, INRAE,
Université de Liège, Université de Lille, Université de Picardie Jules Verne,
80200 Estrées-Mons, France

November 29, 2022

Contents

A Simulations for the exchangeable Gaussian model

A.1 Comparison with the pairwise and the full likelihood methods

We simulate a set of d-dimensional vectors Y_i, $i = 1, \ldots, n$ from an exchangeable multivariate Gaussian distribution with mean vector μ and covariance matrix Σ, where all means μ

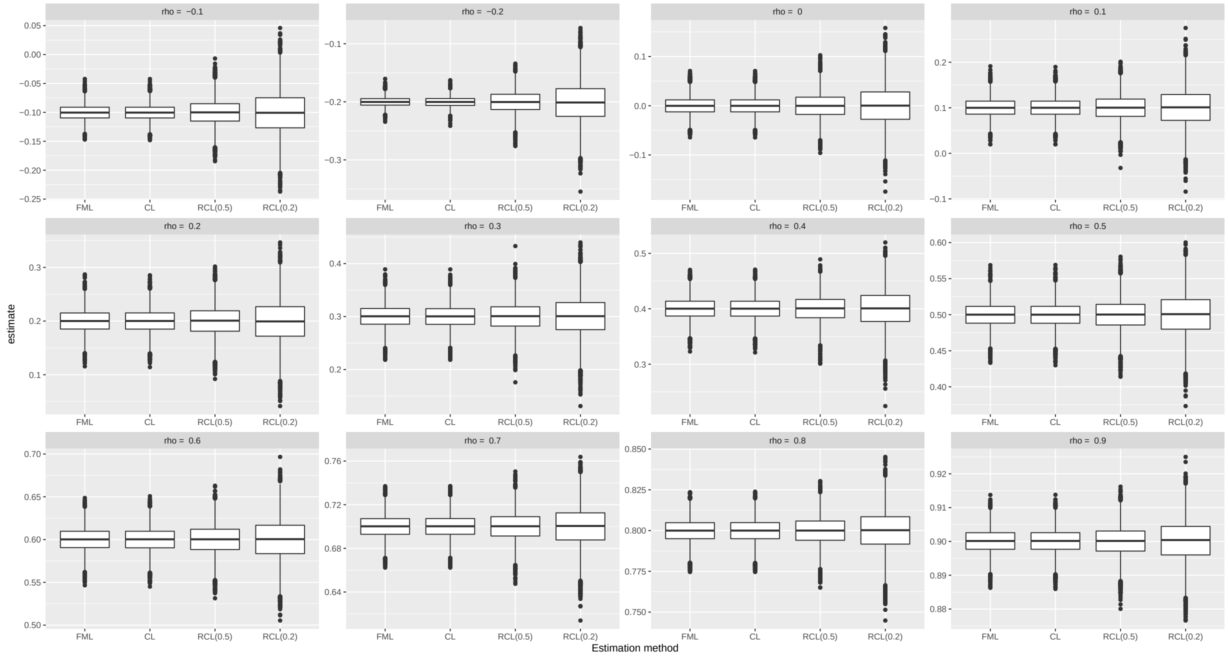

Figure S1: Boxplots of parameter estimates for $n = 500$ across 50,000 simulated datasets using the full maximum likelihood (FML), composite pairwise likelihood (CL), and randomized pairwise composite likelihood (RCL) approaches with $\pi = 0.5$ and 0.2 for $\rho = \{-0.1 \ldots, 0.9\}$.

are considered to be known and set to 0, and all variances and correlations are fixed to 1 and ρ, respectively. In this case, the only parameter to be estimated is thus ρ; in different simulation settings, the true value of ρ 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 considered the sampling parameter values $\pi = 0.5$ and $\pi = 0.2$, and compared the results to those obtained from the full maximum likelihood, and the pairwise likelihood using all pairs of variables and all observations; simulations were repeated 50,000 times. Efficiency was calculated as the ratio of the variance of parameter estimates across simulated datasets in the pairwise

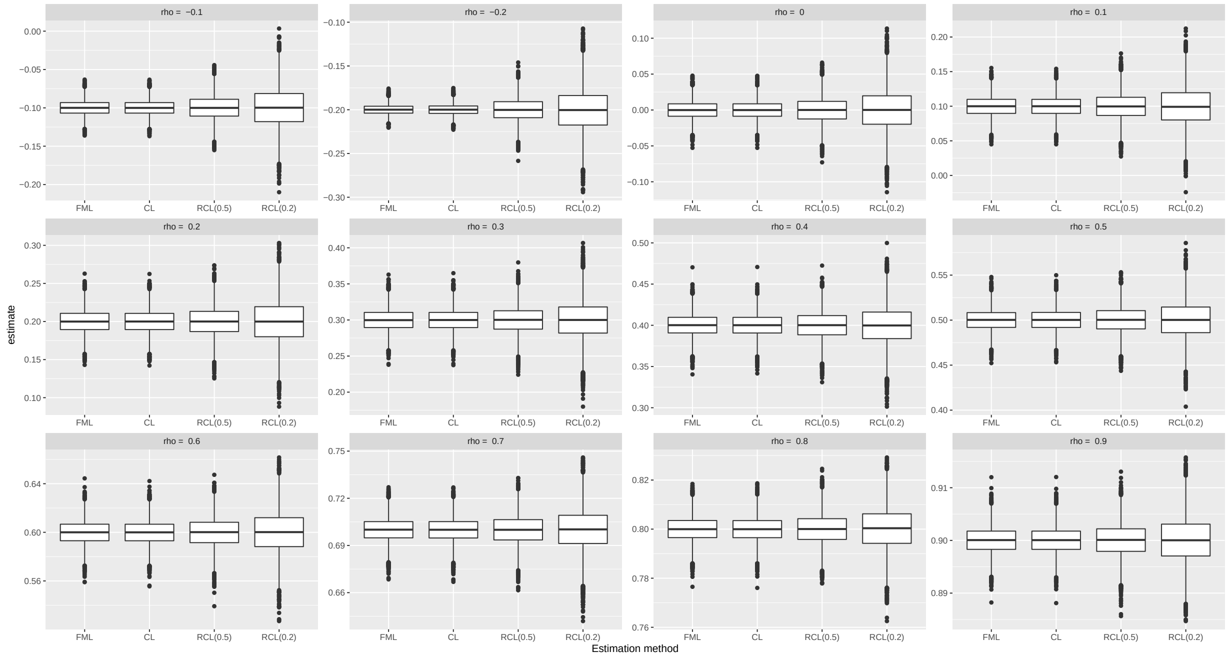

Figure S2: Boxplots of parameter estimates for $n = 1000$ across 50,000 simulated datasets using the full maximum likelihood (FML), composite pairwise likelihood (CL), and randomized pairwise composite likelihood (RCL) approaches with $\pi = 0.5$ and 0.2 for $\rho = \{-0.1 \ldots, 0.9\}$.

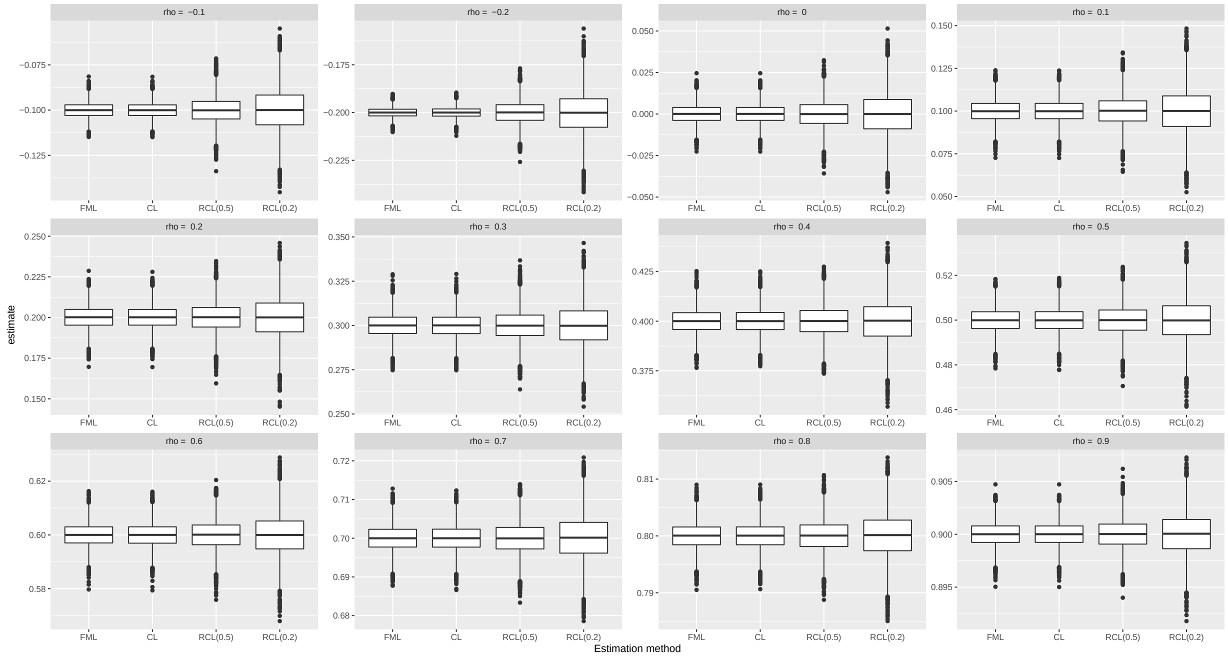

Figure S3: Boxplots of parameter estimates for $n = 5000$ across 50,000 simulated datasets using the full maximum likelihood (FML), composite pairwise likelihood (CL), and randomized pairwise composite likelihood (RCL) approaches with $\pi = 0.5$ and 0.2 for $\rho = \{-0.1\ldots, 0.9\}$.

likelihood and randomized pairwise likelihood methods with respect to the full maximum likelihood approach. For all values of ρ considered, all methods considered successfully recover the true value of ρ, 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 π decreases (see 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 Cox and Reid (2004) 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 π.

A.2 Coverage for the approximate confidence intervals

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 exchangeable Gaussian model model with known means and variances and we estimate the common correlation parameter ρ using randomized pairwise likelihood. Based on Theorem 3 and the derivations of Proposition 4, 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 $V(\hat{\rho}) = 2(1 - \hat{\rho}^2)^4/(d(d-1)(\hat{\rho}^6 - \hat{\rho}^4 - \hat{\rho}^2 + 1))$. One can create an asymptotically $100(1-\alpha)\%$ confidence interval as $\hat{\rho} \pm Z_{1-\alpha/2}\sqrt{V(\hat{\rho})/n\pi}$ 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 π to yield subsample sizes $n\pi$ 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 Figures S4 and S5. 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.

We repeated the simulations in the previous section for datasets with $d \in \{3, 8, 15, 20, 50\}$ to investigate the impact of increasing dimensionality on coverage. Results averaged over 1000 replications are shown in Table S1.

		$d = 3$	$d = 8$	$d = 15$	$d = 20$	$d = 50$
$n = 5000$	$\rho = 0$	0.934	0.953	0.952	0.954	0.946
	$\rho = 0.25$	0.937	0.958	0.949	0.934	0.858
	$\rho = 0.5$	0.932	0.940	0.939	0.934	0.864
	$\rho = 0.75$	0.936	0.945	0.942	0.944	0.909
$n = 10000$	$\rho = 0$	0.930	0.960	0.948	0.951	0.957
	$\rho = 0.25$	0.954	0.941	0.935	0.937	0.864
	$\rho = 0.5$	0.929	0.940	0.929	0.941	0.855
	$\rho = 0.75$	0.937	0.944	0.930	0.951	0.895

Table S1: Average coverage (over 1000 replications) for dimension $d = \{3, 8, 15, 20, 50\}$ for sample sizes $n = 5000$ or 10,000, $\rho \in \{0, 0.25, 0.5, 0.75\}$, and sampling probability $\pi = 0.01$, with $\alpha = 5\%$.

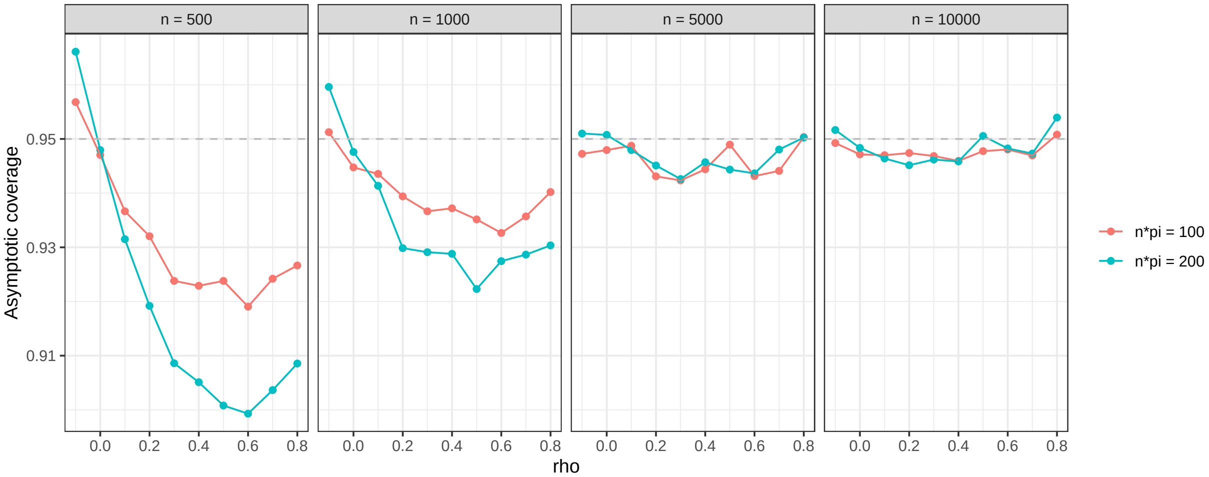

Figure S4: Asymptotic coverage for the exchangeable Gaussian model example, with $\alpha = 5\%$, averaged over 50,000 replications. The values represent the proportion of times the asymptotic interval contains the true value used to simulate the data, with ρ versus asymptotic coverage by sample size n.

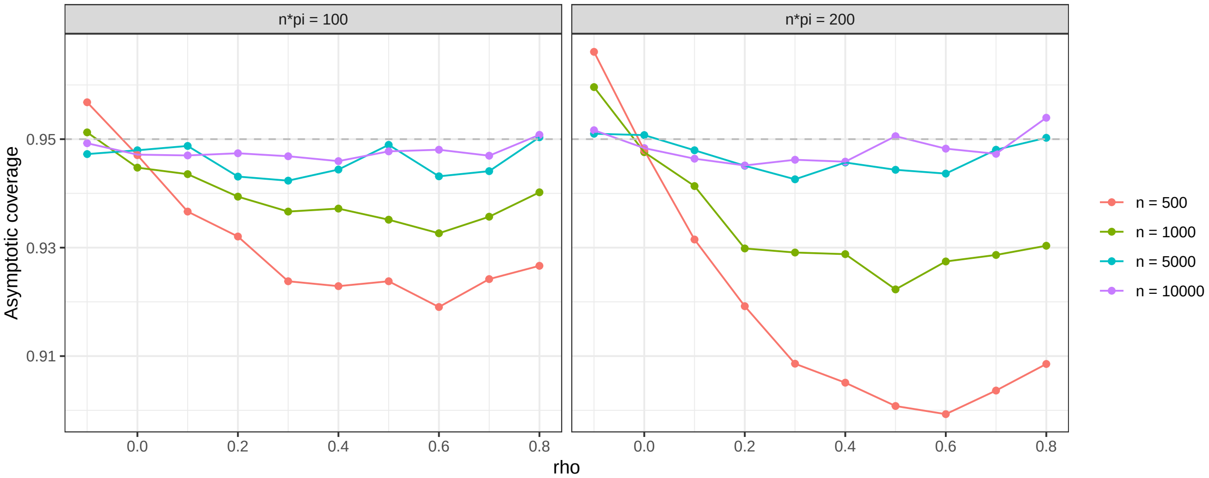

Figure S5: Asymptotic coverage for the exchangeable Gaussian model example, with $\alpha = 5\%$, averaged over 50,000 replications. The values represent the proportion of times the asymptotic interval contains the true value used to simulate the data, for ρ versus asymptotic coverage by subsample size $n\pi$.

B Additional Table

	Phor	Male	Pre-Lay	Arrest	Post-Lay	Young	Emerg	Non-rep	Laying	Lab
Mean	2.3e-07	6.1e-07	2.1e-07	8.7e-07	1.3e-07	8.3e-07	1.6e-07	1.7e-07	6.4e-07	1.8e-07
Phor		1.4e-06	1.2e-06	1.4e-06	1.3e-06	1.4e-06	1.3e-06	2.5e-07	1.1e-06	1.4e-06
Male			1.4e-06	3.3e-06	9.8e-07	1.7e-06	1.0e-06	1.2e-06	2.8e-06	1.1e-06
Pre-Lay				1.9e-06	1.2e-06	1.6e-06	1.1e-06	2.2e-07	8.8e-07	7.7e-07
Arrest					3.0e-06	2.4e-06	1.6e-06	1.4e-04	4.6e-05	3.4e-06
Post-Lay						2.1e-06	2.0e-07	1.9e-06	1.4e-06	2.0e-07
Young							2.6e-06	2.7e-06	2.4e-06	1.3e-06
Emerg								1.6e-06	2.4e-06	1.4e-06
Non-rep									1.1e-04	1.6e-06
Laying										1.9e-06

Table S2: Estimated standard errors for Poisson means (top row) and Gaussian copula parameters (bottom) for the *Varroa* life cycle transcriptome data, using the randomized pairwise likelihood ($\pi = 0.01$) approach.

C Additional Figures

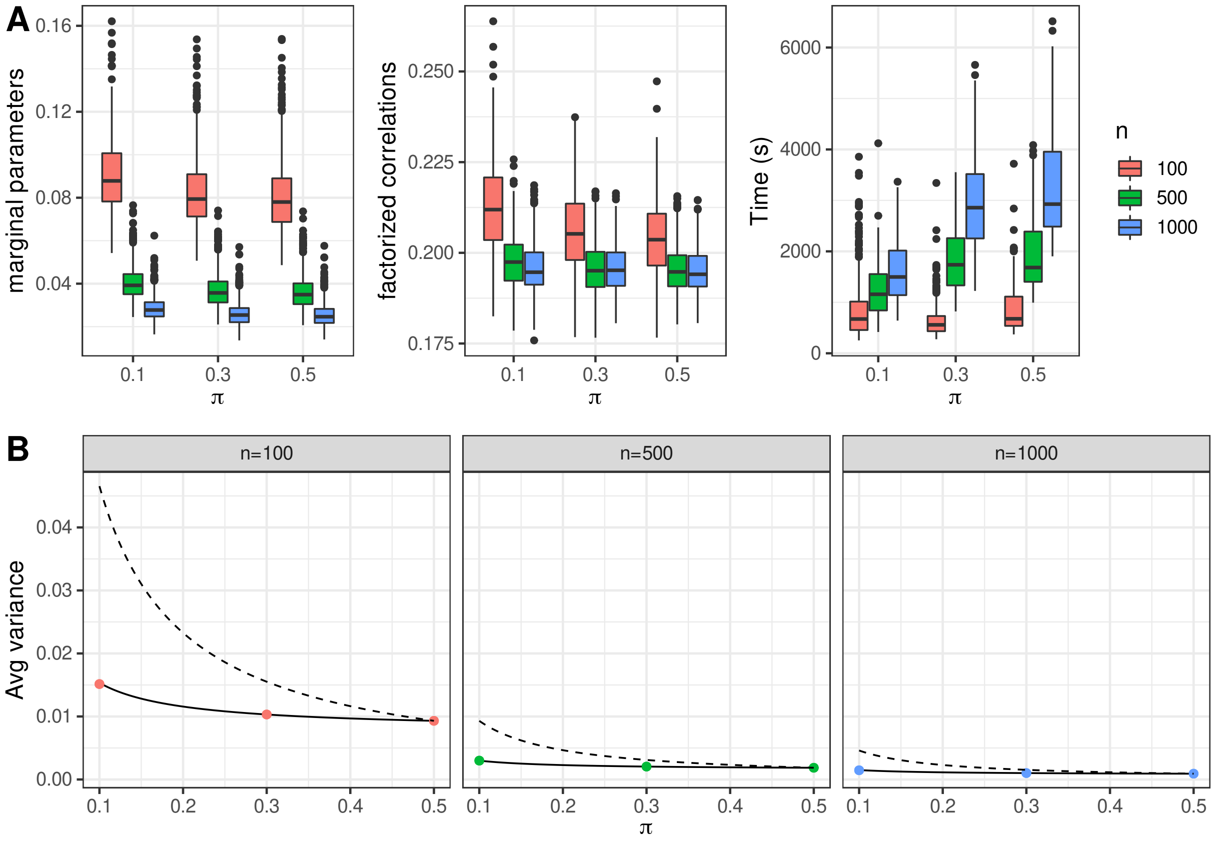

Figure S6: Performance of the randomized pairwise likelihood in the one-factor multivariate Poisson simulations with $d = 30$ over 500 replications. (A) boxplot of the absolute relative errors for the marginal parameters (left) and the factorized correlations (middle), and the corresponding computational times in seconds (right). (B) Averaged variance estimates across parameters (points) for different values of π. The solid line connecting the points corresponds to the theoretical prediction for $\pi = 0.1, 0.3$ knowing the variance at $\pi = 0.5$. The dotted line corresponds to the theoretical prediction under the assumption of a homogeneous inflation factor, knowing the variance at $\pi = 0.5$.

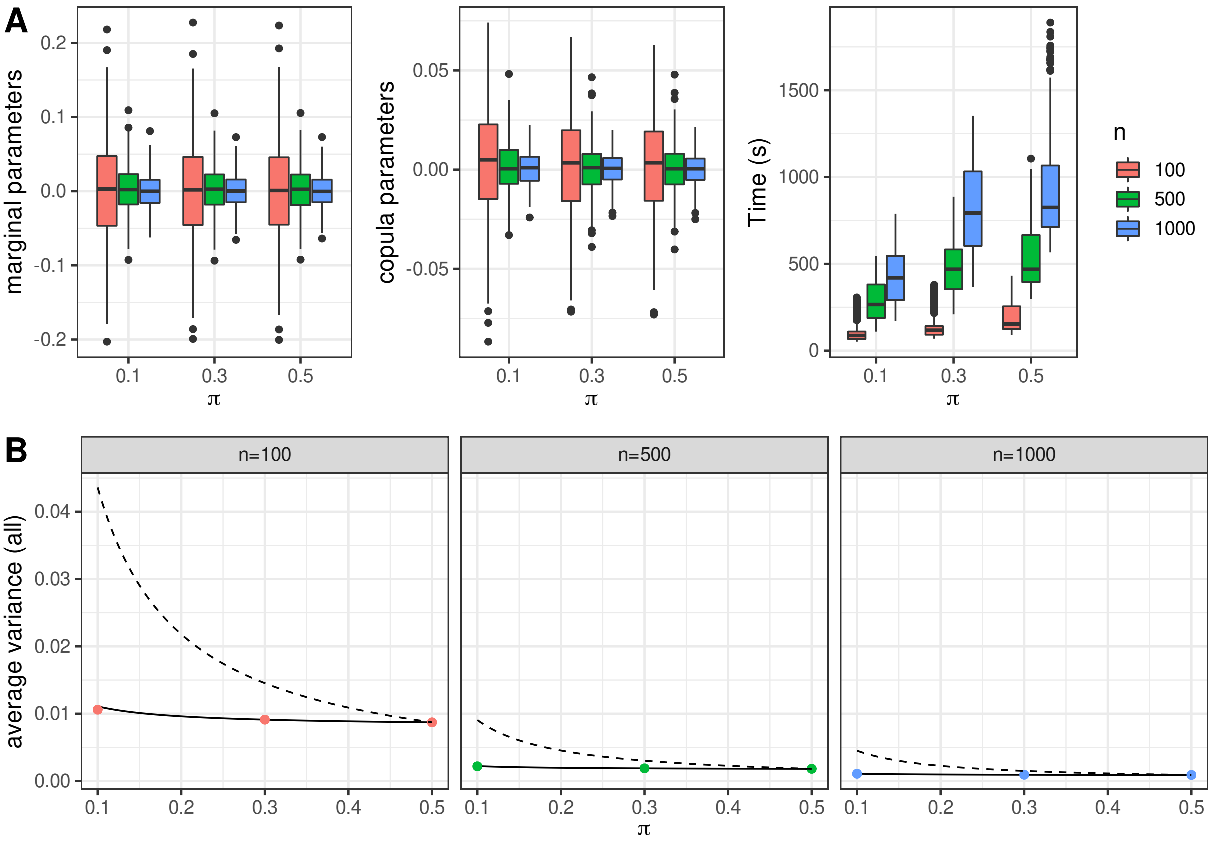

Figure S7: Performance of the randomized pairwise likelihood in the blockwise exchangeable multivariate Poisson simulations with $d = 30$ over 500 replications. (A) boxplot of the averaged centered estimates for the mean parameters (left) and the copula parameters (middle), and the corresponding computational times in seconds (right). (B) Averaged variance estimates across parameters (points) for different values of π. The solid line connecting the points corresponds to the theoretical prediction for $\pi = 0.1, 0.3$ knowing the variance at $\pi = 0.5$. The dotted line corresponds to the theoretical prediction under the assumption of a homogeneous inflation factor, knowing the variance at $\pi = 0.5$.

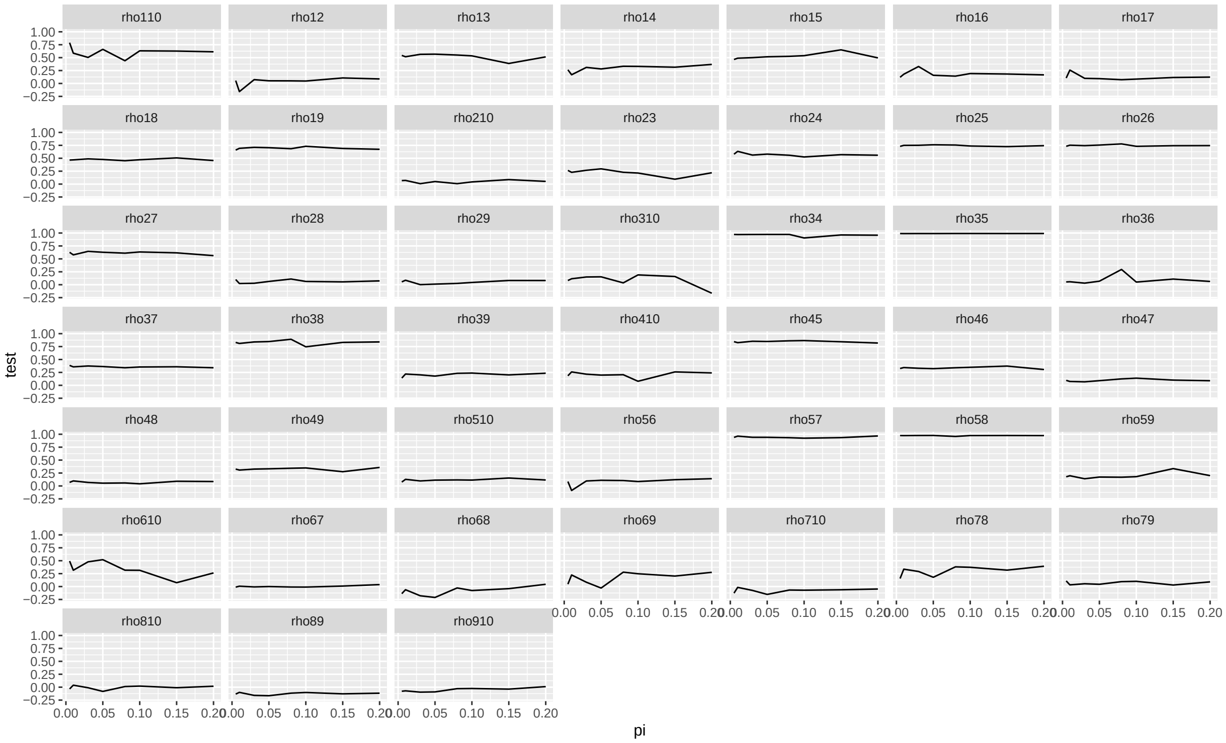

Figure S8: Estimated parameter values of the correlation parameters for the RNA-seq data for varying values of the sampling parameter π.

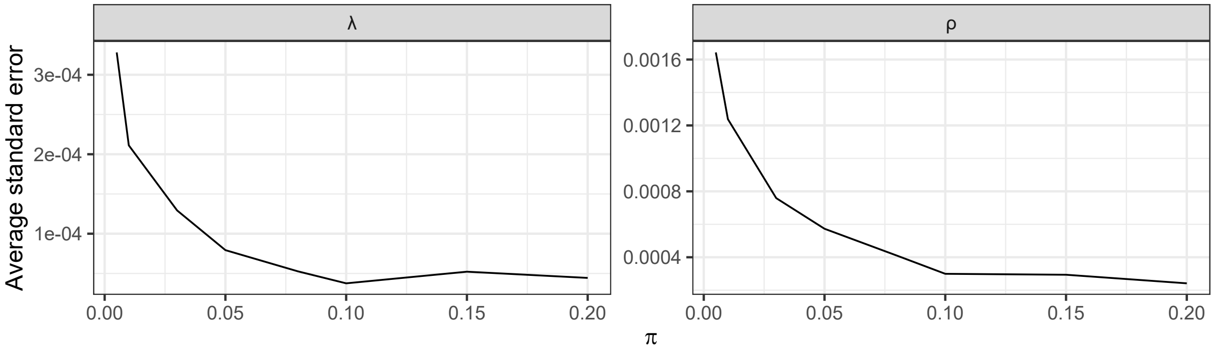

Figure S9: Average standard error of marginal mean (left) and copula (right) parameter estimates from the randomized pairwise likelihood approach for varying values of π.

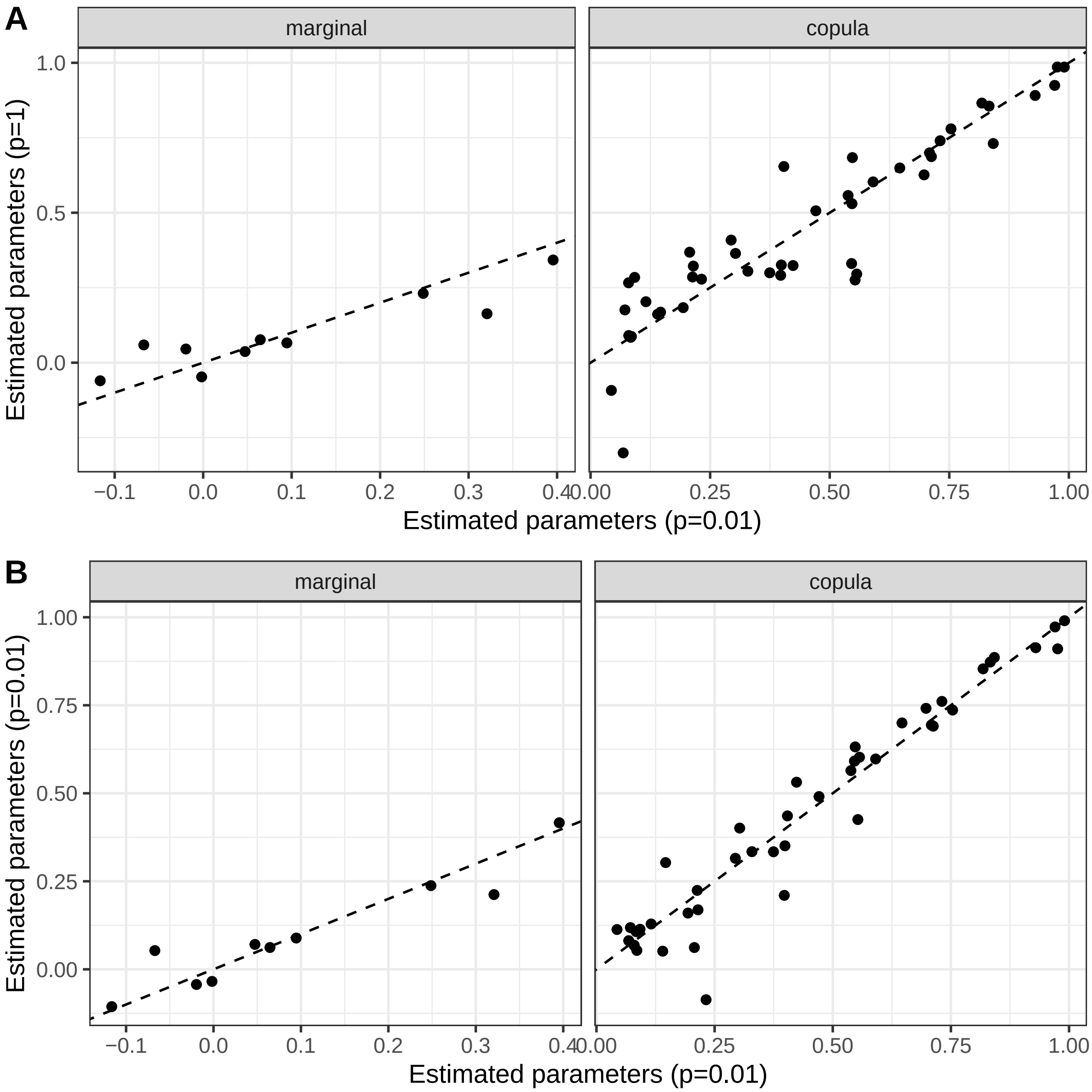

Figure S10: (A) Comparison of estimated marginal (left) and copula (right) parameter values for $\pi = 0.01$ versus $\pi = 1$ for the RNA-seq data. (B) Comparison of estimated marginal (left) and copula (right) parameter values for two independent runs of $\pi = 0.01$ for the RNA-seq data. Dashed lines indicate the identity.

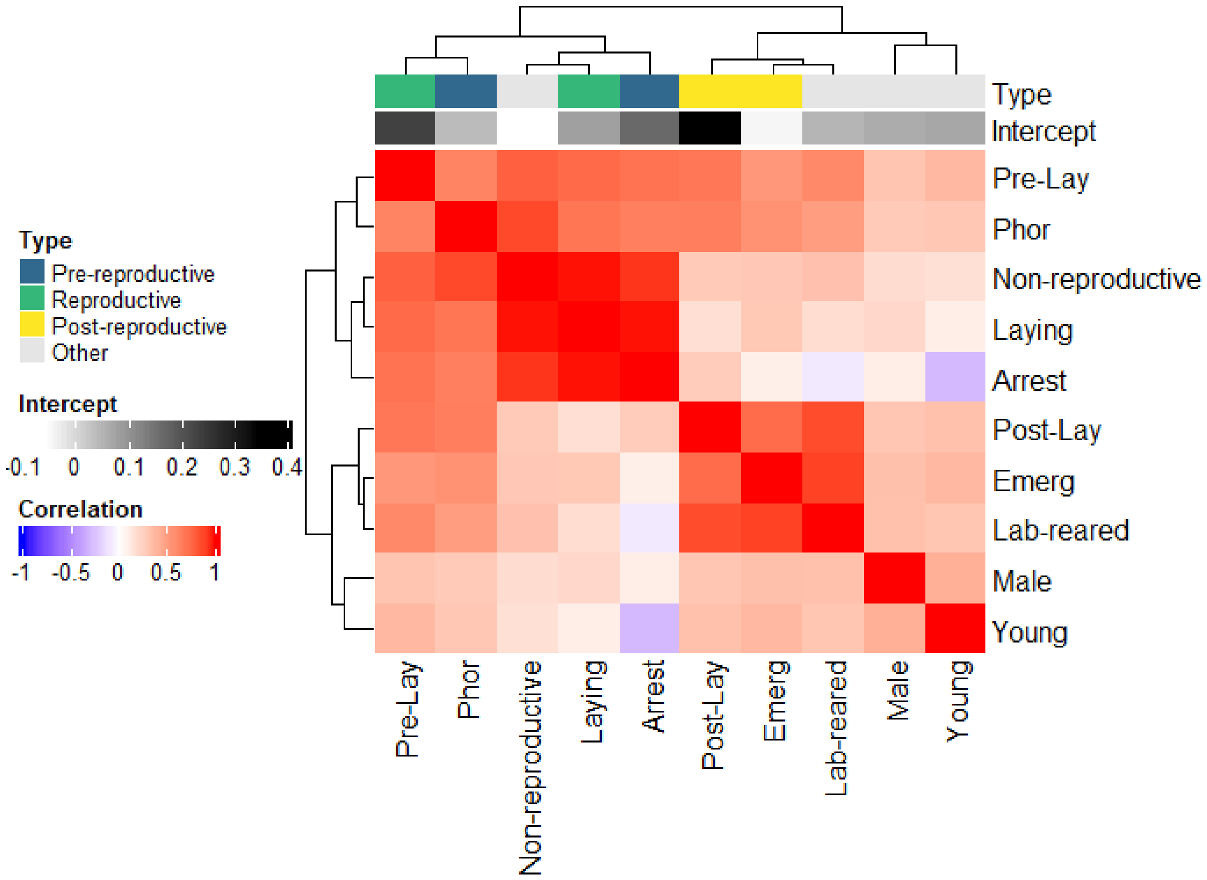

Figure S11: Clustered heatmap of the estimated copula parameters and per-sample intercepts (log-scale) for the *Varroa* life cycle transcriptome data, using the randomized pairwise likelihood ($\pi = 1$). Categorizations of life cycle groups according to reproductive status are included as a column annotation.

D Estimation of the asymptotic variance-covariance matrix S^{-1} in Theorem 3

If $\hat{\theta}_n^{\mathrm{MRPL}}$ is a MRPLE obtained from the sample $X_i = (X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$, and sampling parameter π, then the matrix S defined in Section 2 by

$$S = \sum_{a \in \mathcal{A}} \mathrm{E}\, \dot{\ell}_a(X_1^{(a)}; \theta_0) \dot{\ell}_a(X_1^{(a)}; \theta_0)^\top$$

is estimated by

$$\widehat{S} := \sum_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_a(X_i^{(a)}; \hat{\theta}_n^{\mathrm{MRPL}}) \dot{\ell}_a(X_i^{(a)}; \hat{\theta}_n^{\mathrm{MRPL}})^\top.$$

The matrix $\widehat{S}^{-1}$ of Theorem 3 is obtained by numericall inverting the matrix $\widehat{S}$. Componentwise, an asymptotic confidence interval of level 95% for θ_0 is then given as $\hat{\theta}_n^{\mathrm{MRPL}} \pm 1.96\sqrt{\widehat{S}^{-1}/(n\pi)}$.

E 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 van de Geer (2000); van der Vaart and Wellner (1996); Pollard (1984). If $\mathcal{G}$ is a class of real-valued functions on some Euclidean space equipped with a probability measure P and δ 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 van der Vaart and Wellner (1996); van der Vaart (1998) for a way of addressing this.

Lemma E.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 E.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 E.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*

$$\mathrm{E}\, g_{al}(X_1)^4 \exp\left(- \frac{n\pi_n \kappa}{\sum_{a=1}^{A} g_{al'}(X_1)^2} \right) = o(\pi_n) \tag{S1}$$

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*

$$\sum_{a=1}^{A} \mathrm{E}\, g_a(X_1) g_a(X_1)^\top. \tag{S2}$$

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 E.2 and Lemma E.3.

Proof of Theorem 2

Since $\hat{\theta}_n^{\mathrm{MRPL}}$ is a MRPLE, there is a compact subset $\Lambda \subset \Theta$ that contains θ_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^{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 van der Vaart (1998) 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 Λ. Since the set $\{\theta \in \Lambda : |\theta - \theta_0| \geq \varepsilon\}$ is compact, the supremum of L^{PL} is reached. But this supremum must be less than $L^{\mathrm{PL}}(\theta_0)$, because, by Assumption 2, the point θ_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 E.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 van de Geer (2000) 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 θ by ∇m. Denote the Hessian matrix of L^{PL} at θ_0 by $\nabla^2 L^{\mathrm{PL}}(\theta_0)$. If we can show

$$\sqrt{n\pi_n}(\hat{\theta}^{\mathrm{MRPL}} - \theta_0) = -\left[\nabla^2 L^{\mathrm{PL}}(\theta_0)\right]^{-1} G_{nn}^* \nabla m(\cdot,\cdot,\theta_0) + o_P(1), \tag{S3}$$

then Lemma E.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 π_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 (S3). The map L^{PL} is two times continuously differentiable at θ_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 Λ, 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 θ_0 by Theorem 2. Therefore equation (S3) follows from Theorem 3.2.16 of (van der Vaart and Wellner, 1996, p. 300), which itself is a generalization of an idea of Pollard (1984, 1985), 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

$$\frac{1}{2}\sum_{i_1,i_2}\tilde{h}_{i_1}\tilde{h}_{i_2}G^*_{nn}\nabla^2_{i_1i_2}m(\cdot,\cdot,\theta_0)+\frac{1}{6}\sum_{i_1,i_2,i_3}\tilde{h}_{i_1}\tilde{h}_{i_2}\tilde{h}_{i_3}G^*_{nn}\nabla^3_{i_1i_2i_3}m(\cdot,\cdot,\hat{h})$$
$$=o_P\left(\|\tilde{h}\|+\sqrt{n\pi_n}\|\tilde{h}\|^2+\frac{1}{\sqrt{n\pi_n}}\right), \quad \text{(S4)}$$

where $\hat{h}$ is a point between θ_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 (4), Lemma E.3 implies $G^*_{nn}\nabla^2_{i_1i_2}m(\cdot,\cdot,\theta_0)=O_P(1)$ whether π_n is a constant or $\pi_n\to 0$. Remember that the third derivatives are bounded by the functions Ψ_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 E.3. Thus, in both cases $\pi_n\to 0$ and π_n constant, the left hand side in (S4) is $O_P\left(\|\tilde{h}\|^2\left(1+\|\tilde{h}\|\sqrt{n\pi_n}\right)\right)$. The proof is complete.

F Proofs of the propositions

Proof of Proposition 1

We begin with a lemma.

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

(i) θ_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 θ_0' was 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. $\square$

It is straightforward to show that Lemma F.1 (i) is true. It remains to ensure that Lemma F.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 7

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 $\operatorname{range} v_a = \Theta_i \times \Theta_j \times \operatorname{range} w_a$. 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 $\operatorname{range} v_a$. 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 Θ. 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.

Proof of Proposition 2

Notice that $V(\pi')/V(\pi) \leq \pi/\pi'$ if and only if

$$\frac{\pi S^{-1}CS^{-1}}{\pi S^{-1}CS^{-1} + S^{-1}} \leq \frac{\pi}{\pi'}\left(1 - \frac{S^{-1}}{\pi S^{-1}CS^{-1} + S^{-1}}\right) = \frac{\pi}{\pi'}\left(\frac{\pi S^{-1}CS^{-1}}{\pi S^{-1}CS^{-1} + S^{-1}}\right).$$

Since $S^{-1}CS^{-1} + S^{-1}$ is the asymptotic variance-covariance matrix of Theorem 1 with $\pi = 1$, it must be positive definite, and hence the last inequality is simplified according to the sign of $S^{-1}CS^{-1}$.

Proof of Proposition 3

In this case the functions Φ_a in Theorem 3 are bounded by a constant, say C. Let A be the cardinal of $\mathcal{A}$. The left hand side of (4) 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 4

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 φ 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

$$\mathrm{E}\,\Phi_a(X_1;\theta_0)^2 \leq C(1+\|x\|^2)^2 \tag{S5}$$

for some constant C. (Remember that θ_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 θ 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 5

In view of (S5) and since the the left hand side in (4) is an increasing function of Φ_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 Blumenson (1960). 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 H.2 and hence, by Corollary H.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 6

When d_n goes to infinity, the proof of Theorem 2, which consisted of checking the conditions of van der Vaart's Theorem 5.7 (van der Vaart, 1998, p. 45), is no longer valid. To account for the growth of d_n, one possible avenue is to extend van der Vaart's Theorem 5.7 and its proof. This is done in the next lemma. As in the proof of Theorem 2, $L^{\mathrm{PL}}(\theta)$ stands for $\sum_a L_a(\theta)$; the quantity $L_n^{\mathrm{RPL}}(\theta)$ is the randomized pairwise likelihood evaluated at θ.

Lemma F.2. *If there is a positive sequence* $p_n \to \infty$ *such that*

(i) $\forall \epsilon > 0, \exists \lambda > 0, \forall n \geq 1, \displaystyle\sup_{\|\theta-\theta_0\|\geq\epsilon} \frac{L^{PL}(\theta) - L^{PL}(\theta_0)}{p_n} \leq -\lambda,$

(ii) $\displaystyle\sup_{\theta\in\Theta} \frac{|L_n^{RPL}(\theta) - L^{PL}(\theta)|}{p_n} \xrightarrow{P} 0.$

then $\|\hat{\theta}_n^{MRPL} - \theta_0\| \xrightarrow{P} 0$ *as* $n \to \infty$*.*

The proof of Lemma F.2 is to be found in Section G. For now, let us notice the role played by the sequence p_n. If one chooses a sequence p_n that goes to infinity too fast, then the condition (i) is going to be difficult to satisfy. On the opposite, if one chooses a sequence p_n that goes to infinity too slowly, then it is the condition (ii) that is going to be difficult to satisfy. We therefore must find the correct rate for p_n, if there is one at all.

Checking the first condition

As in Proposition 1, let $v_a(\theta)$ denote the parameters of the marginal density f_a. We say that L^{PL} is *pairwise strongly concave at* θ_0 if there is $\lambda > 0$ such that for all a and all θ, it holds that $L_a(v_a(\theta)) - L_a(v_a(\theta_0)) \leq -\lambda\|v_a(\theta) - v_a(\theta_0)\|_2^2$.

Lemma 1. *If the function* L^{PL} *is pairwise strongly concave at* θ_0 *then*

$$\frac{1}{p_n}(L^{PL}(\theta) - L^{PL}(\theta_0)) \leq -\frac{\lambda \min\{d-1, N_{min}\}}{p_n}\|\theta - \theta_0\|_2^2.$$

Proof. We have

$$
\begin{aligned}
&\frac{1}{p_n}(L^{\mathrm{PL}}(\theta) - L^{\mathrm{PL}}(\theta_0)) \\
&= \frac{1}{p_n}\sum_{j<j'}[L_{(j,j')}(v_{(j,j')}(\theta)) - L_{(j,j')}(v_{(j,j')}(\theta_0))] \\
&\leq -\frac{\lambda}{p_n}\sum_{j<j'}\|v_{(j,j')}(\theta) - v_{(j,j')}(\theta_0)\|_2^2 \\
&= -\frac{\lambda}{p_n}\sum_{j<j'}\|\mu_j - \mu_{0j}\|_2^2 + \|\mu_{j'} - \mu_{0j'}\|_2^2 + |w_{(j,j')}(\theta) - w_{(j,j')}(\theta_0)|^2 \\
&= -\frac{\lambda(d-1)}{p_n}\left(\sum_{j=1}^{d}\|\mu_j - \mu_{0j}\|_2^2\right) - \frac{\lambda}{p_n}\sum_{i=1}^{q}|\theta_i - \theta_{0i}|^2 N_i \\
&\leq -\frac{\lambda(d-1)}{p_n}\left(\sum_{j=1}^{d}\|\mu_j - \mu_{0j}\|_2^2\right) - \frac{\lambda N_{\min}}{p_n}\sum_{i=1}^{q}|\theta_i - \theta_{0i}|^2 \\
&\leq -\frac{\lambda \min\{d-1, N_{\min}\}}{p_n}\|\theta - \theta_0\|_2^2.
\end{aligned}
$$

□

According to Lemma 1, the sequence p_n must satisfy the condition $p_n = O(\min\{d-1, N_{\min}\})$. Assuming the marginals are known and hence ignoring the marginal parameters, the condition is $p_n = O(N_{\min})$.

It remains to show that L^{PL} is pairwise strongly concave. But this is easily seen:

momentarily denoting $v_a(\theta)$ by θ_a and $v_a(\theta_0)$ by θ_{0a} for every pair a, it holds

$$\begin{aligned} L_a(\theta_a) - L_a(\theta_{0a}) &= \frac{1}{2}\log\left(\frac{1-\theta_{0a}^2}{1-\theta_a^2}\right) - \frac{\theta_a(\theta_a-\theta_{0a})}{1-\theta_a^2} \\ &\leq \frac{\theta_a^2-\theta_{0a}^2}{2(1-\theta_{0a}^2)} - \frac{\theta_a(\theta_a-\theta_{0a})}{1-\theta_a^2} \\ &= -\frac{(\theta_a-\theta_{0a})^2}{2}F(\theta_a,\theta_{0a}), \end{aligned}$$

where here $F(\theta_a,\theta_{0a}) = [\theta_a^2 + 2\theta_a\theta_{0a} + 1]/[(1-\theta_a^2)(1-\theta_{0a}^2)] \geq 1$, for all $0 \leq \theta_a, \theta_{0a} < 1$. To sum up, condition (i) of Lemma F.2 is satisfied for all sequences p_n such that $p_n = O(N_{\min})$.

Checking the second condition

Remember that in the proof of Theorem 2, we controled the quantity

$$\sup_{g_{(1,2)}\in\mathcal{G}_{(1,2)},\ldots,g_{(d-1,d)}\in\mathcal{G}_{(d-1,d)}} \left| \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \frac{W_{ni}^{(a)}}{\pi_n} g_a(X_i) - \mathrm{E}\, g_a(X_1) \right|$$

for arbitrary classes $\mathcal{G}_a$. When the dimension grows to infinity, the chances for this strategy to succeed are slim. One needs to take into account the fact that some of the bivariate functions g_a coincide.

In our model, all of the classes $\mathcal{G}_a$ are the same: $\mathcal{G}_a = \mathcal{G}_{(1,2)} = \{\ell_{(1,2)}(\cdot,\cdot;\theta), 0 \leq \theta < 1\}$ for every $a \in \mathcal{A}$. Moreover, all the entries $v_a(\theta)$ of Σ_θ are either one of the parameters $\theta_1, \ldots, \theta_q$ or 0. To each $a \in \mathcal{A}^+$ there corresponds an integer between 1 and q, denoted by $k(a)$, such that $v_a(\theta)$, the parameter of the marginal f_a, is equal to $\theta_{k(a)}$. Notice that when $a \notin \mathcal{A}^+$, the log density $\ell_a(X_i^{(a)}; v_a(\theta)) = \ell_a(X_i^{(a)}; 0)$ plays the role of a constant, and hence does not show up in the sum defining the randomized pairwise likelihood. Therefore, we have

$$\sup_{\theta_1,\ldots,\theta_q} \frac{|L_n^{\mathrm{RPL}}(\theta) - L^{\mathrm{PL}}(\theta)|}{p_n} = \sup_{g_1,\ldots,g_q \in \mathcal{G}_{(1,2)}} \frac{1}{np_n} \left| \sum_i \sum_{k=1}^{q} \sum_{a:k(a)=k} \frac{W_{ni}^{(a)}}{\pi_n} g_k(X_i^{(a)}) - \mathrm{E}\, g_k(X_1^{(a)}) \right|,$$

where above the elements $g_1, \ldots, g_q$ are distinct.

We now proceed as in the proof of Lemma E.2 with $A = d_n(d_n - 1)/2$. Let $\delta > 0$. Let $N = N(\delta, \mathcal{G}_{(1,2)}, P)$ denote the bracketing number of the class $\mathcal{G}_{(1,2)}$. There are brackets $[g_j^{\mathrm{L}}, g_j^{\mathrm{U}}]$, $j = 1, \ldots, N$, such that (i) $\int g_j^{\mathrm{U}} - g_j^{\mathrm{L}} \,\mathrm{d}P < \delta$ for all $j \in \{1, \ldots, N\}$ and (ii) for every $g \in \mathcal{G}_{(1,2)}$, there is $j \in \{1, \ldots, N\}$ such that $g_j^{\mathrm{L}} \leq g \leq g_j^{\mathrm{U}}$. For each $k = 1, \ldots, q$, choose $g_k \in \mathcal{G}_{(1,2)}$ and denote by $[g_{j(k)}^{\mathrm{L}}, g_{j(k)}^{\mathrm{U}}]$ the pair of brackets such that $g_{j(k)}^{\mathrm{L}} \leq g_k \leq g_{j(k)}^{\mathrm{U}}$. It holds that

$$-\frac{|\mathcal{A}^+|\delta}{p_n} + \frac{L_{n,\mathbf{g}}}{p_n} \leq \frac{1}{np_n} \left| \sum_i \sum_{k=1}^{q} \sum_{a:k(a)=k} \frac{W_{ni}^{(a)}}{\pi_n} g_k(X_i^{(a)}) - \mathrm{E}\, g_k(X_1^{(a)}) \right| \leq \frac{U_{n,\mathbf{g}}}{p_n} + \frac{|\mathcal{A}^+|\delta}{p_n},$$

where the random variables

$$\begin{aligned} U_{n,\mathbf{g}} &= \frac{1}{n} \sum_i \sum_{k=1}^{q} \sum_{a:k(a)=k} \left(\frac{W_{n,i}^{(a)}}{\pi_n} g_{j(k)}^{\mathrm{U}}(X_i^{(a)}) - \mathrm{E}\, g_{j(k)}^{\mathrm{U}}(X_1^{(a)}) \right), \\ L_{n,\mathbf{g}} &= \frac{1}{n} \sum_i \sum_{k=1}^{q} \sum_{a:k(a)=k} \left(\frac{W_{n,i}^{(a)}}{\pi_n} g_{j(k)}^{\mathrm{L}}(X_i^{(a)}) - \mathrm{E}\, g_{j(k)}^{\mathrm{L}}(X_1^{(a)}) \right) \end{aligned}$$

depend on $\mathbf{g} = (g_1, \ldots, g_q)$ only through the brackets that enclose them. There are, therefore, at most qN possible bracket combinations and at most as many distinct $U_{n,\mathbf{g}}$ and $L_{n,\mathbf{g}}$. Since N and q are fixed constants, it is sufficient to show that each of the qN possible $U_{n,\mathbf{g}}/p_n$ and $L_{n,\mathbf{g}}/p_n$ vanish in probability and $\limsup |\mathcal{A}^+|/p_n < \infty$. Since it was shown that $p_n = O(N_{\min})$, we take $p_n = N_{\min}$, leading to the condition $|\mathcal{A}^+| = O(N_{\min})$.

It remains to show that $U_{n,\mathbf{g}}/N_{\min}$ and $L_{n,\mathbf{g}}/N_{\min}$ vanish in probability. Let us focus on $U_{n,\mathbf{g}}/N_{\min}$; the reasoning for $L_{n,\mathbf{g}}/N_{\min}$ is similar. By the same arguments as in the proof of Lemma E.1, we have, for every $\epsilon > 0$,

$$\begin{aligned}
&P\left(\frac{1}{nN_{\min}}\left|\sum_i \sum_{k=1}^q \sum_{a:k(a)=k}\left(\frac{W_{n,i}^{(a)}}{\pi_n} g_{j(k)}(X_i^{(a)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a)})\right)\right| > \epsilon\right)\\
&\leq \frac{\mathrm{Var}\left[\sum_k \sum_{a:k(a)=k}\left(\frac{W_{n,1}^{(a)}}{\pi_n} g_{j(k)}(X_1^{(a)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a)})\right)\right]}{nN_{\min}^2\epsilon^2}\\
&= \frac{(1-\pi_n)\sum_k \sum_{a:k(a)=k} \mathrm{E}\, g_{j(k)}(X_1^{(a)})^2}{n\pi_n N_{\min}^2\epsilon^2} + \frac{\mathrm{E}\left(\sum_k \sum_{a:k(a)=k}(g_{j(k)}(X_1^{(a)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a)}))\right)^2}{nN_{\min}^2\epsilon^2},
\end{aligned}$$

where above we have left the superscript "U". Notice that for each $k = 1, \dots, q$, the bivariate vectors $X_1^{(a)}$ are identically distributed for all $a \in \mathcal{A}$ such that $k(a) = k$. Therefore, with every choice $a_1, \dots, a_q \in \mathcal{A}$ such that $k(a_k) = k$, it holds that $\sum_{k=1}^q \sum_{a:k(a)=k} \mathrm{E}\, g_{j(k)}(X_1^{(a)})^2 = \sum_{k=1}^q N_k \,\mathrm{E}\, g_{j(k)}(X_1^{(a_k)})^2 \leq N_{\max} \sum_{k=1}^q \mathrm{E}\, g_{j(k)}(X_1^{(a_k)})^2$. Even though the dimension d is allowed to go to infinity, this sum remains constant. The same reasoning yields

$$
\begin{aligned}
&\mathrm{E}\left(\sum_k \sum_{a:k(a)=k} (g_{j(k)}(X_1^{(a)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a)}))\right)^2 \\
&= \mathrm{E}\left(\sum_k N_k (g_{j(k)}(X_1^{(a_k)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a_k)}))\right)^2 \\
&\leq N_{\max}^2 \,\mathrm{E} \sum_k (g_{j(k)}(X_1^{(a_k)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a_k)}))^2 \\
&\quad + \sum_{k,k'} N_k N_{k'}\, \mathrm{E}(g_{j(k)}(X_1^{(a_k)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a_k)}))(g_{j(k')}(X_1^{(a_{k'})}) - \mathrm{E}\, g_{j(k')}(X_1^{(a_{k'})})) \\
&\leq N_{\max}^2 \left(\mathrm{E} \sum_k (g_{j(k)}(X_1^{(a_k)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a_k)}))^2 \right. \\
&\quad \left. + \sum_{k \neq k'} \sqrt{\mathrm{E}(g_{j(k)}(X_1^{(a_k)}) - \mathrm{E}\, g_{j(k)}(X_1^{(a_k)}))^2} \sqrt{\mathrm{E}(g_{j(k')}(X_1^{(a_{k'})}) - \mathrm{E}\, g_{j(k')}(X_1^{(a_{k'})}))^2} \right).
\end{aligned}
$$

Thus, the conditions under which $U_{n,\mathbf{g}}$ vanishes in probability are

$$
\frac{N_{\max}}{n\pi_n N_{\min}^2} \to 0 \text{ and } \frac{N_{\max}^2}{n N_{\min}^2} \to 0.
$$

G Proofs of the lemmas in Sections E and F

Proof of Lemma E.1

We have

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

Let $\epsilon > 0$. Chebychev's inequality yields

$$P\left(\left|\frac{1}{n}\sum_i\sum_a\left(\frac{W_{ni}^{(a)}}{\pi_n}g_a(X_i) - \mathrm{E}\,g_a(X_1)\right)\right| > \epsilon\right) \le \frac{\mathrm{E}\left|\sum_i\sum_a\left(\frac{W_{ni}^{(a)}}{\pi_n}g_a(X_i) - \mathrm{E}\,g_a(X_1)\right)\right|^2}{n^2\epsilon^2}.$$

Since the random variables $\sum_a(\pi_n^{-1}W_{ni}^{(a)}g_a(X_i) - \mathrm{E}\,g_a(X_1))$, $i = 1, \ldots, n$, are i.i.d. and centered, the upper bound is equal to

$$\begin{aligned}&\frac{\mathrm{Var}\sum_a\left(W_{n1}^{(a)}g_a(X_1)/\pi_n - \mathrm{E}\,g_a(X_1)\right)}{n\epsilon^2}\\ &= \frac{(1-\pi_n)\sum_a \mathrm{E}\,g_a(X_1)^2}{n\pi_n\epsilon^2} + \frac{\mathrm{E}\left(\sum_a g_a(X_1) - \mathrm{E}\,g_a(X_1)\right)^2}{n\epsilon^2},\end{aligned}$$

which goes to zero whether π_n is constant or $\pi_n \to 0$ because $n\pi_n \to \infty$ either way.

Proof of Lemma E.2

We shall follow the track of the proof of Lemma 3.1 in (van de Geer, 2000, p. 26). We have

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

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 $g_{a,j(a)}^{\mathrm{L}} \le g_a \le g_{a,j(a)}^{\mathrm{U}}$. Choose $g_a \in \mathcal{G}_a$ for each

a. We have

$$\begin{aligned}
&-A\delta + L_{n,\mathbf{g}} := -A\delta + \frac{1}{n}\sum_{i,a}\left(\frac{W_{ni}^{(a)}}{\pi_n} g^{\mathrm{L}}_{a,j(a)}(X_i) - \mathrm{E}\, g^{\mathrm{L}}_{a,j(a)}(X_1)\right)\\
&\le \frac{1}{n}\sum_{i,a}\left(\frac{W_{ni}^{(a)}}{\pi_n} g_a(X_i) - \mathrm{E}\, g_a(X_1)\right)\\
&\le \frac{1}{n}\sum_{i,a}\left(\frac{W_{ni}^{(a)}}{\pi_n} g^{\mathrm{U}}_{a,j(a)}(X_i) - \mathrm{E}\, g^{\mathrm{U}}_{a,j(a)}(X_1)\right) + A\delta =: U_{n,\mathbf{g}} + A\delta.
\end{aligned}$$

In the above inequality, the random variable $U_{n,\mathbf{g}}$ depends on the elements g_a that have been chosen in the classes $\mathcal{G}_a$, but only through $\mathbf{g} := \{g^*_{a,j},\ j = 1, \ldots, N_a,\ a = 1, \ldots, A,\ * \in \{\mathrm{U}, \mathrm{L}\}\}$, the brackets "enclosing" the elements g_a. Since the total number of brackets is finite, so is the number of random variables $U_{n,\mathbf{g}}$. (In fact, at most $N_1 + \cdots + N_A$ distinct $U_{n,\mathbf{g}}$ can show up in the inequality.) By Lemma E.1, each one of them vanishes in probability, regardless of the behavior of the sequence π_n. The same chain of arguments applies for the random variables $L_{n,\mathbf{g}}$. Therefore, since δ was arbitrary, the supremum over all possible $g_1 \in \mathcal{G}_1, \ldots, g_A \in \mathcal{G}_A$ of the term lying between $-A\delta + L_{n,\mathbf{g}}$ and $U_{n,\mathbf{g}} + A\delta$ also vanishes in probability. The proof is complete.

Proof of Lemma E.3

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

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

where

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

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. (van der Vaart, 1998, p. 20)), this is true under two conditions:

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

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

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

where above $\mathbf{1}\{\cdot\}$ denotes the indicator function. 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

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

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

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

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

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

The last inequality is an application of Hoeffding's inequality, see e.g (van de Geer, 2000, 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 E.3 implies that the expectation above goes to zero. The proof is complete.

Proof of Lemma F.2

From (i), $|\hat{\theta}_n^{\mathrm{MRPL}}-\theta_0|\geq\epsilon$ implies $(1/p_n)(L^{\mathrm{PL}}(\hat{\theta}_n^{\mathrm{MRPL}})-L^{\mathrm{PL}}(\theta_0))\leq-\lambda$ and hence

$$P\left(|\hat{\theta}_n^{\mathrm{MRPL}}-\theta_0|\geq\epsilon\right)\leq P\left(\frac{L^{\mathrm{PL}}(\theta_0)-L^{\mathrm{PL}}(\hat{\theta}_n^{\mathrm{MRPL}})}{p_n}\geq\lambda\right).$$

The proof will be complete if we can show that in the probability on the right, the random variable in the left-hand side of the inequality vanishes in probability as $n\to\infty$. Thus, let us write

$$\begin{aligned}\frac{L^{\mathrm{PL}}(\theta_0)-L^{\mathrm{PL}}(\hat{\theta}_n^{\mathrm{MRPL}})}{p_n} &= \frac{L^{\mathrm{PL}}(\theta_0)-L_n^{\mathrm{RPL}}(\theta_0)}{p_n}\\ &\quad+\frac{L_n^{\mathrm{RPL}}(\theta_0)-L_n^{\mathrm{RPL}}(\hat{\theta}_n^{\mathrm{MRPL}})}{p_n}\\ &\quad+\frac{L_n^{\mathrm{RPL}}(\hat{\theta}_n^{\mathrm{MRPL}})-L^{\mathrm{PL}}(\hat{\theta}_n^{\mathrm{MRPL}})}{p_n}.\end{aligned}$$

The first and the last terms in the right-hand side of the above inequality vanish in probability by (ii). The term in the middle is nonpositive by definition. Therefore, since the left-hand side is nonnegative, it must go to zero in probability as well. The proof is complete.

H Bound on an integral

Lemma H.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

$$-\alpha(1 - 2\log x)(\beta + \gamma x^4)^2 - \lambda\gamma x^4 \geq 2\lambda\beta. \tag{S6}$$

Note that if $x \leq e^{1/2}$ then $f'(x) \leq 0$. Otherwise, (S6) is equivalent to

$$x^4(\varphi_1(x) + \varphi_2(x) - \lambda\gamma) + \varphi_3(x) \geq 2\lambda\beta, \tag{S7}$$

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 φ_1, φ_2 and φ_3 are increasing and nonnegative on $[e^{1/2}, \infty)$. Thus the function in the left-hand side of (S7) 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 (S7) and hence (S6) 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 (S7) we know that $x^\star \to \infty$. This implies that the limit is as required. $\square$

Lemma H.2. *Let ϖ 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 λ goes to infinity. (See Lemma H.1.) Taking $\nu^2 = 2\sigma^2$ ensures that the inequality is true for all x as soon as λ is greater than some number λ_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 H.1. *The integral $I(\lambda)$ defined in Lemma H.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 H.2, we may take $\eta = \lambda^a$, $a > 0$, because both η and λ 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 $a = 1/6$,

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

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. $\square$

References

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

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

Pollard, D. (1984). *Convergence of stochastic processes.* Springer.

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

van de Geer, S. A. (2000). *Empirical Processes in M-estimation.* Cambridge University Press.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes.* Springer.