A parallel algorithm for ridge-penalized estimation of the multivariate exponential family from data of mixed types

Laman Trip, Diederik S.; Wieringen, Wessel N.van

DOI
10.1007/s11222-021-10013-x

Publication date
2021

Document Version
Final published version

Published in
Statistics and Computing

Citation (APA)
Laman Trip, D. S., & Wieringen, W. N. V. (2021). A parallel algorithm for ridge-penalized estimation of the multivariate exponential family from data of mixed types. *Statistics and Computing*, *31*(4), Article 41. https://doi.org/10.1007/s11222-021-10013-x

Statistics and Computing (2021) 31:41
https://doi.org/10.1007/s11222-021-10013-x

A parallel algorithm for ridge-penalized estimation of the multivariate exponential family from data of mixed types

Diederik S. Laman Trip[1] · **Wessel N. van Wieringen**[2,3]

Received: 25 July 2019 / Accepted: 9 April 2021

Abstract
Computationally efficient evaluation of penalized estimators of multivariate exponential family distributions is sought. These distributions encompass among others Markov random fields with variates of mixed type (e.g., binary and continuous) as special case of interest. The model parameter is estimated by maximization of the pseudo-likelihood augmented with a convex penalty. The estimator is shown to be consistent. With a world of multi-core computers in mind, a computationally efficient parallel Newton–Raphson algorithm is presented for numerical evaluation of the estimator alongside conditions for its convergence. Parallelization comprises the division of the parameter vector into subvectors that are estimated simultaneously and subsequently aggregated to form an estimate of the original parameter. This approach may also enable efficient numerical evaluation of other high-dimensional estimators. The performance of the proposed estimator and algorithm are evaluated and compared in a simulation study. Finally, the presented methodology is applied to data of an integrative omics study.

Keywords Markov random field · Consistency · Pseudo-likelihood · Block-wise Newton–Raphson · Network · Parallel algorithm · Graphical model

1 Introduction

With the increasing capacity for simultaneous measurement of an individual's many traits, networks have become an omnipresent visualization tool to display the cohesion among these traits. For instance, the cellular regulatory network portraits the interactions among molecules like mRNAs and/or proteins. Statistically, a network captures the relationships among variates implied by a joint probability distribution describing the simultaneous random behavior of the variates. These variates may be of different type, representing—for example—traits with continuous, count, or binary state spaces. Generally, the relationship network is unknown and is to be reconstructed from data. To this end, we present methodology that learns the network from data with variates of mixed types in a computationally efficient manner.

A collection of p variates of mixed type is mostly modeled by a pairwise Markov random field (MRF) distribution (a special case of the multivariate exponential family). A Markov random field is a set of random variables $Y_1, \ldots, Y_p$ that satisfies certain conditional independence properties specified by an undirected graph. This is made more precise by introduction of the relevant notions. A graph is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a finite set of vertices or nodes $\mathcal{V}$ and a collection of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ that join node pairs. In an undirected graph, any edge is undirected, i.e., $(v_1, v_2) \in \mathcal{E}$ is an unordered pair implying that $(v_2, v_1) \in \mathcal{E}$. A subgraph $\mathcal{G}' \subseteq \mathcal{G}$ with $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$ is a clique if $\mathcal{G}'$ is complete, i.e., all nodes are directly connected to all other nodes. The neighborhood of a node $v \in \mathcal{V}$, denoted $N(v)$, is the collection of nodes in $\mathcal{V}$ that are adjacent to v: $N(v) = \{v' \in \mathcal{V} \mid (v, v') \in \mathcal{E}, v \neq v'\}$. The closed neighborhood is simply $v \cup N(v)$ and denoted by $N[v]$. Now let $\mathbf{Y}$ be a p-dimensional random vector. Represent each variate of $\mathbf{Y}$ with a node in a graph $\mathcal{G}$ with $\mathcal{V} = \{1, \ldots . p\}$. Node names thus index the elements of $\mathbf{Y}$. Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be exhaustive and mutually exclusive subsets of $\mathcal{V} = \{1, \ldots . p\}$. Define the random vectors $\mathbf{Y}_a$, $\mathbf{Y}_b$ and $\mathbf{Y}_c$ by restricting the

✉ Wessel N. van Wieringen
w.vanwieringen@amsterdamumc.nl

[1] Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands

[2] Department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam UMC, location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands

[3] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Published online: 17 May 2021

p-dimensional random vector $\mathbf{Y}$ to the elements of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, respectively. Then $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are conditionally independent given random vector $\mathbf{Y}_c$, written as $\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b \,|\, \mathbf{Y}_c$, if and only if their joint probability distribution factorizes as $P(\mathbf{Y}_a, \mathbf{Y}_b \,|\, \mathbf{Y}_c) = P(\mathbf{Y}_a \,|\, \mathbf{Y}_c) \cdot P(\mathbf{Y}_b \,|\, \mathbf{Y}_c)$. The random vector $\mathbf{Y}$ satisfies the local Markov property with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $Y_j \perp\!\!\!\perp \mathbf{Y}_{\mathcal{V}\setminus N[j]} \,|\, \mathbf{Y}_{N(j)}$ for all $j \in \mathcal{V}$. Graphically, conditioning on the neighbors of j detaches j from $\mathcal{V} \setminus N[j]$. A Markov random field (or undirected graphical model) is a pair $(\mathcal{G}, \mathbf{Y})$ consisting of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with associated random variables $\mathbf{Y} = \{Y_j\}_{j\in\mathcal{V}}$ that satisfy the local Markov property with respect to $\mathcal{G}$ (cf. Lauritzen 1996). For strictly positive probability distributions of $\mathbf{Y}$ and by virtue of the Hammersley–Clifford theorem (Hammersley and Clifford 1971), the local Markov property may be assessed through the factorization of the distribution in terms of clique functions, i.e., functions of variates that correspond to a clique's nodes of the associated graph $\mathcal{G}$.

In this work, we restrict ourselves to cliques of size at most two. Thus, only pairwise interactions between the variates of $\mathbf{Y}$ are considered. Although restrictive, many higher-order interactions can be approximated by pairwise interactions (confer, e.g., Gallagher et al. 2011). Under the restriction to pairwise interactions and the assumption of a strictly positive distribution, the probability distribution can be written as:

$$P(\mathbf{Y}) = \exp\Big[-\sum\nolimits_{j,j'\in\mathcal{V}} \phi_{j,j'}(Y_j, Y_{j'}) - D\Big], \tag{1}$$

with log-normalizing constant or log-partition function D and pairwise log-clique functions $\{\phi_{j,j'}\}_{j,j'\in\mathcal{V}}$. The pairwise MRF distribution $P(\mathbf{Y})$, and therefore the graphical structure, is fully known once the log-clique functions are specified. In particular, nodes $j, j' \in \mathcal{V}$ are connected by an edge whenever $\phi_{j,j'} \neq 0$ as the probability distribution of $\mathbf{Y}$ would then not factorize in terms of the variates, Y_j and $Y_{j'}$, constituting this clique.

The estimation of the strictly positive MRF distribution (1) with pairwise interactions will be studied here. This is hampered by the complexity of the log-partition function. Although analytically known, for example for the multivariate normal distribution, it is—in general—computationally not feasible to evaluate. Indeed, the partition function is computationally intractable for MRFs that have variables with a finite state space (Welsh 1993; Höfling and Tibshirani 2009), or more generally for MRFs with variables of mixed type (Lee and Hastie 2013). In effect, maximum likelihood estimation is prohibited computationally. Instead, parameters will be estimated by means of pseudo-likelihood estimation. In particular, as the number of parameters is often of the same order—if not larger—as the sample size, the pseudo-likelihood will be augmented with a penalty. An overview of related work, which concentrates mainly on ℓ_1-penalization, can be found in Supplementary Material A (henceforth SM).

The contribution of this work to existing literature is threefold. In short, (i) we present machinery for estimation of the mixed variate graphical model with a quadratic penalty, (ii) we propose an efficient parallel algorithm for evaluating this estimator, and (iii) we created a software package that implements this algorithm to learn graphical models from data of more than two different variable types.

Specifically, we present machinery for estimation of the mixed variate graphical model with a quadratic penalty, i.e., ridge or ℓ_2. Our motivation for ridge penalized estimation is multifold: (i) ridge estimators are unique, and (ii) an analytic expression or a stable algorithm is available for their evaluation, preventing convergence problems often exhibited by lasso-type estimators. (iii) Ridge estimators generally yield a better fit than those of lasso-type, as has been observed in the graphical model context (van Wieringen and Peeters 2016; Miok et al. 2017; Bilgrau et al. 2020). (iv) The dominant paradigm of sparsity is not necessarily valid in all fields of application. For example, more dense (graphical) structures are advocated in molecular biology (Boyle et al. 2017). (v) If desired, the smoothness and strict convexity of the ridge penalty can be used to approximate other penalties (Fan and Li 2001), as previously done for the generalized lasso/elastic net in graphical model context (van Wieringen 2019). SM B contains a more elaborate motivation.

The second contribution of our work is to be found in the efficient algorithm for the evaluation of the presented estimator. This exploits the high degree of parallelization allowed by modern computing systems. We developed a Newton–Raphson procedure that uses *full* (instead of partial) second-order information with comparable computational complexity to existing methods that use only limited second-order information. Our approach translates to other high-dimensional estimators that may profit in their numerical evaluation.

Thirdly, this work is complemented with a software implementation to learn graphical models from data of more than two different variable types. This is a practical and relevant contribution as medical and biological fields measure more and more different types of traits of samples. This can be witnessed from the TCGA (The Cancer Genome Atlas) repository, where many types of the molecular traits of cancer samples are measured. In current developments, this molecular information is augmented with imaging data (referred to as radiomics, Gillies et al. 2015). Additionally, these data are further complemented with a sample's exposome, i.e., a quantification its environmental exposure (Wild 2012). Thus, there is a need for methods and implementations that can deal with data comprising more than two types.

The paper is structured as follows. First, Sect. 2 recaps the pairwise MRF distribution for variates of mixed types

as a special case of the more general exponential family, along with parameter constraints that ensure its well-definedness. Next, Sect. 3 presents a consistent penalized pseudo-likelihood estimator for the exponential family model parameter—thereby also for that of the pairwise MRF distribution. Then, Sect. 4 introduces a form of the Newton–Raphson algorithm to numerically evaluate this estimator. The algorithm is parallelized to exploit the multi-core capabilities of modern computing systems, and conditions that ensure convergence of the algorithm are described. Finally, Sect. 5 presents (a) an *in silico* comparison of the estimator to related ones and (b) a simulation study into the computational performance of the algorithm.

2 Model

This section describes the graphical model for data of mixed types. In its most general form, it is any exponential family distribution. Within the exponential family the model is first specified variate-wise, conditionally on all other variates. The parametric form of this conditionally formulated model warrants that the implied joint distribution of the variates is also an exponential family member. This correspondence between the variate-wise and joint model parameters endows the former (by way of zeros in the parameter) with a direct relation to conditional independencies between variate pairs, thus linking it to the underlying graph. Finally, parameter constraints are required to ensure that the proposed distribution is well-defined.

The multivariate exponential family is a broad class of probability distributions that describe the joint random behavior of a set of variates (possibly of mixed type). It encompasses many distributions for variates with a continuous, count and binary outcome space. All distributions share the following functional form:

$$f_{\boldsymbol{\Theta}}(\mathbf{y}) = h(\mathbf{y}) \exp\{\eta(\boldsymbol{\Theta}) T(\mathbf{y}) - D[\eta(\boldsymbol{\Theta})]\},$$

where $\boldsymbol{\Theta}$ is a $p \times p$-dimensional parameter matrix, $h(\mathbf{y})$ is a nonnegative base measure, $\eta(\boldsymbol{\Theta})$ is the natural or canonical parameter, $T(\mathbf{y})$ the sufficient statistic, and $D[\eta(\boldsymbol{\Theta})]$, the log-partition function or the normalization factor, which ensures $f_{\boldsymbol{\Theta}}(\mathbf{y})$ is indeed a probability distribution. The log-partition function $D[\eta(\boldsymbol{\Theta})]$ needs to be finite to ensure a well-defined distribution. Standard distributions are obtained for specific choices of η, T and h. Theoretical results presented in Sects. 3 and 4 are stated for the multivariate exponential family and therefore apply to all encompassing distributions. To provide for the envisioned practical purpose of reconstruction of the conditional dependence graph, we require and outline next a Markov random field in which the variates follow a particular exponential family member conditionally. This is thus a special case of the delineated class of exponential family distributions, as will be obvious from the parametric form of the Markov random field distribution.

Following Besag (1974) and Yang et al. (2014), the probability distribution of each individual variate of Y_j of $\mathbf{Y}$ conditioned on all remaining variates $\mathbf{Y}_{\setminus j}$ is assumed to be a (potentially distinct) univariate exponential family member, e.g., a Gaussian or Bernoulli distribution. Its (conditional) distribution is:

$$P(Y_j \mid \mathbf{Y}_{\setminus j}) \propto h_j(Y_j) \times \exp\left[\eta_j(\boldsymbol{\Theta}_{j,\setminus j}; \mathbf{Y}_{\setminus j}) T_j(Y_j) - D_j(\eta_j)\right]. \tag{2}$$

Theorem 1 specifies the joint distribution for graphical models of variates that have conditional distribution (2). In particular, it states that there exists a joint distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ of $\mathbf{Y}$ such that $(\mathcal{G}, \mathbf{Y})$ is a Markov random field if and only if each variate depends conditionally on the other variates through a linear combination of their univariate sufficient statistics.

Theorem 1 *(after Yang et al. 2014)*

Consider a p-variate random variable $\mathbf{Y} = \{Y_j\}_{j \in \mathcal{V}}$. *Assume the distributions of each variate* Y_j, $j \in \mathcal{V}$, *conditionally on the remaining variates to be an exponential family member as in (2). Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *be a graph which decomposes into* $\mathcal{C}$, *the set of cliques of size at most two. Finally, the off-diagonal support of the MRF parameter* $\boldsymbol{\Theta}$ *matches the edge structure of* $\mathcal{G}$. *Then, the following assumptions are equivalent:*

i) For $j \in \mathcal{V}$, *the natural parameter* η_j *of the variate-wise conditional distribution (2) is:*

$$\eta_j(\boldsymbol{\Theta}_{j,\setminus j}; \mathbf{Y}_{\setminus j}) = \boldsymbol{\Theta}_{j,j} + \sum_{\{j' \in \mathcal{V} : (j,j') \in \mathcal{E}_C, C \in \mathcal{C}\}} \boldsymbol{\Theta}_{j,j'} T_{j'}(Y_{j'}). \tag{3}$$

ii) There exists a joint distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ *of* $\mathbf{Y}$ *such that* $(\mathcal{G}, \mathbf{Y})$ *is a Markov random field.*

Moreover, by either assumption the joint distribution of $\mathbf{Y}$ *is:*

$$P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \prod_{j \in \mathcal{V}} h_j(Y_j) \exp\left\{T_j(Y_j)\left[\boldsymbol{\Theta}_{j,j} + \sum_{\{j' \in \mathcal{V} : (j,j') \in \mathcal{E}_C, C \in \mathcal{C}\}} \boldsymbol{\Theta}_{j,j'} T_{j'}(Y_{j'})\right]\right\}. \tag{4}$$

The theorem above differs from the original formulation in Yang et al. (2014) in the sense that here it is restricted to pairwise interactions (i.e., cliques of size at most two).

For the reconstruction of the graph underlying the Markov random field, the edge set $\mathcal{E}$ is captured by the parameter $\boldsymbol{\Theta}$: nodes $j, j' \in \mathcal{V}$ are connected by a direct edge $(j, j') \in \mathcal{E}$ if and only if $\boldsymbol{\Theta}_{j,j'} \neq 0$ [by the Hammersley-Clifford theorem, Lauritzen (1996)]. This gives a simple parametric criterion to assess local Markov (in)dependence. Moreover, the parameter $\boldsymbol{\Theta}_{j,j'}$ can be interpreted as an interaction parameter between variables Y_j and $Y_{j'}$.

We refer to distribution (4) as the pairwise MRF distribution. After normalization of (4), the joint distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is fully specified by sufficient statistics and base measures of the exponential family members. For practical and illustrative purposes, the remainder will feature—but is not limited to—only four common exponential family members, the *GLM family*: the Gaussian (with unknown variance), exponential, Poisson and Bernoulli distributions.

The joint distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ formed from the variate-wise conditional distributions need not be well-defined for arbitrary parameter choices. In order for $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ to be well-defined, the log-normalizing constant $D[\eta(\boldsymbol{\Theta})]$ needs to be finite. For example, for the Gaussian graphical model, a special case of the pairwise MRF distribution under consideration, this is violated when the covariance matrix is singular. Lemma 1 of Chen et al. (2015) specifies the constraints on the parameter $\boldsymbol{\Theta}$ that ensure a well-defined pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ when the variates of $\mathbf{Y}$ are GLM family members conditionally (see SM C for details).

These parameter constraints are restrictive on the structure of the graph and the admissible interactions. As the graph is implicated by the off-diagonal support of $\boldsymbol{\Theta}$, the constraints for well-definedness imply that the nodes corresponding to conditionally Gaussian random variables cannot be connected to the nodes representing exponential and/or Poisson random variables. Moreover, when Y_j and $Y_{j'}$ are assumed to be Poisson and/or exponential random variables conditionally on the other variates, their interaction can only be negative. However, these restrictions could be relaxed by modeling data with, for example, a truncated Poisson distribution (Yang et al. 2014).

3 Estimation

The parameter $\boldsymbol{\Theta}$ of the multivariate exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is now to be learned from (high-dimensional) data. Straightforward maximization of the penalized loglikelihood is impossible due to the fact that the log-partition function cannot be evaluated in practice. For example, the partition function of the Ising model with p binary variates sums over all 2^p configurations. For large p, this becomes computationally intractable for almost all Ising models. This is circumvented by the replacement of the likelihood by the pseudo-likelihood comprising the variate-wise conditional distributions (Besag 1974; Höfling and Tibshirani 2009). We show that the maximum penalized pseudo-likelihood estimator of the exponential family model parameter is—under conditions—consistent. Finally, we present a computationally efficient algorithm for the numerical evaluation of this proposed estimator. Both results carry over to the pairwise MRF parameter as special case of the multivariate exponential family.

Consider an identically and independently distributed sample of p-variate random variables $\{\mathbf{Y}_i\}_{i=1}^n$ all drawn from $P_{\boldsymbol{\Theta}}$. The associated (sample) pseudo-loglikelihood is a composite loglikelihood of all variate-wise conditional distributions averaged over the observations:

$$\mathcal{L}_{\mathrm{PL}}(\boldsymbol{\Theta}, \{\mathbf{Y}_i\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \mathcal{V}} \log[P_{\boldsymbol{\Theta}}(Y_{ij} \mid \mathbf{Y}_{i,\setminus j})]. \tag{5}$$

The maximum penalized pseudo-loglikelihood augments this by a strictly convex, continuous penalty function $f_{\mathrm{pen}}(\boldsymbol{\Theta}; \lambda)$ with penalty parameter $\lambda > 0$. Hence, $\mathcal{L}_{\mathrm{penPL}}(\boldsymbol{\Theta}, \{\mathbf{Y}_i\}_{i=1}^n) := \mathcal{L}_{\mathrm{PL}}(\boldsymbol{\Theta}, \{\mathbf{Y}_i\}_{i=1}^n) - f_{\mathrm{pen}}(\boldsymbol{\Theta}; \lambda)$. Then, the maximum penalized pseudo-likelihood estimator of $\boldsymbol{\Theta}$ is:

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{pen}}(\lambda) = \arg\max_{\boldsymbol{\Theta}} \mathcal{L}_{\mathrm{penPL}}(\boldsymbol{\Theta}, \{\mathbf{Y}_i\}_{i=1}^n). \tag{6}$$

The next theorem shows that the maximum penalized pseudo-likelihood estimator (6) is consistent in the traditional sense, i.e., a regime of fixed dimension p and an increasing sample size n. It is a minimum requirement of a novel estimator. A motivation for refraining from consistency results in high-dimensional regimes is provided in SM D.

Theorem 2 *Let $\{\mathbf{Y}_{i=1}^n\}$ be n independent draws from a p-variate exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})]$. Temporarily supply $\widehat{\boldsymbol{\Theta}}$ and λ with an index n to explicate their sample size dependence. Then the maximum penalized pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n^{\mathrm{pen}}$ maximizing the penalized pseudo-likelihood is consistent, i.e., $\widehat{\boldsymbol{\Theta}}_n^{\mathrm{pen}} \xrightarrow{p} \boldsymbol{\Theta}$ as $n \to \infty$ if,*

- *i) The parameter space is compact and such that $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined for all $\boldsymbol{\Theta}$,*
- *ii) $\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})$ can be bounded by a polynomial, $|\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})| \leq c_1 + c_2 \sum_{j \in \mathcal{V}} |Y_j|^{\beta}$ for constants $c_1, c_2 < \infty$ and $\beta \in \mathbb{N}$,*
- *iii) The penalty function $f_{\mathrm{pen}}(\boldsymbol{\Theta})$ is strict convex, continuous, and the penalty parameter λ_n converges in probability to zero: $\lambda_n \xrightarrow{p} 0$ as $n \to \infty$.*

Proof Refer to SM E. □

Theorem 2 differs from related theorems on ℓ_1-estimators in two respects. Most importantly, (i) it holds uniformly over

all (well-defined) models, i.e., it does not require a sparsity assumption. Moreover, (ii) the assumption on the penalty parameter is of a probabilistic rather than a specific deterministic nature, which we consider to be more suited as λ is later chosen in a data-driven fashion.

Theorem 2 warrants—under conditions—the convergence of the maximum penalized pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}$ as the sample size increases ($n \to \infty$). These conditions require a compact parameter space, a common assumption in the field of graphical models (Lee et al. 2015). Theorem 2 holds in general for any multivariate exponential family distribution and is therefore generally applicable with the pairwise MRF distribution as special case.

Finally, if the penalty function $f_{\text{pen}}(\boldsymbol{\Theta}; \lambda)$ is proportional to the sum of the square of the elements of the parameter, $f_{\text{pen}}(\boldsymbol{\Theta}; \lambda) = \frac{1}{2}\lambda\|\boldsymbol{\Theta}\|_F^2$ with $\|\cdot\|_F$ the Frobenius norm, it is referred to as the ridge penalty. With the ridge penalty, the estimator (6) is called the maximum ridge pseudo-likelihood estimator. Then, when $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined for the GLM family, we obtain the following corollary.

Corollary 1 *Let $\{\mathbf{Y}_i\}_{i=1}^n$ be p-variate independent draws from a well-defined pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ with parameter $\boldsymbol{\Theta}$. The ridge pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n^{\text{ridge}}$ that maximizes the ridge pseudo-likelihood is consistent, i.e., $\widehat{\boldsymbol{\Theta}}_n^{\text{ridge}} \xrightarrow{p} \boldsymbol{\Theta}$ as $n \to \infty$, if the parameter space is compact, and the penalty parameter λ_n converges in probability to zero: $\lambda_n \xrightarrow{p} 0$ as $n \to \infty$.*

Proof Refer to SM E. □

Note that, in practice—as recommended by Höfling and Tibshirani (2009)—we employ $f_{\text{pen}}(\boldsymbol{\Theta}; \lambda) = \frac{1}{2}\lambda \sum_{j,j'=1, j\neq j'}^p \boldsymbol{\Theta}_{j,j'}^2$, thus leaving the diagonal unpenalized. Empirically, we observed this yields a better model fit, which is intuitively understandable as the estimator is then able to (unconstrainedly) account for (at least) the marginal variation in each variate.

4 Algorithm

Maximization of the ridge pseudo-loglikelihood presents a convex optimization problem (a concave pseudo-loglikelihood and convex parameter space, SM E). We present a parallel block-wise Newton-Raphson algorithm for numerical evaluation of the penalized pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}(\lambda)$. We show that this algorithm yields a sequence of updated parameters that converge to $\widehat{\boldsymbol{\Theta}}(\lambda)$ and terminates after a finite number of steps. The results for the algorithm presented in this section hold for maximizing the penalized pseudo-loglikelihood for any multivariate exponential family and are not restricted to the pairwise MRF distribution.

Strict concavity of the optimization problem (6) and smoothness of $\mathcal{L}_{\text{penPL}}$ permit the application of the Newton-Raphson algorithm to find the estimate. The Newton–Raphson algorithm starts with an initial guess $\widehat{\boldsymbol{\Theta}}^{(0)}(\lambda)$ and—motivated by a Taylor series approximation—updates it sequentially. This generates a sequence $\{\widehat{\boldsymbol{\Theta}}^{(k)}(\lambda)\}_{k\geq 0}$ that converges to $\widehat{\boldsymbol{\Theta}}(\lambda)$ (Fletcher 2013). However, the Newton–Raphson algorithm requires inversion of the Hessian matrix and is reported to be slow for pseudo-loglikelihood maximization (Lee and Hastie 2013; Chen et al. 2015): It has computational complexity $O(p^6)$ for p variates. Instead of a naive implementation of the Newton-Raphson algorithm to solve (6), the remainder of this section describes a block-wise approach (Xu and Yin 2013), that speeds up the evaluation of the estimator by exploiting the structure of the pseudo-likelihood and splitting the optimization problem (6) into multiple simpler subproblems. These subproblems are then solved in parallel fashion. This parallel block-wise Newton–Raphson algorithm makes optimal use of available multi-core processing systems and is necessary to answer to the increasing size of data sets. Finally, in contrast to other pseudo-likelihood approaches (Höfling and Tibshirani 2009; Lee and Hastie 2013), the presented approach allows for the use of all second-order information (i.e., the Hessian) with the benefit of potentially faster convergence, but without increasing the computational complexity.

In order to describe the block-wise approach some notation is introduced. Define $q = \frac{1}{2}p(p+1)$, the number of unique parameters of $\boldsymbol{\Theta}$. The set of unique parameter indices is denoted by $\mathcal{Q} = \{(j, j') : j \leq j' \in \mathcal{V}\}$ and we use $\boldsymbol{\theta}$ as shorthand for the q-dimensional vector of unique parameters $\{\boldsymbol{\Theta}_{j,j'}\}_{(j,j')\in\mathcal{Q}}$. Furthermore, write $\boldsymbol{\theta}_j$ for $\boldsymbol{\Theta}_{*,j} = (\boldsymbol{\Theta}_{j,*})^\top$, the p-dimensional vector of all unique parameters of $\boldsymbol{\Theta}$ that correspond to the j-th variate. Consequently, for $j \neq j'$ the corresponding $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j'}$ have parameter(s) of $\boldsymbol{\Theta}$ in common. Finally, let $\mathbf{H}_j$ be the $p \times p$-dimensional submatrix of the Hessian limited to the elements that relate to the j-th variate, i.e., $\mathbf{H}_j = \partial^2 \mathcal{L}_{\text{penPL}} / \partial\boldsymbol{\theta}_j \partial\boldsymbol{\theta}_j^\top$.

The block-wise approach maximizes the penalized pseudo-loglikelihood with respect to the parameter subvector $\boldsymbol{\theta}_j$ for $j \in \mathcal{V}$, while all other parameters are temporarily kept constant at their current value. Per block we maximize by means of the Newton-Raphson algorithm, with initial guess $\hat{\boldsymbol{\theta}}^{(0)}(\lambda)$ and current parameter value $\hat{\boldsymbol{\theta}}_j^{(k)}(\lambda)$, updating to $\hat{\boldsymbol{\theta}}_j^{(k+1)}(\lambda)$ through:

$$\hat{\boldsymbol{\theta}}_j^{(k+1)}(\lambda) = \hat{\boldsymbol{\theta}}_j^{(k)}(\lambda) - \left(\frac{\partial^2 \mathcal{L}_{\text{penPL}}}{\partial\boldsymbol{\theta}_j \partial\boldsymbol{\theta}_j^\top}\right)^{-1}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k)}(\lambda)} \times \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial\boldsymbol{\theta}_j}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k)}(\lambda)}. \tag{7}$$

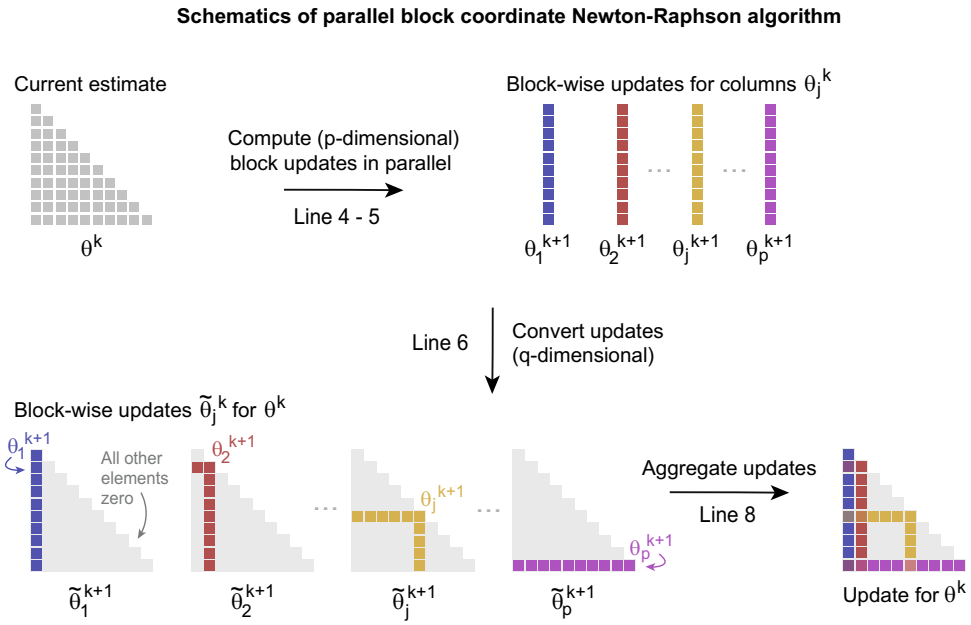

Fig. 1 Parameter updating by Algorithm 1. First, the p subvectors $\boldsymbol{\theta}_j$ are updated to novel $\hat{\boldsymbol{\theta}}_j$ (line 5 of Algorithm 1). These updates are interspersed with zero's to form the q-dimensional vectors $\tilde{\boldsymbol{\theta}}_j$ (line 6 of Algorithm 1). Finally, the $\tilde{\boldsymbol{\theta}}_j$'s are averaged weightedly to produce the update of $\boldsymbol{\theta}$ (line 8 of Algorithm 1)

Block coordinate-wise the procedure converges to the optimum, that is, the maximum of $\mathcal{L}_{\text{penPL}}$ given the other parameters of $\boldsymbol{\theta}$. Sequential application of the block-wise approach is—by the concavity of $\mathcal{L}_{\text{penPL}}$—then guaranteed to converge to the desired estimate. Sequential application of the block-wise approach may be slow and is ran in parallel for all $j \in \mathcal{V}$ simultaneously. This means that all $\{\hat{\boldsymbol{\theta}}_j^{(k+1)}\}_{j\in\mathcal{V}}$ are computed in parallel during a single step. As some elements of $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j'}$ map to the same element of $\boldsymbol{\theta}$, multiple estimates of the latter are thus available. Hence, the results of each parallel step need to be combined in order to provide a single update of the full estimate $\hat{\boldsymbol{\theta}}^{(k)}$. This update of $\hat{\boldsymbol{\theta}}^{(k)}$ should increase $\mathcal{L}_{\text{penPL}}$ and iteratively solve the concave optimization problem (6). We find such an update in the direction of the sum of the block-wise updates of $\{\hat{\boldsymbol{\theta}}_j^{(k+1)}\}_{j\in\mathcal{V}}$. A well-chosen step size in this direction then provides a suitable update of $\hat{\boldsymbol{\theta}}^{(k)}$. Alternatively, to avoid the need for combining block-wise updates one may seek a split of the elements of $\boldsymbol{\Theta}$ into blocks without overlap. This, however, raises several issues. First, there is no straightforward choice for coordinate blocks without overlap. Second, as the algorithm is parallelized one can only use the estimate from the previous step. Non-overlapping coordinate blocks optimize the pseudo-likelihood for their respective blocks, but are suboptimal for the entire parameter affecting the convergence. Finally, removing overlap requires a choice: which coordinate block provides the estimate for a shared parameter. There is no obvious rationale that tells which one should prevail. Moreover, there is no guarantee that the coordinate block with the overlapping elements removed still increases the pseudo-likelihood.

Algorithm 1 gives a pseudo-code description of the parallel block-wise Newton–Raphson algorithm (the combination of block-wise estimates is visualized in Fig. 1). Theorem 3 states that Algorithm 1 converges to the maximum penalized pseudo-likelihood estimator and terminates. While Theorem 3 is a rather general result for the maximum penalized pseudo-likelihood estimator of exponential family distributions, as special case the same result follows for the maximum ridge pseudo-likelihood estimator of the pairwise MRF distribution with the GLM family.

Theorem 3 *Let* $\{\mathbf{Y}_i\}_{i=1}^n$ *be* n *independent draws from a* p*-variate exponential family distribution* $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})]$. *Assume that the parameter space of* $\boldsymbol{\Theta}$ *is compact. Let* $\widehat{\boldsymbol{\Theta}}(\lambda)$ *be the unique global maximum of the penalized pseudo-likelihood* $\mathcal{L}_{\text{penPL}}(\boldsymbol{\Theta}, \{\mathbf{Y}_i\}_{i=1}^n)$. *Then, for any initial parameter* $\boldsymbol{\theta}^{(0)}$, *threshold* $\tau > 0$ *and sufficiently large multiplier* $\alpha \geq p$, *Algorithm 1 terminates after a finite number of steps and generates a sequence of parameters* $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ *that converge to* $\widehat{\boldsymbol{\Theta}}(\lambda)$.

Proof Refer to SM F. □

The presented Algorithm 1 balances computational complexity, convergence rate and optimal use of available information. The algorithm terminates after a finite number of steps and one step, i.e., lines 3–10, has computational complexity $O(p^3)$ when run in parallel. Moreover, Algorithm 1 uses all available second-order information (the Hessian of $\mathcal{L}_{\text{penPL}}$) and its convergence rate is at least linear. Furthermore, the convergence rate is quadratic when the multiple updates for each parameter are identical.

input : $n \times p$ data matrix $\mathbf{Y}$;
p exponential family members;
initial parameter $\boldsymbol{\theta}^{(0)}$;
penalty parameter $\lambda_n \in \mathbb{R}_{>0}$;
threshold $\tau \in \mathbb{R}_{>0}$;
multiplier $\alpha > 0$.
output: sequence $\{\boldsymbol{\theta}^{(k)}\}_{k \geq 0}$.

1 **initialize** $k = 0$, $err_0 = 2\tau$.
2 **while** $err_k > \tau$ **do**
3 **for** $j \in \mathcal{V}$ **do in parallel**
4 calculate the gradient $\partial \mathcal{L}_{\text{penPL}} / \partial \boldsymbol{\theta}_j$ and Hessian $\mathbf{H}_j$.
5 compute a single Newton-Raphson update of $\boldsymbol{\theta}_j$.
6 formulate the update as a q-dimensional vector $\tilde{\boldsymbol{\theta}}_j$ by:
7 $(\tilde{\boldsymbol{\theta}}_j)_q = \begin{cases} (\boldsymbol{\theta}_j)_{j'} & \text{for } q \in \mathcal{Q} \text{ s.t. } q = (j, j') \text{ or } q = (j', j), \\ 0 & \text{otherwise.} \end{cases}$
8 **end synchronize**
9 define the parameter estimate $\hat{\boldsymbol{\theta}}^{(k+1)} := \hat{\boldsymbol{\theta}}^{(k)} + \frac{1}{\alpha} \sum_{j \in \mathcal{V}} \tilde{\boldsymbol{\theta}}_j$.
10 [optional] compute the reciprocal conditional variance $\{\boldsymbol{\Omega}_j\}_{j \in \mathcal{V}}$ of Gaussian variates (SM G).
11 assess error $err_k = \| \partial \mathcal{L}_{\text{penPL}} / \partial \boldsymbol{\theta} \big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{k+1}} \|_F$ and set $k = k + 1$.
12 **end**

Algorithm 1: Pseudocode of the parallel block-wise Newton-Raphson algorithm for evaluation of the maximum penalized pseudo-likelihood estimator.

As comparison, other work uses either the pseudo-likelihood or a node-wise regression for optimization. The pseudo-likelihood method has previously been reported to be computationally intensive with slow algorithms (Chen et al. 2015). For instance, the computational complexity of pseudo-likelihood maximization is $O(p^6)$ per step for a naive implementation of the Newton-Raphson algorithm. When maximizing the pseudo-loglikelihood, existing methods therefore use a diagonal Hessian or an approximation thereof, or only first-order information (Höfling and Tibshirani 2009; Lee and Hastie 2013). Such approaches achieve linear convergence at best and have a computational complexity of at least $O(np^2)$ per step as the gradient of the pseudo-loglikelihood must be evaluated. Alternatively, the computational complexity of node-wise regression methods is $O(p^4)$ per step for existing algorithms, which could be optimized to $O(p^3)$ with a parallel implementation. However, node-wise regression methods estimate each parameter twice and subsequently need to aggregate their node-wise estimates. This aggregated estimate does not exhibit quadratic convergence. Moreover, these node-wise estimates are potentially contradictory and their quality depends on the type of the variable (Chen et al. 2015).

In short, we expect Algorithm (1) to perform no worse than other pseudo-likelihood maximization approaches, since its computational complexity of $O(p^3)$ is comparable or better than existing methods and all available second-order information is used.

We can, in addition to the pairwise MRF distribution parameter, analytically estimate the variance of the Gaussian variates from the pseudo-loglikelihood (SM G). We perform this additional estimation at the end of each parallel update of the algorithm (line 9, Algorithm 1). This allows the variance of Gaussian variates to be unknown and also aids the intuitive understanding of the estimated parameter as follows. Suppose that we have a multivariate Gaussian distribution with precision matrix $\boldsymbol{\Omega}$. The off-diagonal elements of the MRF distribution parameter $\boldsymbol{\Theta}$ correspond to the off-diagonal of $\boldsymbol{\Omega}$. The elements of the diagonal of $\boldsymbol{\Omega}$ represent the reciprocal of the conditional variances. In contrast, and by definition of the pairwise MRF distribution, the diagonal of $\boldsymbol{\Theta}$ represents the marginal mean of the variates. This non-intuitive relationship between the precision matrix $\boldsymbol{\Omega}$ and parameter $\boldsymbol{\Theta}$ is remedied by substituting the diagonal elements of $\boldsymbol{\Theta}$ corresponding to Gaussian variates with the reciprocal of the conditional (estimated) variances. Then, if the data consist of only Gaussian variates, the algorithm estimates the precision matrix, and additionally returns the estimated means as intuitively expected. This extends to data of mixed types.

Finally, the condition on the multiplier α in Theorem 3 may be relaxed when using the ridge penalty (cf. Lemma 1), thereby appropriately increasing the step size of the parameter update and the convergence speed of Algorithm 1.

Lemma 1 *Let* $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^n$ *be p-variate independent draws from an exponential family distribution* $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})]$. *Let* $\boldsymbol{\theta}^{(0)}$ *be any initial parameter unequal to the maximum ridge pseudo-likelihood estimator* $\widehat{\boldsymbol{\Theta}}^{\text{ridge}}(\lambda)$. *A single step of Algorithm (1) initiated with* $\boldsymbol{\theta}^{(0)}$ *yields the block solutions* $\{\boldsymbol{\theta}_j\}_{j \in \mathcal{V}}$ *(Line 5, Algorithm 1) and block-wise updates* $\{\tilde{\boldsymbol{\theta}}_j\}_{j \in \mathcal{V}}$ *(Line 6, Algorithm 1). Let* $\alpha > 0$ *and define* $\boldsymbol{\theta}^{(1)}$ *as* $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \frac{1}{\alpha} \sum_{j \in \mathcal{V}} \tilde{\boldsymbol{\theta}}_j$. *Next, define the p-dimensional difference vectors* $\{\boldsymbol{\delta}_j\}_{j \in \mathcal{V}}$ *with elements:*

$$(\boldsymbol{\delta}_j)_{j'} = \begin{cases} (\boldsymbol{\theta}_{j'})_j - (\boldsymbol{\theta}_j)_{j'} & \text{if } \; j \neq j' \\ -(\boldsymbol{\theta}_j)_{j'} & \text{if } \; j = j', \end{cases} \tag{8}$$

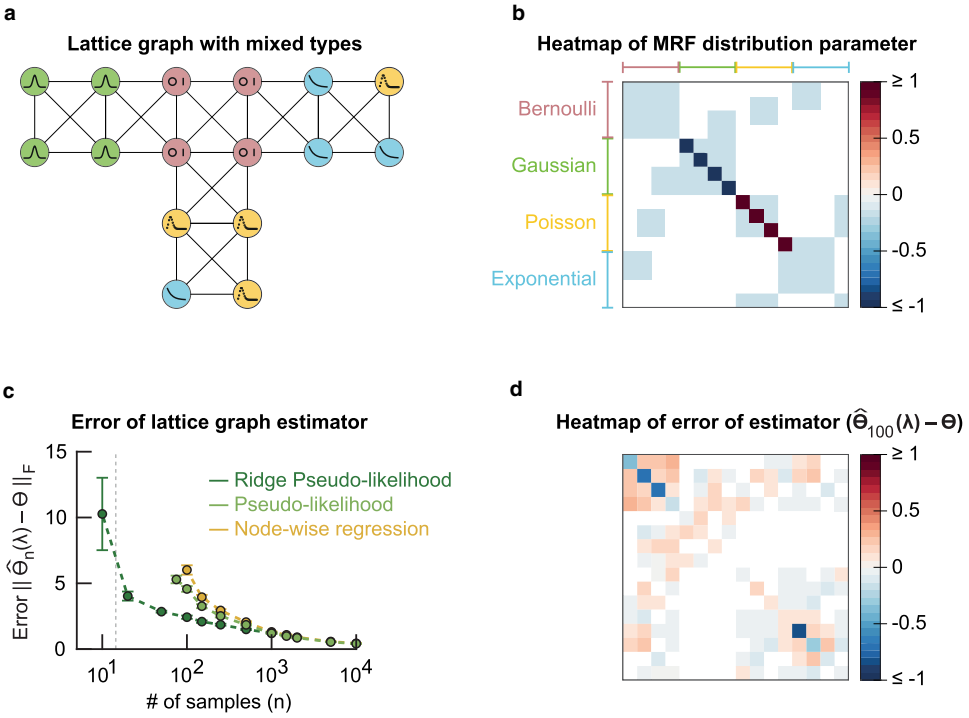

Fig. 2 Pseudo-likelihood estimator of a lattice graph having Bernoulli, Gaussian, Poisson and Exponential data types. **a** The synthetic simulation lattice graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $p = 16$ nodes. Variates have a conditional distribution represented by the shape of the node's pictogram. Distributions are GLM family members Gaussian (green), Bernoulli (red), Poisson (yellow) and exponential (blue). **b** Heatmap of the pairwise MRF distribution parameter. The parameter has $\frac{1}{2}(p+1) \cdot p = 136$ elements allowing for 120 edges (off-diagonal). The parameter constraints reduce the number of allowed edges to 88, with 36 edges present. Diagonal elements for Gaussian variates represent the variance (specifically, the negative reciprocal of the conditional variances). **c** Scaling of the error of the estimator with the sample size. Included are the cross-validated ridge pseudo-likelihood ($k = 10$, dark green), its unpenalized counterpart (light green) and the 'averaged' node-wise regression coefficients (yellow). All curves show mean ± standard error of the mean (20 replicates per condition). (**d**) Heatmap of the error of a cross-validated ridge pseudo-likelihood estimator for $n = 100$. Also see SM M, Figures S1-S4

for all $j, j' \in \mathcal{V}$. Let $L(\cdot; \boldsymbol{\theta}^{(0)})$ be the second-order Taylor approximation of $\mathcal{L}_{\text{penPL}}$ at $\boldsymbol{\theta}^{(0)}$. Then $L(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) > L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)})$ if,

$$\alpha \geq \alpha_{\min} = 3 + \frac{3}{2} \cdot \frac{\sum_{j \in \mathcal{V}} \boldsymbol{\delta}_j^\top \mathbf{H}_j \boldsymbol{\delta}_j}{\sum_{j \in \mathcal{V}} \boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j}, \tag{9}$$

where $\mathbf{H}_j$ is the j-th block Hessian matrix.

Proof Refer to SM H. □

Lemma 1 presents a lower bound $\alpha_{\min} > 3$ on α which warrants, when $\alpha > \alpha_{\min}$, an increase of the penalized pseudo-likelihood $\mathcal{L}_{\text{penPL}}(\boldsymbol{\Theta}, \{\mathbf{Y}_i\}_{i=1}^n)$ at each step of Algorithm 1. In practice, we used $\alpha_{\min}$ throughout as it significantly speeds up the convergence of Algorithm 1. Similarly, we noticed that updating the diagonal elements of the pairwise MRF distribution parameter $\boldsymbol{\Theta}$ more than once could also enhance convergence (see SM H, Corollary 5 for details).

4.1 Implementation

We implemented Algorithm 1 in C++ using the OpenMP API that supports multi-threading with a shared memory. For convenience of the user, the algorithm is wrapped in an R-package as extension for the R statistical computing software. To ensure the estimated parameter always produces a well-defined pairwise MRF distribution, the constraints on the parameter space are implemented using additional convex border functions (SM I). The package includes some auxiliary functions such as a Gibbs sampler to draw samples from the pairwise MRF distribution (SM J) and k-fold cross-validation to select the penalty parameter λ for the maximum ridge pseudo-likelihood estimator. The simplicity, generality and good prediction performance of k-fold cross-validation make it a natural choice for ridge-type estimators that do not induce sparsity (SM K), although we considered sparsification procedures (SM L). The package is made publicly available on GitHub.

5 Simulations

In a numerical study with synthetic data, we evaluate the performance of the proposed Algorithm 1 for numerical evaluation of the maximum ridge pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n(\lambda_{\text{opt}})$ of parameter $\boldsymbol{\Theta}$. We also assess the quality of $\widehat{\boldsymbol{\Theta}}_n(\lambda_{\text{opt}})$ using the convex and twice differentiable ridge penalty $\|\boldsymbol{\Theta}\|_F^2$, leaving the diagonal of $\boldsymbol{\Theta}$ unpenalized (as recommended by Höfling and Tibshirani 2009). Unless stated otherwise, we use threshold $\tau = 10^{-10}$ and multiplier $\alpha = \alpha_{\min}$ for Algorithm 1.

5.1 Performance Illustration

We illustrate the capabilities of the estimator and our algorithm with a simulation of a lattice graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, thus following Yang et al. 2014; Lee and Hastie 2013, and Chen et al. 2015. The lattice graph's layout represents the most general setting encompassed by the outlined theory. Each GLM family member is present with an equal number of (four) variates (Fig. 2a). In short, the nodes are laid out on a lattice, each node being connected to all of its neighbors (e.g., the Gaussian nodes form a complete subgraph, and similarly all Bernoulli nodes, or combinations of three Poisson nodes and an Exponential node form complete subgraphs). The interactions between nodes obey the parameter restrictions for well-definedness of the pairwise MRF distribution. The resulting lattice graph for $p = 16$ nodes has $|\mathcal{E}| = 36$ edges, and it contains 40% of all possible edges. Consequently, the nodes have an average degree of 4.5, while correct graphical model selection is no longer guaranteed (asymptotically) when the maximum vertex degree is larger than $\sqrt{p/\log(p)} = \sqrt{16/\log(16)} = 2.4$ (Das et al. 2012). The lattice graph thus represents a setting where previous work on (sparse) graphical models with data of mixed types fails when the sample size is small, relative to the number of parameters. To ensure the resulting pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ adheres to the described lattice graph $\mathcal{G}$, we choose its parameter $\boldsymbol{\Theta}$ as follows (Fig. 2b):

$$\boldsymbol{\Theta}_{j,j'} = \begin{cases} -0.2 & j, j' \in \mathcal{V} \text{ such that } j \neq j' \text{ and } (j, j') \in \mathcal{E}, \\ -0.2 & j, j' \in \mathcal{V} \text{ such that } j = j' \text{ and } Y_j \text{ follows} \\ & \text{either a Bernoulli or an exponential,} \\ 2 & j, j' \in \mathcal{V} \text{ such that } j = j' \text{ and } Y_j \text{ follows} \\ & \text{a Poisson,} \\ 0 & \text{Otherwise.} \end{cases}$$

This parameter choice ensures the pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined and all edges share the same edge weight. Finally, the variance of the conditional Gaussian variates is set to $\sigma^2 = 1$.

We compare the performance—in terms of the error—of the cross-validated ridge pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n(\lambda_{\text{opt}})$ of $\boldsymbol{\Theta}$ to the unpenalized pseudo-likelihood estimator and the averaged node-wise regression coefficients, whenever the sample size allows. The error is defined as the Frobenius norm of the difference between the parameter and its estimate, e.g., $\|\widehat{\boldsymbol{\Theta}}(\lambda_{\text{opt}}) - \boldsymbol{\Theta}\|_F$. Hereto we generate data for $n \in [10, 10^4]$ samples from the 'lattice graph' distribution (SM J). From these data, the estimators are evaluated and their errors calculated (Fig. 2c). The error of the cross-validated ridge pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}(\lambda_{\text{opt}})$ decreases slowly with the sample size n in the low-dimensional regime as expected, while a sharp increase of its error of is observed in a high-dimensional setting. The error of the ridge pseudo-likelihood is generally on a par with its unpenalized counterpart and the node-wise regression in the low-dimensional regime. More refined, both the maximum ridge and unpenalized pseudo-likelihood estimator outperform the averaged node-wise regression for all sample sizes. The full information and simultaneous parameter estimation approaches are thus preferable. Finally, the proposed ridge pseudo-likelihood estimator clearly shows better performance in the sample domain of (say) $n < 150$. Hence, regularization aids (in the sense of error minimization) when the dimension p approaches or exceeds the sample size n.

To gain further insight in the quality of the estimator, we compute the per-element error of the parameter. This is visualized by means of a heatmap of the estimator's error for a representative example (Fig. 2d). The Bernoulli variates have the largest per-element error in the (ridge) pseudo-likelihood estimator, predominantly amongst Bernoulli-Bernoulli interactions. This is observed across sample sizes. This is intuitive as precise estimation of the parameter of a Bernoulli distribution requires a larger sample size than that of (say) the exponential distribution. Although the error of all types of pairwise interactions decreases with sample size (SM M, Figure S1), the relative contribution of each type of pairwise interaction to the error remains surprisingly constant (SM M, Figure S2). Thus, an increase of the sample size reduces the per-element error for any interaction type but leaves their relative contributions to the total error approximately unaltered.

We also study model selection via the lasso penalty and compare the errors of the ridge and lasso pseudo-likelihood estimators (SM M, Figures S3 and S4).

5.2 Comparison

In a Gaussian graphical model context, we compare the performance of the proposed maximum (ridge) pseudo-likelihood estimator to that of the ridge precision matrix estimator (van Wieringen and Peeters 2016). The latter assumes normality, $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$, and estimates $\boldsymbol{\Omega}$ through ridge penalized likelihood maximization. The maximum ridge pseudo-likelihood estimator too estimates $\boldsymbol{\Omega}$, but does so in a limited information approach. Here we compare

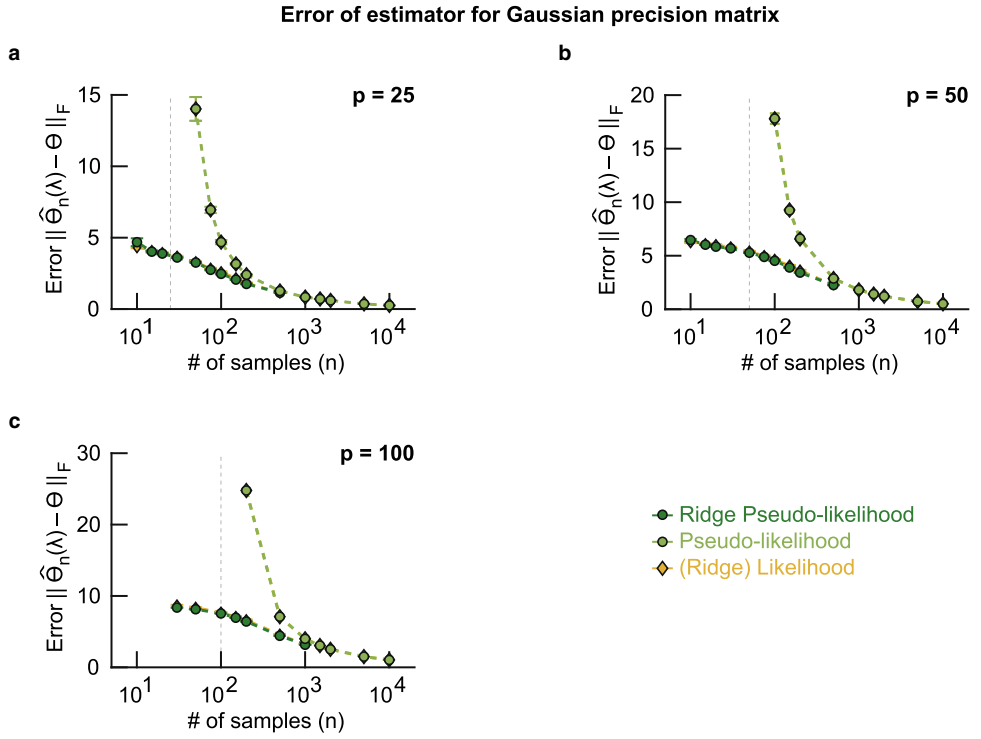

Fig. 3 Scaling of the error of the precision matrix estimator with the sample size. Compared are the errors of the cross-validated maximum ridge pseudo-likelihood estimator (dark-green), its unpenalized counterpart (light-green), the maximum ridge likelihood estimator (yellow) and its unpenalized counterpart (yellow). Shown is the error for dimensions $p = 25$ (**a**), $p = 50$ (**b**) and $p = 100$ (**c**). All curves show mean $\pm$ standard error of the mean (10 replicates for $n = 10^4$, 20 replicates per condition for $n < 10^4$). Also see SM M, Figure S5

the quality of these full and limited information approaches *in silico*. Define a three-banded precision matrix $\mathbf{\Omega}$ with a unit diagonal, $\mathbf{\Omega}_{j,j+1} = 0.5 = \mathbf{\Omega}_{j+1,j}$ for $j = 1, \ldots, p-1$, $\mathbf{\Omega}_{j,j+2} = 0.2 = (\mathbf{\Omega}_{j+2,j}$ for $j = 1, \ldots, p-2$, $\mathbf{\Omega}_{j,j+3} = 0.1 = \mathbf{\Omega}_{j+3,j}$ for $j = 1, \ldots, p-4$, and all other entries equal to zero. The number of variates p ranges from $p = 25$ to $p = 150$ to test the performance of the proposed estimator for its intended use in the context of a large number of variates. Data are sampled from the thus defined multivariate normal $\mathcal{N}(\mathbf{0}_p, \mathbf{\Omega}^{-1})$ and used to evaluate both the maximum (ridge) likelihood and (ridge) pseudo-likelihood estimators for various sample sizes n.

We compare the performance of the precision estimators by means of their error, defined as $||\widehat{\mathbf{\Omega}}(\lambda_{\text{opt}}) - \mathbf{\Omega}||_F$, and analogously for their unpenalized counterparts. In the low-dimensional regime, for $p = 25$ and $n > 100$, the errors of all estimators are very close and decrease slowly as the sample size n increases (Fig. 3a). Specifically, the maximum unpenalized likelihood and maximum unpenalized pseudo-likelihood estimators are identical for all sample sizes and all data sets, resulting in identical errors. In the high-dimensional regime, for $p = 25$ and $n < 100$, the penalized estimators clearly outperform their unpenalized counterparts as can be witnessed from their diverging error when n approaches p. Moreover, the maximum ridge pseudo-likelihood estimator appears to slightly outperform its maximum ridge likelihood counterpart. This is probably due to the penalty or implementation of the cross-validation methods, as both estimates are generally very close. This corroborates the results of previous simulation studies into the maximum (lasso) penalized pseudo-likelihood estimator (Lee and Hastie 2013; Höfling and Tibshirani 2009). With the application of large data sets and parallel computing in mind, we consider the performance of the maximum penalized pseudo-likelihood estimator for a higher number of variates up to $p = 150$ next (Fig. 3b–d). Generally, while an increase of the dimension p increases the error of the estimators, qualitatively their relative behavior remains largely unchanged. Specifically, (i) the errors of all estimators are very close in the low-dimensional regime, (ii) the unpenalized likelihood and unpenalized pseudo-likelihood estimators are identical for all dimensions and sample sizes, (iii) the penalized estimators outperform their unpenalized counterparts in the high-dimensional regime, and (iv) the errors of the maximum ridge pseudo-likelihood estimator are very close to the errors of the maximum ridge likelihood estimator. We further study the error of the penalized estimators as function of the degree of nodes in the underlying graph (SM M, Figure S5).

5.3 Speed-up and benchmark

Here we pursue to speed up Algorithm 1 by further reducing its computational complexity. To this end, we modified the parallel block-wise Newton–Raphson to a block-wise quasi-Newton approach using a chord method that computes the inverse of the block-wise Hessian matrices $\{\mathbf{H}_j\}_{j\leq p}$ only every $k_0 = p$ steps of the algorithm. This vastly reduces the computational complexity of the algorithm—without significantly reducing convergence— by alleviating the burden of the rate-limiting substep (SM M, Figure S6).

We benchmark the proposed algorithm by studying its run time and the required number of steps. For this we consider a pairwise MRF distribution having variates of the binary (Bernoulli) and continuous (Gaussian) type. The two types are equally represented among the p variates. The parameter $\boldsymbol{\Theta}$ of this distribution is chosen such that it satisfies the parameter restrictions for well-definedness of the pairwise MRF distribution (SM M, Figure S7). The conditional precision matrix of the Gaussian variates is three-banded as in the previous comparison study. Each Bernoulli variate has an interaction with every $\sqrt{p}$-th other variate, and the corresponding interaction parameters are set equal to ± 0.1, in alternating fashion. Data with $n = 1.000$ samples with a dimension ranging from $p = 16$ to $p = 200$ variates are sampled from this mixed Bernoulli-Gaussian distribution and used for benchmarking.

First we study the effect of the step size of the algorithm on convergence of the parameter estimate. The update changes the estimate $\hat{\boldsymbol{\theta}}^{(k)} \rightarrow \hat{\boldsymbol{\theta}}^{(k+1)}$ with $\frac{1}{\alpha}\sum_{j\in\mathcal{V}}\hat{\boldsymbol{\theta}}_j$ (Line 8 of Algorithm 1). Thus, the step size is proportional to α^{-1}. Conventionally, one would take $\alpha = p$ and naively average the block-wise updates (a convex combination) to update $\hat{\boldsymbol{\theta}}^{(k)}$. We showed that the approach of combining block updates theoretically allows for a larger step $\alpha < p$ (Lemma 1). Indeed, we find that increasing the step size—decreasing α—reduces the number of required steps for Algorithm 1 to find the estimator (Fig. 4a). Specifically, doubling the step size approximately halves the number of required steps for all dimensions p. However, a step size that is too large prevents convergence of the parameter estimate (hence the endpoints of the curves at $\alpha < p$). Lemma 1 circumvents this problem by evaluating an (in some sense) optimal $\alpha_{\min}$ at every step of the algorithm and using this $\alpha_{\min}$ for the update of that step. This approach further reduces the required number of steps to find the estimator (diamonds in Fig. 4a).

We next assess the effect of having multiple processors to compute the parallel part of Algorithm 1 (the gradient and inverse of the block-wise Hessian matrices). Doubling the number of processors approximately halves the required time per step of the algorithm (Fig. 4b), especially at high dimensions (e.g., $p \geq 64$). This is expected as the rate-limiting substep (inverting the Hessians) increasingly dominates the run time as the problem size increases, and almost perfectly parallelizes with the number of processors. Note that a single processor computes everything sequentially, but an update for $\hat{\boldsymbol{\theta}}^{(k)}$ still represents the aggregated block-wise updates.

With both the number of steps and the time per step optimized, we compare the performance of our proposed algorithm with naive approaches. Hereto we compute the time required to find the maximum pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n$ for three methods: using (i) Newton-Raphson, (ii) sequential block-wise and (iii) parallel block-wise algorithms. The Newton–Raphson approach computes and inverts the full $q \times q$-dimensional Hessian matrix (with $q = \frac{1}{2}(p+1)p$). The sequential block-wise approach sequentially picks variates $j \in \{1, \ldots, p\}$ and then only updates the block of elements $\boldsymbol{\Theta}_{j,*}$ of the parameter estimate, thus inverting only the j-th $p \times p$-dimensional block-wise Hessian $\mathbf{H}_j$ at a given step. The parallel approach is Algorithm 1, inverting all $p \times p$-dimensional Hessians $\{\mathbf{H}_j\}_{j\leq p}$ in parallel and aggregating the block-wise estimates using a step size determined by $\alpha_{\min}$. Note that using the Newton-Raphson algorithm that inverts the full Hessian at each step is computationally too intensive, while a diagonal Hessian requires too many steps to convergence (see SM M, Figure S6 for a comparison). For a fair comparison each approach inverts their respective Hessian matrices only every $k_0 = p$-th step.

Algorithm 1 outperforms the other approaches, especially for large problem sizes (Fig. 4c). This is also independent of whether the variance of Gaussian variates is estimated or not (SM M, Figure S7). Specifically, sequential block-wise is the slowest approach for all dimensions p as the number of required steps increases very fast with p (SM M, Figure S7). The Newton–Raphson approach, requiring very few steps, is fastest for small dimensions $p < 20$, although the computational complexity of inverting the full Hessian quickly becomes prohibitively large (for $p > 20$). To appreciate the computational efficiency of the parallel algorithm, note that one step of the full Newton–Raphson algorithm has computational complexity $O(p^6)$ compared to $O(p^3)$ of the parallel Algorithm 1. This permits the latter $O(p^3)$ steps before exceeding the computational complexity of one step of full Newton–Raphson. Indeed, the parallel algorithm was found to always terminate within $O(p^3)$ steps. In terms of actual run time, the parallel algorithm typically finds the estimator (converges with threshold $\tau = 10^{-10}$) in under one minute for $p = 150$. In contrast, the sequential approach already takes over 6 minutes for $p = 64$, while Newton–Raphson takes over 20 minutes for $p = 100$. In summary, the proposed parallel algorithm is orders of magnitude faster compared to naive approaches for large dimensions $p > 100$ by (i) reducing the required number of steps and (ii) reducing the time taken per step for the algorithm to find the estimator (Fig. 4d).

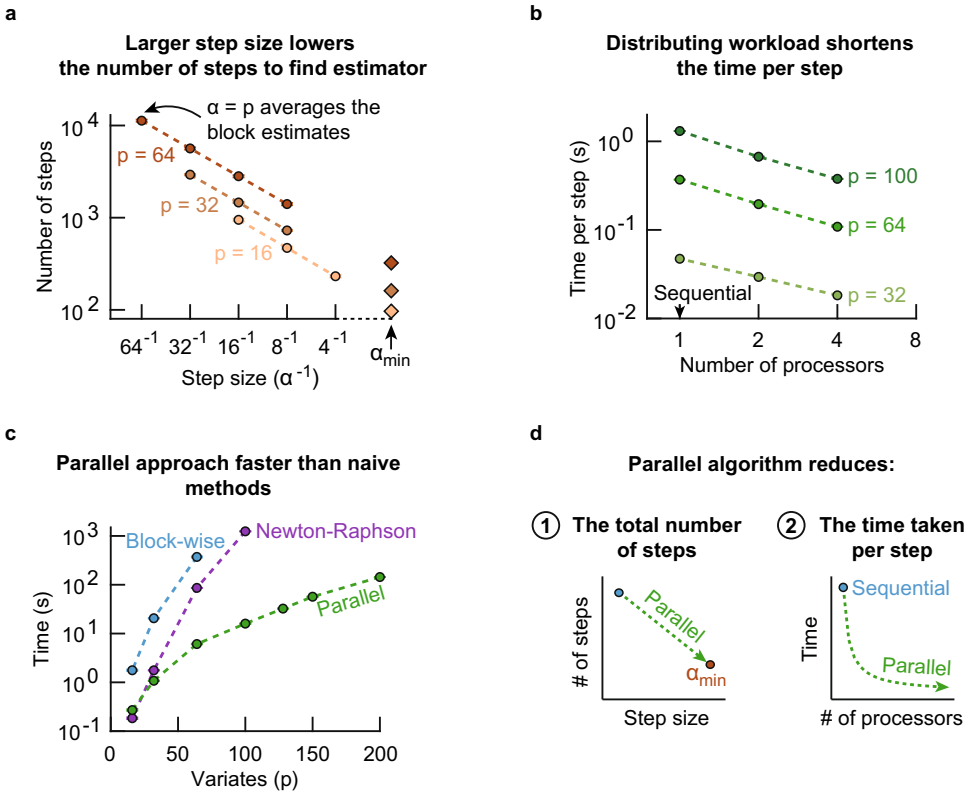

Fig. 4 Parallel algorithm outperforms naive algorithms. **a** The number of steps required for Algorithm 1 to find the estimator as function of the step size (α^{-1}). Shown are dimensions $p = 64$ (dark red), $p = 32$ (red) and $p = 16$ (orange). A step size p^{-1} is equivalent to naively averaging the block-wise estimates. The optimal (average) $\alpha_{\min}$ from Lemma 1 for each dimension is indicated with a diamond. **b** The time per step of Algorithm 1 as function of the number of available processors. Shown are dimensions $p = 100$ (dark green), $p = 64$ (green) and $p = 32$ (light green). One processor corresponds to performing all computations sequentially. **c** The time taken to find the estimator for the naive Newton–Raphson algorithm (purple), a sequential block-wise algorithm (blue) and the parallel Algorithm 1 (green) as function of the number of variates p. The inverse Hessian matrices were computed every $k = p$ steps to reduce computation time for each method. **a–c** All curves show mean ± standard error of the mean (5 replicates per condition). Processor cores were artificially limited to 2 GHz for a fair comparison (SM N). **d** Schematic summary. The parallel approach reduces the time to find the estimator by (1) lowering the required number of steps (scaling with step size) and (2) shortening the time taken per step (scaling with number of processors). Also see SM M, Figures S6 and S7

6 Conclusion

We presented methodology for the maximum penalized pseudo-likelihood estimation of multivariate exponential family distributions. As special case of interest, the employed class of distributions encompasses the pairwise Markov random field that describes stochastic relations among variates of various types. The presented estimator was shown to be consistent under mild conditions. Our algorithm for its evaluation allows for efficient computation on multi-core systems and accommodates for a large number of variates. The algorithm was shown to converge and terminate. A simulation study showed that the performance of the proposed (ridge-penalized) pseudo-likelihood estimator was very close to the maximum ridge likelihood estimator. Moreover, our benchmark showed that the proposed parallel algorithm is superior to naive approaches. Finally, our methodology was demonstrated with an application to an integrative omics study using data from various molecular levels (and types) (see SM O).

Envisioned extensions of the presented ridge pseudo-likelihood estimator allow—among others—for variate type-wise penalization. Technically, this is a minor modification of the algorithm but brings about the demand for an efficient penalty parameter selection procedure. Furthermore, when quantitative prior information of the parameter is available it may be of interest to accommodate shrinkage to nonzero values.

Foreseeing a world with highly parallelized workloads, our algorithm provides a first step towards a theoretical framework that allows for efficient parallel evaluation of (high-dimensional) estimators. Usually and rightfully most effort concentrates on the mathematical optimization of the computational aspects of an algorithm. Once that has reached its limits, parallelization may push further. This amounts to simultaneous estimation of parts of the parameter fol-

lowed by careful—to ensure convergence—recombination to construct a fully updated parameter estimate. Such parallel algorithms may bring about a considerable computational gain. For example, in the presented case this gain was exploited to incorporate full second-order information without inferior computational complexity compared to existing algorithms.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-021-10013-x.

References

Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Methodol. **36**(2), 192–236 (1974)

Bilgrau, A.E., Peeters, C.F.W., Eriksen, P.S., Bøgsted, M., van Wieringen, W.N.: Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. J. Mach. Learn. Res. **21**(26), 1–52 (2020)

Boyle, E.A., Li, Y.I., Pritchard, J.K.: An expanded view of complex traits: from polygenic to omnigenic. Cell **169**(7), 1177–1186 (2017)

Chen, S., Witten, D.M., Shojaie, A.: Selection and estimation for mixed graphical models. Biometrika **102**(1), 47–64 (2015)

Das, A K., Netrapalli, P., Sanghavi, S., Vishwanath, S.: Learning Markov graphs up to edit distance. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2731–2735. IEEE, 2012

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)

Fletcher, R.: Practical Methods of Optimization, 2nd edn. Wiley, Hoboken (2013)

Gallagher, A C., Batra, D., Parikh, D.: Inference for order reduction in Markov random fields. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1857–1864. IEEE, 2011

Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **2**(278), 563–577 (2015)

Hammersley, J M., Clifford, P.: Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971

Höfling, H., Tibshirani, R.: Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. J. Mach. Learn. Res. **10**, 883–906 (2009)

Lauritzen, S.: Graphical Models. Oxford University Press, Oxford (1996)

Lee, J., Hastie, T.: Learning the structure of mixed graphical models. J. Comput. Graph. Stat. **24**(1), 230–253 (2013)

Lee, J.D., Sun, Y., Taylor, J.: On model selection consistency of regularized M-estimators. Electron. J. Stat. **9**(1), 608–642 (2015). https://doi.org/10.1214/15-EJS1013

Miok, V., Wilting, S.M., van Wieringen, W.N.: Ridge estimation of the var (1) model and its time series chain graph from multivariate time-course omics data. Biom. J. **59**(1), 172–191 (2017)

van Wieringen, W.N.: The generalized ridge estimator of the inverse covariance matrix. J. Comput. Graph. Stat. **28**(4), 932–942 (2019)

van Wieringen, W.N., Peeters, C.F.W.: Ridge estimation of inverse covariance matrices from high-dimensional data. Comput. Stat. Data Anal. **103**, 284–303 (2016)

Welsh, D.J.A.: Complexity: Knots. Cambridge University Press, Colourings and Counting (1993)

Wild, C.P.: The exposome: from concept to utility. Int. J. Epidemiol. **1**(41), 24–32 (2012)

Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**(3), 1758–1789 (2013)

Yang, E., Baker, Y., Ravikumar, P., Allen, G., Liu, Z.: Mixed graphical models via exponential families. Artif. Intel. Stat. **33**, 1042–1050 (2014)