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

Supplementary Material to: A parallel algorithm for ridge penalized estimation of the multivariate exponential family from data of mixed types

Diederik S. Laman Trip · Wessel N. van Wieringen

Diederik S. Laman Trip
Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands

Wessel N. van Wieringen
Department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam UMC, location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands; and Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
E-mail: w.vanwieringen@amsterdamumc.nl

Supplementary Material A: Related work

There is extensive literature on (learning) graphical models. Here we point to the sources relevant for our study. Attention originally focussed on a set of variates with a continuous state space following a multivariate normal distribution and thus giving rise to the Gaussian graphical model. The penalized estimation approaches of this model either maximize the likelihood augmented with a penalty (be it ℓ_1 or ℓ_2, (Banerjee et al., 2008, Friedman et al., 2008, van Wieringen and Peeters, 2016)), or comprise node-wise penalized regressions (Meinshausen and Bühlmann, 2006, Ravikumar et al., 2011). The Ising model is the classic case for graphical models with discrete variables. Its estimation is prohibited by the computationally intractability of its partition function, which is generally the case for distributions of variates with a finite state space (e.g. Bernoulli). This may be circumvented by a stringent sparsity assumption (Höfling and Tibshirani, 2009, Lee et al., 2007). Alternatively, an approximation to the partition function gives rise to approaches that amount to node-wise regressions (Ravikumar et al., 2010, Wainwright et al., 2007, Friedman et al., 2010). This approach is extended to node-conditional distributions of multinomial or Poisson distributed variates (Jalali et al., 2011, Allen and Liu, 2013).
Estimation with pseudo-likelihood. A final alternative would be to maximize the penalized pseudo-loglikelihood (Höfling and Tibshirani, 2009, building on Besag, 1974), which has the advantage over the node-wise regression of producing a single parameter estimate for the whole graphical structure instead of multiple (possible contradictory) local ones. Moreover, reported simulations indicate that the resulting estimate is close to its penalized maximum likelihood counterpart (Höfling and Tibshirani, 2009).
Variates of mixed type. Originating in Besag [1974], work on graphical models of mixed type focusses on the derivation of the MRF distribution by the specification of the node-conditional distributions. Lauritzen [1996] and Lauritzen and Wermuth [1989] produced the first examples for combinations of Bernoulli and Gaussian variates. Later, Yang et al. [2012] extended this work and specified the node-conditional distribution of each variate as an univariate exponential family member (all of the same type), from which a corresponding joint MRF distribution was derived. These univariate exponential families include multi-parameter members with all but one parameter constant (e.g. Gaussian variables with constant variance). Subsequently, Yang et al. [2012] estimate the MRF parameter by a limited information approach exploiting the node-conditional distributions. Yang et al. [2014] further generalized their work on MRF distributions to allow variates of mixed type with node-conditional distribution of any univariate exponential family member. An maximum ℓ_1 penalized pseudo-likelihood estimator for this model, limited to a combination of Bernoulli and Gaussian variates, was put forward by Lee and Hastie [2013]. In parallel effort, Chen et al. [2015] derived a similar joint MRF distribution that also includes multi-parameter exponential family members (e.g. Gaussian variables with unknown mean and variance), and specify conditions on its parameter space. However, after model specification, existing works (Yang et al., 2014 and Chen et al., 2015) consider only one-parameter variates of at most two different types in their estimation procedure.

Supplementary Material B: Motivation for ridge penalization

Our motivation for ridge penalized estimation is multifold:

1. For starters, it fills the lacuna in the literature which is mainly orientated towards lasso penalization.
2. Ridge estimators are, in contrast to lasso estimators, unique, due to their strict convex penalty.
3. For ridge estimators, due to the smoothness and strict convexness of the ridge penalty, either an analytic expression of the estimator exists or a stable algorithm for the evaluation the corresponding estimator is conceivable. Algorithms for the evaluation of lasso-type estimator, however, often exhibit convergence problems, in particular when either the model is not extremely sparse, the data are medium to high-dimensional, and/or the penalization is not severe.
4. Biostatistical folkore has it that ridge estimators generally yield a better fit than those of lasso-type. This has also been observed in the graphical model context by for example van Wieringen and Peeters [2016], Miok et al. [2017] or Bilgrau et al. [2020]. The simulations presented in those works also show that ridge estimation combined with posthoc selection generally performs well – if not better – compared to those of lasso estimation.
5. The use of a ridge penalty translates the existing lasso-associated framework of sparse network estimation by allowing for estimation of dense networks. The need for such procedures can be found in the fact that the dominant paradigm of sparsity is not necessarily valid in all fields of application. In particular, in molecular biology this paradigm has recently come under fire (Boyle et al., 2017), and more dense (graphical) structures are advocated. Moreover, the severe degree of sparsity required by the theoretical resulting underpinning ℓ_1 estimation procedures may lead to a too stringent inference on the underlying graph, potentially oversimplifying. Therefore an ℓ_2 estimation procedure might be more appropriate, especially in molecular biology applications, as it makes no sparsity assumption.
6. The smoothness and strict convexity of the ridge penalty can be used to approximate other penalties (as was first pointed by Fan and Li, 2001). Within the graphical model context we have used this ourselves to derive a generalized lasso/elastic net estimator of the precision matrix of a Gaussian graphical model (see van Wieringen, 2019).

Supplementary Material C: Parameter constraints

j \ j'	Bernoulli	Poisson	Gaussian	Exponential
Bernoulli	$\boldsymbol{\Theta}_{j,j'} \in \mathbb{R}$	$\boldsymbol{\Theta}_{j,j'} \in \mathbb{R}$	$\boldsymbol{\Theta}_{j,j'} \in \mathbb{R}$	$\sum_{j \in \mathcal{V}_B} \max\{\boldsymbol{\Theta}_{j,j'}, 0\} < -\boldsymbol{\Theta}_{j,j'}$
Poisson		$\boldsymbol{\Theta}_{j,j'} \leq 0$	$\boldsymbol{\Theta}_{j,j'} = 0$	$\boldsymbol{\Theta}_{j,j'} \leq 0$
Gaussian			$\boldsymbol{\Theta}_G \prec 0$	$\boldsymbol{\Theta}_{j,j'} = 0$
Exponential				$\boldsymbol{\Theta}_{j,j'} \leq 0$

Table 1 Parameters constraints for a well-defined pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ for GLM family members, for completeness reproduced from Chen et al. [2015]. More general results for exponential family members may be derived, see e.g. Yang et al. [2014]. Here, $\boldsymbol{\Theta}_G$ refers to the submatrix of $\boldsymbol{\Theta}$ that correspond to variates that are conditionally Gaussian distributed, while $\mathcal{V}_B \subset \mathcal{V}$ refers the subset of variates with a conditional Bernoulli distribution. In additional to the constraints in the table the rate parameters of variates with a conditional exponential distribution are required to be negative: $\boldsymbol{\Theta}_{j,j} < 0$. If desired, these restrictions can be relaxed by modeling data with, for example, a truncated Poisson distribution (Yang et al., 2014), or, while possibly not resulting in a well-defined joint distribution, might even be ignored altogether for network estimation (Chen et al., 2015).

Supplementary Material D: Motivation of traditional consistency

The motivation for a traditional consistency result, i.e. in the p fixed while $n \to \infty$ regime, is three-fold.

1. It suffices for practical purposes in the 'omics-field' where the system's dimensions are known and fixed at the outset of the analysis.
2. Almost all existing high-dimensional results (cf. e.g. Lee and Hastie, 2013, Chen et al., 2015, Lee et al., 2015) show 'sparsistency' – a contraction of sparse and consistency – which amounts to the probability of correct (sparse) network structure selection tending to one as $p \to \infty$. Sparsistency thus addresses the selection property of an estimator but not the (limiting) quality of the estimator in terms of the parameter values. In this light sparsistency results have – understandably – only been proven for ℓ_1-penalized estimators (or a generalization thereof, e.g. Lee et al., 2015). Under the assumptions of *i)* strong convexity of the loss function and *ii)* irrepresentability of the model (for details see Lee et al., 2015), sparsistency is proven for sparse models (with the maximum vertex degree d such that $d = o(n)$, Chen et al., 2015), with a minimal parameter value (in the absolute sense) for those parameters being nonzero, and a minimum sample size dependent on the dimension p and the degree of sparsity (Lee et al., 2015). These assumptions and requirements are hard – if not impossible – to verify for any practical case at hand (Bento and Montanari, 2009, Anandkumar et al., 2012), thus questioning the usefulness of sparsistency results (Lee and Hastie, 2013). Moreover, for the proposed ridge estimator, which does not select, sparsistency appears not to be the relevant concept.
3. An information theoretic analysis shows that learning graphical models, even for simple cases such as the Ising or Gaussian graphical model, requires at least $n = \Omega(d^2 \log p)$ samples (in words, n is at least of order $d^2 \log p$) as $p \to \infty$ (Das et al., 2012). For example, when $p > n$, one finds that $p > n > d^2 \log p$. This, when solved for d, yields: $d < \sqrt{p/\log p}$. For $p = 100$ the maximum vertex degree d would then be bounded by 4.6, which rules out dense networks. Therefore, there is no hope for successful graphical model selection in the high-dimensional setting ($n < p$) when the maximum vertex degree satisfies $d = \Omega(\sqrt{p/\log p})$, which can be assumed to be the case for molecular biology applications or scale-free and dense networks in general.

Hence, the focus on traditional consistency which is a minimum requirement of a novel estimator.

Supplementary Material E: Consistency of the penalized pseudo-likelihood estimator (proof Theorem 2 of the main text)

Here we study the consistency of the penalized pseudo-likelihood estimator. This asymptotic behaviour is shown under the assumption of any multivariate exponential family distribution, of the form $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})]$. It is immediate that the pairwise MRF distribution is a special case.

Theorem 1 *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be p-variate draws from a exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})]$. The pseudo-likelihood estimator, $\widehat{\boldsymbol{\Theta}}_n := \arg\max_{\boldsymbol{\Theta}} \mathcal{L}_{PL}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, that maximizes the pseudo-likelihood is consistent, i.e., $\widehat{\boldsymbol{\Theta}}_n \xrightarrow{p} \boldsymbol{\Theta}$, 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 $\beta \in \mathbb{N}$.

The proof of this theorem relies on the following theorem:

Theorem 2 *(Theorem 5.7, Van der Vaart, 1998)*
Let M_n be random functions and let M be a fixed function of $\boldsymbol{\Theta}$ such that for every $\varepsilon > 0$,

$$\sup_{\boldsymbol{\Theta}} \big| M_n(\boldsymbol{\Theta}) - M(\boldsymbol{\Theta}) \big| \xrightarrow{p} 0, \tag{1}$$

$$\sup_{\boldsymbol{\Theta}: d(\boldsymbol{\Theta}, \boldsymbol{\Theta}_0) \geq \varepsilon} M(\boldsymbol{\Theta}) < M(\boldsymbol{\Theta}_0). \tag{2}$$

Then any sequence of estimators $\widehat{\boldsymbol{\Theta}}_n$ with $M_n(\widehat{\boldsymbol{\Theta}}_n) \geq M_n(\boldsymbol{\Theta}_0) - o_P(1)$ converges in probability to $\boldsymbol{\Theta}_0$.

Proof (of Theorem 1)
The proof of Theorem (1) infers Theorem (2), that specifies the conditions for an estimator's consistency. It thus suffices to check these conditions for the estimator at hand:

- The first 'stochastic' condition of Theorem (2) requires that the pseudo-likelihood $\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ converges to its (asymptotic) mean $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$ uniformly. This is shown in Theorem (4) by use of the uniform law of large numbers and the conditions *i)* and *ii)* specified in Theorem (1).
- For the second 'deterministic' condition of Theorem (2) it is sufficient to show that the (asymptotic) mean $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$ is strictly concave. This is shown in Corollary (2) using condition *i)* of Theorem (1). Consequently, $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$ has a global maximum. In particular, but not required, the true parameter $\boldsymbol{\Theta}$ of $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is equal to (one of) these global maxima (cf. Lemma 2).

Finally, let $\widehat{\boldsymbol{\Theta}}_n$ be the unique maximizer of the strictly concave sample pseudo-likelihood $\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Then, by applying Theorem (2), the pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n$ converges to the true parameter $\boldsymbol{\Theta}$ in probability. □

Proof Consistency: Uniform Convergence of the Pseudo-likelihood

Sufficient conditions for uniform convergence (the stochastic condition of Theorem 2) of the pseudo-likelihood $\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ of $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ to its asymptotic mean $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$ are provided. It relies on the uniform law of large numbers (Jennrich, 1969).

Theorem 3 *(Uniform law of large numbers, Jennrich, 1969)*
Suppose that the parameter space of $\boldsymbol{\Theta}$ is compact and let $f(\mathbf{Y}, \boldsymbol{\Theta})$ be a continuous function of $\boldsymbol{\Theta}$ for all $\mathbf{Y}$. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be random draws from $f(\mathbf{Y}, \boldsymbol{\Theta})$. Then:

$$\sup_{\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{Y}_i, \boldsymbol{\Theta}) - \mathbb{E}[f(\mathbf{Y}, \boldsymbol{\Theta})] \right| \xrightarrow{p} 0,$$

if $f(\cdot, \cdot)$ is dominated by a function $b(\mathbf{Y})$ such that $|f(\mathbf{Y}, \boldsymbol{\Theta})| \leq b(\mathbf{Y})$ and $\mathbb{E}[b(\mathbf{Y})] < \infty$.

Theorem 4 *(Uniform convergence of pseudo-likelihood)*
Let $\mathbf{Y}_1, ..., \mathbf{Y}_n$ be n independent random draws from a well-defined p-variate exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})]$. The pseudo-likelihood then converges uniformly to its asymptotic mean:

$$\sup_{\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{PL}(\boldsymbol{\Theta}, \mathbf{Y}_i) - \mathbb{E}_{P_{\boldsymbol{\Theta}}}[\mathcal{L}_{PL}(\boldsymbol{\Theta}, \mathbf{Y})] \right| \xrightarrow{p} 0,$$

if,

i) The parameter space of $\boldsymbol{\Theta}$ *is compact,*
ii) $\boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})$ *can be bounded by a polynomial,* $|\boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})| \leq c_1 + c_2 \sum_{j \in \mathcal{V}} |Y_j|^\beta$ *for constants* $c_1, c_2 < \infty$ *and* $\beta \in \mathbb{N}$.

Proof It suffices to check the conditions of the uniform law of large numbers (Theorem 3): *a)* the continuity of the pseudo-likelihood, and *b)* the existence of a dominating function with finite expectation for the pseudo-likelihood. Let $E(\boldsymbol{\Theta}, \mathbf{Y}) = \boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})$. With respect to condition *a)*, recall the definition of the pseudo-likelihood:

$$\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}) = \sum_{j \in \mathcal{V}} \log[P_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})] \;\; = \;\; \sum_{j \in \mathcal{V}} E(\boldsymbol{\Theta}, Y_j \,|\, \mathbf{Y}_{\setminus j}) - D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j}),$$

where E is linear in $\boldsymbol{\Theta}$. With D_j a logarithmic sum of continuous functions over all possible outcomes, it is itself continuous. Hence, the pseudo-likelihood, being a composition of continuous functions, is continuous in $\boldsymbol{\Theta}$.

Rests condition *b)* to show the existence of a bound $b(\mathbf{Y})$ for $|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})|$ with finite expectation. Existence of a bound is provided by Lemma (1) which tells us that, when assuming conditions *i)* and *ii)*, for all $\boldsymbol{\Theta}$, $|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| \leq c_1 b(\mathbf{Y}) + c_2$ with constants $c_1, c_2 < \infty$ and dominating function $b(\mathbf{Y}) := \sum_{j \in \mathcal{V}} |Y_j|^\beta$. We conclude by checking its finite expectation. By the linearity of expectation we have $\mathbb{E}_{P_{\boldsymbol{\Theta}}}[b(\mathbf{Y})] = \sum_{j \in \mathcal{V}} \mathbb{E}_{P_{\boldsymbol{\Theta}}}\big(|Y_j|^\beta\big)$. The finite cardinality of $\mathcal{V}$ implies that it suffices to verify that $\mathbb{E}_{P_{\boldsymbol{\Theta}}}(|Y_j|^\beta) < \infty$ for all $j \in \mathcal{V}$. Marginalization yields,

$$\mathbb{E}_{P_{\boldsymbol{\Theta}}}\big(|Y_j|^\beta\big) \;\; = \;\; \int |y_j|^\beta P_{\boldsymbol{\Theta}}(y_j) \; dy_j.$$

Note that,

$$|Y_j|^\beta = \int_0^{|Y_j|} \beta \; t^{\beta - 1} \; dt \;\; = \;\; \int_0^\infty \beta \; t^{\beta - 1} \mathbb{1}_{\{t < |Y_j|\}} \; dt,$$

and Tonelli's theorem (Fubini's theorem for non-negative functions) then yields,

$$\begin{aligned} \mathbb{E}_{P_{\boldsymbol{\Theta}}}(|Y_j|^\beta) \;\; &= \;\; \int |y_j|^\beta P_{\boldsymbol{\Theta}}(y_j) \; dy_j \;\; = \;\; \int \int_0^\infty \beta t^{\beta - 1} \mathbb{1}_{\{t < |y_j|\}} P_{\boldsymbol{\Theta}}(y_j) \; dt \; dy_j \\ &\overset{\text{(Tonelli)}}{=} \beta \int_0^\infty t^{\beta - 1} \int \mathbb{1}_{\{t < |y_j|\}} P_{\boldsymbol{\Theta}}(y_j) \; dy_j \; dt \;\; = \;\; \beta \int_0^\infty t^{\beta - 1} P_{\boldsymbol{\Theta}}(|Y_j| > t) \; dt. \end{aligned}$$

For any $\varepsilon > 0$, Markov's inequality yields,

$$P_{\boldsymbol{\Theta}}(|Y_j| > t) \leq e^{-\epsilon t} \cdot \mathbb{E}_{P_{\boldsymbol{\Theta}}}\Big[\exp(\varepsilon |Y_j|)\Big],$$

as the exponential function is non-negative and monotonically increasing on the non-negative reals. Combine the two preceeding displays to arrive at:

$$\mathbb{E}_{P_{\boldsymbol{\Theta}}}\big(|Y_j|^\beta\big) \leq \mathbb{E}_{P_{\boldsymbol{\Theta}}}\Big[\exp(\varepsilon |Y_j|)\Big] \; \beta \int_0^\infty t^{\beta - 1} \exp(-\varepsilon t) \; dt. \tag{3}$$

The integral in (3) evaluates to,

$$\int_0^\infty t^{\beta - 1} \exp(-\varepsilon t) \; dt = \varepsilon^{-\beta} \; \Gamma(\beta) \;\; < \;\; \infty,$$

so we are left to assess that $\mathbb{E}_{P_{\boldsymbol{\Theta}}}[\exp(\varepsilon |Y_j|)] < \infty$ for $j \in \mathcal{V}$. To this end bound the absolute value as,

$$\int P_{\boldsymbol{\Theta}}(\mathbf{Y}) \exp(\varepsilon |Y_j|) \; d\mathbf{Y} \leq \int P_{\boldsymbol{\Theta}}(\mathbf{Y}) \exp(-\varepsilon Y_j) \; d\mathbf{Y} + \int P_{\boldsymbol{\Theta}}(\mathbf{Y}) \exp(\varepsilon Y_j) \; d\mathbf{Y}.$$

Recall that $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is an exponential family distribution, hence log-linear in $\boldsymbol{\Theta}$. The integrals on the right-hand side of the preceeding display can then be rewritten as,

$$\int P_{\boldsymbol{\Theta}}(\mathbf{Y}) \exp(\pm \varepsilon Y_j) \, d\mathbf{Y} = \frac{\exp[D(\tilde{\boldsymbol{\Theta}})]}{\exp[D(\boldsymbol{\Theta})]} \cdot \int P_{\tilde{\boldsymbol{\Theta}}}(\mathbf{Y}) \, d\mathbf{Y} \tag{4}$$

where $\tilde{\boldsymbol{\Theta}}$ is a perturbation of $\boldsymbol{\Theta}$ only in the element that forms the coefficient of the linear Y_j-term in the exponent of the probability distribution such that $\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty < \varepsilon$. Finally, notice that, by well-definedness of the exponential family distribution, the normalization factor $D(\boldsymbol{\Theta}) < \infty$ for any $\boldsymbol{\Theta}$ in the compact parameter space. Then, for any $\boldsymbol{\Theta}$ inside the parameter space, a small enough $\varepsilon > 0$ can be chosen such that the perturbation $\tilde{\boldsymbol{\Theta}}$ is also inside the parameter space, and thus: $D(\tilde{\boldsymbol{\Theta}}) < \infty$. Then the right-hand side of (4) marginalizes and is finite. Hence, $\mathbb{E}_{P_{\boldsymbol{\Theta}}}[\exp(\varepsilon|Y_j|)] < \infty$ for small enough $\varepsilon > 0$.

Put together, we now have that $|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| \leq c_1 b(\mathbf{Y}) + c_2$ with $\mathbb{E}_{P_{\boldsymbol{\Theta}}}[b(\mathbf{Y})] < \infty$. By Theorem (3) the pseudo-likelihood converges to its asymptotic mean. □

The next lemma finalizes the details of the proof of Theorem (4), as it provides the dominating bound $b(\mathbf{Y})$ for $|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})|$.

Lemma 1 *(Boundedness of the pseudo-likelihood)*
Let $\mathbf{Y}_1, ..., \mathbf{Y}_n$ be n independent random draws from a well-defined p-variate exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta}\, T(\mathbf{Y}) + h(\mathbf{Y})]$. Then, the pseudo-likelihood $\mathcal{L}_{PL}(\boldsymbol{\Theta}, \mathbf{Y})$ of $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is dominated as:

$$|\mathcal{L}_{PL}(\boldsymbol{\Theta}, \mathbf{Y})| \leq c_1 + c_2 \sum_{j \in \mathcal{V}} |Y_j|^\beta, \qquad \textit{for all } \boldsymbol{\Theta}, \tag{5}$$

if i) the parameter space of $\boldsymbol{\Theta}$ is compact, and ii) $\boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})$ is dominated by a polynomial: $|\boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})| \leq c_1' + c_2' \sum_{j \in \mathcal{V}} |Y_j|^\beta$, for $c_1', c_2' < \infty$ and some $\beta \in \mathbb{N}$.

Proof The proof decomposes the pseudo-likelihood in terms of the node-wise unnormalized distributions and the log-normalizing constants. Subsequently, bounds are derived for the terms in the decomposition. A first bound on the pseudo-likelihood is found by iterative application of the triangle inequality:

$$\begin{aligned} |\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| &\leq \sum_{j \in \mathcal{V}} \big| \log[P_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})] \big| \;=\; \sum_{j \in \mathcal{V}} \big| \log[p_j(Y_j \,|\, \mathbf{Y}_j)] - D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j}) \big| \\ &\leq \sum_{j \in \mathcal{V}} \big| \log[p_j(Y_j \,|\, \mathbf{Y}_{\setminus j})] \big| + \big| D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j}) \big|, \end{aligned} \tag{6}$$

where $p_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})$ is the unnormalized j-th conditional distribution and $D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j})$ its conditional log-normalizing constant. Next, we find a bound for the last term in equality (6).

Notice that the log-normalizing constant $D(\boldsymbol{\Theta})$ of the pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is related to the conditional log-normalizing constant $D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j})$ via,

$$\begin{aligned} \exp[D(\boldsymbol{\Theta})] = \sum_{\mathbf{Y}} p_{\boldsymbol{\Theta}}(\mathbf{Y}) \;&=\; \sum_{\mathbf{Y}} p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j})\; p_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j}) \\ &= \sum_{\mathbf{Y}_{\setminus j}} p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j}) \sum_{Y_j} p_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j}) \\ &= \sum_{\mathbf{Y}_{\setminus j}} p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j}) \exp[D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j})] \end{aligned}$$

With each summand on the right-hand side of the preceeding display being positive, it may be bounded further from below by any of the individual summands. Take the logarithm of the resulting inequality and obtain:

$$D_j(\boldsymbol{\Theta}, \mathbf{Y}_{\setminus j}) \leq D(\boldsymbol{\Theta}) - \log[p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j})] \;\leq\; D(\boldsymbol{\Theta}) + \big| \log[p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j})] \big| \qquad \text{for all } j \in \mathcal{V}.$$

Substitution of this into inequality (6) gives:

$$|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| \leq p|D(\boldsymbol{\Theta})| + \sum_{j \in \mathcal{V}} \big| \log[p_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})] \big| + \big| \log[p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j})] \big|. \tag{7}$$

Now use the factorization of the unnormalized distribution, $p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\mathcal{V}}) = p_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j}) \cdot p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j}), \quad j \in \mathcal{V}$, to bound the unnormalized probabilities of inequality (7):

$$\begin{aligned} \log[p_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})] + \log[p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\setminus j})] &= \log[p_{\boldsymbol{\Theta}}(\mathbf{Y}_{\mathcal{V}})] \\ &= \boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y}) \;=\; \sum_{\mathcal{U} \subseteq \mathcal{V}} f_{\mathcal{U}}(\mathbf{Y}_{\mathcal{U}}, \boldsymbol{\Theta}_{\mathcal{U}}), \end{aligned} \tag{8}$$

for a collection of functions $\{f_{\mathcal{U}}\}_{\mathcal{U}\subseteq\mathcal{V}}$ linear in $\boldsymbol{\Theta}$. As each term on the right-hand side of (8) is contained in either $\log[P(Y_j \,|\, \mathbf{Y}_{\setminus j})]$ or $\log[P(\mathbf{Y}_{\setminus j})]$, we conclude by the triangle inequality that:

$$|\log[P(Y_j \,|\, \boldsymbol{\Theta}, \mathbf{Y}_{\setminus j})]| + |\log[P(\mathbf{Y}_{\setminus j})]| \leq \sum_{\mathcal{U}\subseteq\mathcal{V}} |f_{\mathcal{U}}(\mathbf{Y}_{\mathcal{U}}, \boldsymbol{\Theta}_{\mathcal{U}})| \qquad \text{for all } j \in \mathcal{V}.$$

Substitute this into inequality (7) and find:

$$\frac{1}{p}|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| \leq |D(\boldsymbol{\Theta})| + \sum_{\mathcal{U}\subseteq\mathcal{V}} |f_{\mathcal{U}}(\mathbf{Y}_{\mathcal{U}}, \boldsymbol{\Theta}_{\mathcal{U}})|. \tag{9}$$

Assumption *ii)* of the lemma implies that $|f_U(\mathbf{Y}_{\mathcal{U}}, \boldsymbol{\Theta}_{\mathcal{U}})| \leq c_1 + c_2 \sum_{j\in\mathcal{V}} |Y_j|^\beta$, which, when used in inequality (9), yields:

$$\frac{1}{p}|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| \leq |D(\boldsymbol{\Theta})| + c_1' + c_2' \sum_{j\in\mathcal{V}} |Y_j|^\beta, \tag{10}$$

for some $\beta \in \mathbb{N}$. Finally, the exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is assumed to be well-defined. Hence, by the continuity and the compactness of its support, there exists a constant $M < \infty$ such that $|D(\boldsymbol{\Theta})| < M$ for all $\boldsymbol{\Theta}$. Hence,

$$|\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})| \leq c_1'' + c_2'' \sum_{j\in\mathcal{V}} |Y_j|^\beta, \tag{11}$$

for suitably chosen constants $c_1'', c_2'' < \infty$. □

Proof Consistency: Unique Maximum of the Pseudo-likelihood

For the proof of Theorem (1) we are left to show that $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$ is strict concave with the true parameter $\boldsymbol{\Theta}$ of $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ as unique global maximum. To this end, we first show that the exponent of any exponential family distribution is strictly concave.

Theorem 5 *(The exponential family distribution has a strict concave log-likelihood)*
Let $\mathbf{Y}$ *be a p-variate random variable following a well-defined exponential family distribution* $P_{\boldsymbol{\Theta}}(\mathbf{Y}) \propto \exp[\boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})]$. *Then the log-likelihood* $\log[P_{\boldsymbol{\Theta}}(\mathbf{Y})]$ *is strictly concave.*

Proof The proof first shows strict convexity of the log-normalizing constant of the exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$, followed by strict concavity of the log-likelihood $\log[P_{\boldsymbol{\Theta}}(\mathbf{Y})]$

The convexity of the log-normalizing constant of the exponential family distribution is shown using Hölder's Inequality (Holder, 1889). Define $\boldsymbol{\Theta}_0 = (1-\nu)\boldsymbol{\Theta}_1 + \nu\boldsymbol{\Theta}_2$ with $\nu \in (0,1)$ and $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$ in the parameter space. Let $E(\boldsymbol{\Theta}, \mathbf{Y}) = \boldsymbol{\Theta} \cdot T(\mathbf{Y}) + h(\mathbf{Y})$. Then the unnormalized exponential family distribution factorizes in accordance with:

$$\begin{aligned} \exp[-E(\boldsymbol{\Theta}_0, \mathbf{Y})] &= \exp\Big[-\nu E(\boldsymbol{\Theta}_1, \mathbf{Y}) - (1-\nu)E(\boldsymbol{\Theta}_2, \mathbf{Y})\Big] \\ &= \Big\{ \exp[-E(\boldsymbol{\Theta}_1, \mathbf{Y})]\Big\}^{\nu} \Big\{ \exp[-E(\boldsymbol{\Theta}_2, \mathbf{Y})]\Big\}^{1-\nu}. \end{aligned} \tag{12}$$

Summation over all outcomes and application of Hölder's inequality yields,

$$\sum_{\mathbf{Y}} \exp[-E(\boldsymbol{\Theta}_0, \mathbf{Y})] \leq \Big\{ \sum_{\mathbf{Y}} \exp[-E(\boldsymbol{\Theta}_1, \mathbf{Y})]\Big\}^{\nu} \Big\{ \sum_{\mathbf{Y}} \exp[-E(\boldsymbol{\Theta}_2, \mathbf{Y})]\Big\}^{1-\nu}. \tag{13}$$

Take the logarithm of the preceeding inequality to obtain one for the the log-normalizing constant of the exponential family distribution:

$$D(\boldsymbol{\Theta}_0) \leq \nu D(\boldsymbol{\Theta}_1) + (1-\nu)D(\boldsymbol{\Theta}_2). \tag{14}$$

from which it follows that $D(\boldsymbol{\Theta})$ is convex. For $\nu \in (0,1)$, Hölder's inequality specifies that equality in (14) holds if and only if $E(\boldsymbol{\Theta}_2, \mathbf{Y}) = E(\boldsymbol{\Theta}_1, \mathbf{Y})$ for all $\mathbf{Y}$. The linearity of E in $\boldsymbol{\Theta}$ implies that this holds if and only if $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$. The log-normalizing constant is thus strict convex in $\boldsymbol{\Theta}$.

Strict concavity of the log-likelihood now follows from the linearity of the energy function and strict convexity of the log-normalizing constant of the exponential family distribution,

$$\begin{aligned}\log[P_{\boldsymbol{\Theta}_0}(\mathbf{Y})] &= \log\Big\{\exp[-E(\boldsymbol{\Theta}_0,\mathbf{Y})]\Big\} - D(\boldsymbol{\Theta}_0)\\ &> \nu\log\Big\{\exp[-E(\boldsymbol{\Theta}_1,\mathbf{Y})]\Big\} + (1-\nu)\log\Big\{\exp[-E(\boldsymbol{\Theta}_2,\mathbf{Y})]\Big\}\\ &\quad -\nu D(\boldsymbol{\Theta}_1) - (1-\nu)D(\boldsymbol{\Theta}_2)\\ &= \nu\log[P_{\boldsymbol{\Theta}_1}(\mathbf{Y})] + (1-\nu)\log P_{\boldsymbol{\Theta}_2}(\mathbf{Y}),\end{aligned}$$

for all $\mathbf{Y}$ and parameters $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$. □

The following corollary shows that the exponential family distribution is well-defined on a convex parameter space, which will be exploited in the construction of a parallel Newton-Raphson algorithm to find the maximum of a concave function on a convex set.

Corollary 1 *(The exponential family distribution is well-defined on a convex parameter space)*
Let $\mathbf{Y}$ be a p-variate random variable following an exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$. Then $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined on a convex parameter space.

Proof The proof is a direct corollary of Theorem (5), and checks the definition of convexity for the parameter space.

Choose $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$ in the parameter space with $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$ and consider $\boldsymbol{\Theta}_0 = \nu\boldsymbol{\Theta}_1 + (1-\nu)\boldsymbol{\Theta}_2$ for any $\nu \in (0,1)$. Assume that $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined for $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, such that the log-normalizing constant of $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ satisfies $D(\boldsymbol{\Theta}_1) < \infty$ and $D(\boldsymbol{\Theta}_2) < \infty$. The distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ has an energy function that is linear in $\boldsymbol{\Theta}$, such that $D(\boldsymbol{\Theta})$ is convex (proof Theorem 5). Then,

$$D(\boldsymbol{\Theta}_0) \leq \nu D(\boldsymbol{\Theta}_1) + (1-\nu)D(\boldsymbol{\Theta}_0) < \infty.$$

It follows that $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined for $\boldsymbol{\Theta}_0$ for any $\nu \in (0,1)$, and thus well-defined on a convex parameter space. □

Recall that the proof of main Theorem (1) leaves to show that $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta},\mathbf{Y})]$ is strictly concave and uniquely maximized by the true parameter $\boldsymbol{\Theta}$ of the distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$. The following corollary addresses the first part of the remaining proof and shows that both the pseudo-likelihood $\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta},\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$ and its (asymptotic) mean $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta},\mathbf{Y})]$ are concave and therefore have a global maximum.

Corollary 2 *(Strict concavity of the pseudo-likelihood)*
Let $\mathbf{Y}_1,...,\mathbf{Y}_n$ be p-variate random draws from a well-defined exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$. Then, the pseudo-likelihood $\mathcal{L}_{PL}(\boldsymbol{\Theta},\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$ and its asymptotic mean $\mathbb{E}[\mathcal{L}_{PL}(\boldsymbol{\Theta},\mathbf{Y})]$ are concave and have a global maximum.

Proof Recall that the pseudo-likelihood is the average over the observations of the sum of all conditional log-likelihoods,

$$\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta},\mathbf{Y}_1,...,\mathbf{Y}_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{V}}\log[P_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})]. \tag{15}$$

Moreover, the conditional log-likelihood $\log[P_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})]$ is itself the log-likelihood of an exponential family distribution with energy function linear in $\boldsymbol{\Theta}$, hence strictly concave in the parameter (Theorem 5). Then the pseudo-likelihood, being an average of the sum of strictly concave functions, is strict concave itself.

Finally, the asymptotic mean,

$$\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta},\mathbf{Y})] = \lim_{n\to\infty}\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta},\mathbf{Y}_1,\ldots,\mathbf{Y}_n),$$

is the limit of an uniformly convergent sequence of strictly concave functions (Theorem 4). Its limit must therefore also be strictly concave. □

To complete the proof of Theorem (1), the following lemma shows that the true parameter $\boldsymbol{\Theta}$ of the distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ maximizes $\mathbb{E}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$.

Lemma 2 *(Unique, global maximizer of the asymptotic mean of the pseudo-likelihood)*
Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be p-variate draws from an exponential family distribution $P_{\boldsymbol{\Theta}^*}(\mathbf{Y})$ with parameter $\boldsymbol{\Theta}^*$. Then $\boldsymbol{\Theta}^*$ is a unique, global maximizer of the asymptotic mean of the pseudo-likelihood, i.e. $\mathbb{E}[\mathcal{L}_{PL}(\boldsymbol{\Theta}^*, \mathbf{Y})] > \mathbb{E}[\mathcal{L}_{PL}(\boldsymbol{\Theta}, \mathbf{Y})]$ for all $\boldsymbol{\Theta}$ with $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}^*$.

Proof Rewrite the asymptotic mean of the pseudo-likelihood by means of Bayes' rule,

$$\begin{aligned}
\mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y})}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})] &= \sum_{j \in \mathcal{V}} \mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y})} \Big\{ \log[P_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})] \Big\}. \qquad (16) \\
&= \sum_{j \in \mathcal{V}} \sum_{\mathbf{y}_{\setminus j}} P_{\boldsymbol{\Theta}^*}(\mathbf{y}_{\setminus j}) \sum_{y_j} P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{y}_{\setminus j}) \log[P_{\boldsymbol{\Theta}}(y_j \,|\, \mathbf{y}_{\setminus j})] \\
&= \sum_{j \in \mathcal{V}} \mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y}_{\setminus j})} \left\{ \sum_{y_j} P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{Y}_{\setminus j}) \log[P_{\boldsymbol{\Theta}}(y_j \,|\, \mathbf{Y}_{\setminus j})] \right\}.
\end{aligned}$$

Then, from the nonnegativity of the Kullback-Leibler divergence, we obtain:

$$\sum_{y_j} P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{Y}_{\setminus j}) \log[P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{Y}_{\setminus j})] \geq \sum_{y_j} P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{Y}_{\setminus j}) \log[P_{\boldsymbol{\Theta}}(y_j \,|\, \mathbf{Y}_{\setminus j})],$$

for any model $P_{\boldsymbol{\Theta}}(y_j \,|\, \mathbf{Y}_{\setminus j})$ and all $\mathbf{Y}_{\setminus j}$. Moreover, equality holds if and only if $P_{\boldsymbol{\Theta}^*}(Y_j \,|\, \mathbf{Y}_{\setminus j}) = P_{\boldsymbol{\Theta}}(Y_j \,|\, \mathbf{Y}_{\setminus j})$ for all Y_j. This only happens when $\boldsymbol{\Theta} = \boldsymbol{\Theta}^*$. As $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}^*$ by assumption, substitution of the preceeding display into (16) yields,

$$\begin{aligned}
\mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y})}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})] &< \sum_{j \in \mathcal{V}} \mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y}_{\setminus j})} \left\{ \sum_{y_j} P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{Y}_{\setminus j}) \log[P_{\boldsymbol{\Theta}^*}(y_j \,|\, \mathbf{Y}_{\setminus j})] \right\} \\
&= \sum_{j \in \mathcal{V}} \mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y})} \Big\{ \log[P_{\boldsymbol{\Theta}^*}(Y_j \,|\, \mathbf{Y}_{\setminus j})] \Big\} \\
&= \mathbb{E}_{P_{\boldsymbol{\Theta}^*}(\mathbf{Y})}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}^*, \mathbf{Y})],
\end{aligned}$$

and therefore the true parameter $\boldsymbol{\Theta}^*$ of $P_{\boldsymbol{\Theta}^*}(\mathbf{Y})$ uniquely maximizes the asymptotic mean of the pseudo-likelihood. □

Consistency of the Penalized Pseudo-Likelihood Estimator

The following corollary shows that the more general penalized pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_n^{\text{pen}}$ of $\boldsymbol{\Theta}$ is consistent for a strictly convex and continuous penalty function.

Corollary 3 *(Consistency of the penalized pseudo-likelihood)*
Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be p-variate draws from an exponential family distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$. Then the penalized pseudo-likelihood estimator, $\widehat{\boldsymbol{\Theta}}_n^{pen} := \arg\max_{\boldsymbol{\Theta}} \mathcal{L}_{penPL}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, that maximizes the penalized pseudo-likelihood is consistent, i.e., $\widehat{\boldsymbol{\Theta}}_n^{pen} \xrightarrow{p} \boldsymbol{\Theta}$ as $n \to \infty$ if,
i) The conditions of Theorem (1) are satisfied,
ii) The penalty function $f_{pen}(\boldsymbol{\Theta})$ is strict convex and continuous,
iii) The penalty parameter λ_n converges in probability to zero: $\lambda_n \xrightarrow{p} 0$ as $n \to \infty$.

Proof The proof verifies the conditions of Theorem (2), for which it uses the consistency of the pseudo-likelihood estimator. For the uniform convergence of the penalized likelihood (the 'stochastic' condition (1) of Theorem 2) use:

$$\begin{aligned}
&\sup_{\boldsymbol{\Theta}} \Big| \mathcal{L}_{\text{penPL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) - \mathbb{E}_{P_{\boldsymbol{\Theta}}}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})] \Big| \qquad (17) \\
&\quad = \sup_{\boldsymbol{\Theta}} \Big| \mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) - \mathbb{E}_{P_{\boldsymbol{\Theta}}}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})] - \lambda_n f_{\text{pen}}(\boldsymbol{\Theta}) \Big| \\
&\quad \leq \sup_{\boldsymbol{\Theta}} \Big| \mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) - \mathbb{E}_{P_{\boldsymbol{\Theta}}}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})] \Big| + \sup_{\boldsymbol{\Theta}} \lambda_n |f_{\text{pen}}(\boldsymbol{\Theta})|.
\end{aligned}$$

The first supremum on the right-hand side of the preceeding display converges to zero in probability (cf. the proof of Theorem 1). Consider the second supremum in this expression. The penalty function $f_{pen}(\boldsymbol{\Theta})$ is continuous with a compact domain. Its image is therefore bounded. Hence, there exists a finite M such that $|f_{\text{pen}}(\boldsymbol{\Theta})| < M$ for all $\boldsymbol{\Theta}$ in the parameter space. Together with the convergence (in probability) of λ_n to zero, it follows that $\sup_{\boldsymbol{\Theta}} \lambda_n |f_{\text{pen}}(\boldsymbol{\Theta})| \xrightarrow{p} 0$. The penalized pseudo-likelihood $\mathcal{L}_{\text{penPL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ thus converges uniformly to the asymptotic mean of the (unpenalized) pseudo-likelihood.

The 'deterministic' condition (2) of Theorem (2) is satisfied for the asymptotic mean $\mathbb{E}_{P_{\boldsymbol{\Theta}}}[\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y})]$ of the unpenalized pseudo-likelihood as shown in the proof of Theorem (1). Finally, $\mathcal{L}_{\text{penPL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) - \lambda_n f_{\text{pen}}(\boldsymbol{\Theta})$ is strictly concave since the penalty function $f_{\text{pen}}(\boldsymbol{\Theta})$ is strictly convex and $\mathcal{L}_{\text{PL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ is (strictly) concave (cf. Corollary 2). Consequently, $\mathcal{L}_{\text{penPL}}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ has an unique global maximum. The corresponding maximizer $\widehat{\boldsymbol{\Theta}}_n^{\text{pen}}$ converges to the true parameter $\boldsymbol{\Theta}$ in probability (by Theorem 2). □

Finally, the following corollary specifies the consistency result for the setting considered in the main text: variates from the GLM family and the ridge penalty function.

Corollary 4 *(The ridge pseudo-likelihood estimator of the pairwise MRF distribution is consistent)*
Let the p-variate random variables $\mathbf{Y}_1, ..., \mathbf{Y}_n$ *be drawn from a well-defined pairwise MRF distribution* $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ *with parameter* $\boldsymbol{\Theta}$*. The ridge pseudo-likelihood estimator, defined as* $\widehat{\boldsymbol{\Theta}}_n^{ridge} := \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{PL}^{ridge}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$*, that maximizes the ridge-penalized pseudo-likelihood is consistent, i.e.,*

$$\widehat{\boldsymbol{\Theta}}_n^{ridge} \xrightarrow{p} \boldsymbol{\Theta}, \qquad \text{as } n \to \infty,$$

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 Corollary (3) is applied. Notice that $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is by construction of the pairwise MRF distribution an exponential family distribution. Moreover, the parameter space is assumed to be compact and such that $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ is well-defined. Then the conditions of Theorem (1) are satisfied as the exponent of the pairwise MRF distribution is at most quadratic in its variables $\mathbf{Y}$. Finally, the ridge penalty $\|\cdot\|_2^2$ is strictly convex and continuous, such that all conditions of Corollary (3) hold. □

Supplementary Material F: Convergence of the parallel block-wise Newton-Raphson algorithm (proof Theorem 3 of the main text)

The penalized pseudo-likelihood is maximized by means of a parallel implementation of the Newton-Raphson algorithm. At each step of the algorithm the parallelization yields multiple updates for the same parameter. With Theorem (6) we study how these updates are to be combined in order to ensure convergence of the parallel algorithm. The proof hinges upon the fact that the maximization of the penalized pseudo-likelihood is a strict concave optimization problem over a convex set. Moreover, it exploits the fact that the maximizer is found iteratively by means of the Newton-Raphson algorithm, which at each step improves upon the current approximation of the maximizer. Given the current approximation and a direction in which the penalized pseudo-likelihood increases, it is left to decide upon the step size to proceed in this direction. The presented parallel implementation of the Newton-Raphson algorithm effectively proposes several directions, each updating partially overlapping subsets of parameters, in which to proceed. The proof shows how these directions may be combined. A simple average of these directions can, by the concavity of the penalized pseudo-likelihood, be shown (cf. Lemma 5) to yield a novel direction in which all parameters are updated and that yields a better approximation. Theorem (6) then tells us that iterative updating in this 'average direction' produces a sequence of updates that converges to the maximum of the penalized pseudo-likelihood. Finally, the convergence in the 'average direction' may be slow and can improved upon by taking larger (than average) step sizes, which is stated in Lemma (7).

The proof of Theorem (6) comprises the iterative application of Lemma (5), which rests on Lemma's (3) and (4) that are presented and proven prior to Lemma (5). In a schematic description, Lemma (3) shows that one step of Algorithm (1) with $\boldsymbol{\Theta}^{(k)}$ strictly improves the second-order Taylor approximation of the pseudo-likelihood with some $\epsilon > 0$. Next, lemma (4) shows that the same improvement $\epsilon > 0$ holds for any $\boldsymbol{\psi}^{(k)}$ (in some sense) close to $\boldsymbol{\Theta}^{(k)}$. Lemma (5) subsequently shows that this result generalizes to the penalized pseudo-likelihood $\mathcal{L}_{\text{penPL}}$. Finally, Theorem (6) shows that iterative application of Lemma (5) forces convergence of the sequence $\{\boldsymbol{\Theta}^{(k)}\}_{k\geq 0}$ to $\widehat{\boldsymbol{\Theta}}(\lambda)$.

Theorem 6 (Algorithm (1) converges to $\hat{\Theta}_n^{\text{pen}}$) *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ 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 $\mathcal{L}_{penPL}(\boldsymbol{\Theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. 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 The proof follows a common approach used to show convergence of algorithms, but is modified for the pseudo-likelihood and parallel approach. The proof crucially relies on Lemma (5) and the remainder of the proof modifies e.g. Theorem (1) from Lee et al. [2006] or Theorem (1) by Höfling and Tibshirani [2009]. As in Höfling and Tibshirani [2009], consider the compact set,

$$P_\delta = \{\boldsymbol{\theta} \,:\, ||\boldsymbol{\theta} - \widehat{\boldsymbol{\Theta}}(\lambda)||_2 \geq \delta\} \cap \boldsymbol{\Omega}. \tag{18}$$

with parameter space $\boldsymbol{\Omega}$. Then the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ converges to $\widehat{\boldsymbol{\Theta}}(\lambda)$ if, for all $\delta > 0$, there exists a constant $K_\delta \in \mathbb{N}_0$ such that $\boldsymbol{\theta}^{(k)} \notin P_\delta$ for $k \geq K_\delta$. Moreover, it follows that Algorithm (1) terminates in a finite number of steps for any threshold $\tau > 0$ since $||\partial \mathcal{L}_{\text{penPL}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}||_2 \to 0$ as $\boldsymbol{\theta}^{(k)} \to \widehat{\boldsymbol{\Theta}}(\lambda)$. The remainder of the proof shows that the constant K_δ exists.

To this end, let $\delta > 0$. For any $\boldsymbol{\theta} \in P_\delta$, Lemma (5) provides a $\delta_{\boldsymbol{\theta}} > 0$ and $\alpha_{\boldsymbol{\theta}} \geq p$ such that for all,

$$\boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}} := \{\tilde{\boldsymbol{\theta}} : \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2 < \delta_{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^q\} \quad \text{and} \quad \alpha \geq \alpha_{\boldsymbol{\theta}},$$

a single step of Algorithm (1) initiated by $\boldsymbol{\psi}^{(0)}$ yields an updated parameter estimate $\boldsymbol{\psi}^{(1)}$ that satisfies, for some $\varepsilon_{\alpha,\boldsymbol{\theta}}$,

$$\mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^{(1)}) \geq \mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^{(0)}) + \varepsilon_{\alpha,\boldsymbol{\theta}}. \tag{19}$$

Notice that $P_\delta \subseteq \cup_{\boldsymbol{\theta}\in P_\delta} N_{\boldsymbol{\theta}}$ by definition of the neighborhoods $\{N_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in P_\delta}$. Moreover, P_δ is compact such that the Heine-Borel theorem yields a finite subset $Q_\delta \subset P_\delta$ for which $\{N_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in Q_\delta}$ is a finite open cover

of P_δ, i.e. $P_\delta \subseteq_{\boldsymbol{\theta} \in Q_\delta} N_{\boldsymbol{\theta}}$. This construction of the finite set Q_δ is the key idea of the theory in Lee et al. [2006] and Höfling and Tibshirani [2009]. As Q_δ is finite, the following maximum exists,

$$\alpha_{\max} := \max_{\boldsymbol{\theta} \in Q_\delta} \alpha_{\boldsymbol{\theta}}.$$

Now choose $\alpha \geq \alpha_{\max}$. For all $\boldsymbol{\theta} \in Q_\delta$ there exist some $\epsilon_{\alpha,\boldsymbol{\theta}} > 0$ such that (19) holds since $\alpha \geq \alpha_{\max} \geq \alpha_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in Q_\delta$. As Q_δ is a finite set, the following minimum exists,

$$\epsilon_{min} := \min_{\boldsymbol{\theta} \in Q_\delta} \epsilon_{\alpha,\boldsymbol{\theta}} > 0.$$

Now let $\boldsymbol{\theta}^{(k)} \in P_\delta$. Then any step $k \geq 0$ of Algorithm (1) with multiplier $\alpha_{\max}$ and $\boldsymbol{\theta}^{(k)} \in P_\delta$, generates a parameter $\boldsymbol{\theta}^{(k+1)}$ that satisfies (19),

$$\mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}^{(k+1)}) \geq \mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}^{(k)}) + \epsilon_{min}. \tag{20}$$

It follows from (20) that every step of Algorithm (1) inside P_δ increases $\mathcal{L}_{\text{penPL}}$ with some constant $\epsilon_{min} > 0$. Moreover, $\mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ is strictly concave (Corollary 2) with unique global maximum $\widehat{\boldsymbol{\Theta}}(\lambda)$. Therefore there exists a $K_\delta \in \mathbb{N}_0$ such that $||\boldsymbol{\theta}^{(k)} - \widehat{\boldsymbol{\Theta}}(\lambda)||_2 < \delta$ for all $k \geq K_\delta$, hence $\boldsymbol{\theta}^{(k)} \notin P_\delta$ for $k \geq K_\delta$, which concludes the proof. □

The remainder of the proof of Theorem (4) of the main text comprises the iterative application of Lemma (5), which rests on Lemma's (3) and (4) that are presented and proven prior to Lemma (5). We show that the Taylor expansion of the pseudo-loglikelihood is maximized 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. Thus, we can maximize by means of a block-wise Newton-Raphson approach, starting from an initial guess $\hat{\boldsymbol{\theta}}^{(0)}$ and updating the current parameter value of block $\hat{\boldsymbol{\theta}}_j^{(k)}$ with $\hat{\boldsymbol{\theta}}_j^{(k+1)}$ through,

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

Then, the following Lemma shows that any convex combination of the block-wise solutions $\hat{\boldsymbol{\theta}}_j^{(k+1)}$ also increases the pseudo-loglikelihood.

Lemma 3
Let $\mathbf{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$ be a p-variate 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 penalized pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}(\lambda)$. A single step of Algorithm (1) initiated with $\boldsymbol{\theta}^{(0)}$ yields 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). Define $\boldsymbol{\theta}^{(1)}$ as:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \sum_{j \in \mathcal{V}} \omega_j \tilde{\boldsymbol{\theta}}_j, \tag{21}$$

where $\{\omega_j\}_{j \in \mathcal{V}} \in [0,1]^p$ with $\sum_{j \in \mathcal{V}} \omega_j = 1$. Let $L(\cdot; \boldsymbol{\theta}^{(0)})$ be the second-order Taylor approximation of the penalized pseudo-likelihood at $\boldsymbol{\theta}^{(0)}$. Then, $L(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) > L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)})$.

Proof The proof first shows that each block-wise update yields an increase of the second-order Taylor approximation, and, subsequently, that any convex combination of these block-wise updates does the same. Define the second-order Taylor approximation of the penalized pseudo-likelihood as:

$$L(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}) = \mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}^{(0)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^\top (\partial \mathcal{L}_{\text{penPL}}/\partial \boldsymbol{\theta})\big|_{\boldsymbol{\theta}^{(0)}} + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^\top \mathbf{H}\big|_{\boldsymbol{\theta}^{(0)}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}), \tag{22}$$

where $\mathbf{H}$ is the Hessian of the penalized pseudo-likelihood. To show the majorization of a block-wise update of the current parameter value, note that by construction of $\tilde{\boldsymbol{\theta}}_j$ (which extends the block-wise update $\tilde{\boldsymbol{\theta}}_j$ with zeros):

$$\tilde{\boldsymbol{\theta}}_j^\top \frac{d\mathcal{L}_{\text{penPL}}}{d\boldsymbol{\theta}} = \boldsymbol{\theta}_j^\top \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial \boldsymbol{\theta}_j} \qquad \text{and} \qquad \tilde{\boldsymbol{\theta}}_j^\top \mathbf{H} \tilde{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j. \tag{23}$$

Substitute this into the 2^{nd} order Taylor approximation:

$$L(\boldsymbol{\theta}^{(0)}+\tilde{\boldsymbol{\theta}}_j;\boldsymbol{\theta}^{(0)}) = L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)}) + \tilde{\boldsymbol{\theta}}_j^\top \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial \boldsymbol{\theta}^{(0)}} + \frac{1}{2}\tilde{\boldsymbol{\theta}}_j^\top \mathbf{H}\tilde{\boldsymbol{\theta}}_j \;\; = \;\; L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)}) - \tfrac{1}{2}\tilde{\boldsymbol{\theta}}_j^\top \mathbf{H}\tilde{\boldsymbol{\theta}}_j,$$

in which we have used that the block-wise update – being the solution of a Newton-Raphson update – satisfies $\mathbf{H}_j\boldsymbol{\theta}_j = -\partial\mathcal{L}_{\text{penPL}}/\partial\boldsymbol{\theta}_j$. The majorization now follows as $\frac{1}{2}\tilde{\boldsymbol{\theta}}_j^\top \mathbf{H}\tilde{\boldsymbol{\theta}}_j < 0$ due the fact that Hessian of a strictly concave function is negative definite (see Corollary 2).

The majorization extends to any convex combination of block-wise updates as defined by (21). Hereto apply Jensen's inequality (Jensen, 1906) to the strictly concave 2^{nd} order Taylor approximation,

$$L(\boldsymbol{\theta}^{(1)};\boldsymbol{\theta}^{(0)}) = L\Big[\sum_{j\in\mathcal{V}} \omega_j\big(\boldsymbol{\theta}^{(0)}+\tilde{\boldsymbol{\theta}}_j\big);\boldsymbol{\theta}^{(0)}\Big] \;\; \geq \;\; \sum_{j\in\mathcal{V}} \omega_j L(\boldsymbol{\theta}^{(0)}+\tilde{\boldsymbol{\theta}}_j;\boldsymbol{\theta}^{(0)}),$$

where the weights ω_j sum to one: $\sum_{j\in\mathcal{V}}\omega_j = 1$. The block-wise majorization then gives:

$$L(\boldsymbol{\theta}^{1};\boldsymbol{\theta}^{(0)}) \geq L(\boldsymbol{\theta}^{0};\boldsymbol{\theta}^{(0)}) - \sum_{j\in\mathcal{V}} \frac{1}{2}\omega_j\tilde{\boldsymbol{\theta}}_j^\top \mathbf{H}\tilde{\boldsymbol{\theta}}_j \;\; > \;\; L(\boldsymbol{\theta}^{0};\boldsymbol{\theta}^{(0)}), \tag{24}$$

which warrants the majorization by any convex combination of the block-wise updates. □

Thus, the procedure of block-wise updating the parameter $\boldsymbol{\theta}$ (either sequentially, or by any convex combination of block-wise updates (parallel)) maximizes the pseudo-loglikelihood. The following lemma extends the preceeding one to hold for parameter values chosen from a neighborhood of $\boldsymbol{\theta}^{(0)}$.

Lemma 4
Let $\mathbf{Y} = \{\mathbf{Y}_1,\ldots,\mathbf{Y}_n\}$ be p-variate 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 penalized pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}(\lambda)$. Let $L(\cdot;\boldsymbol{\theta})$ be the second-order Taylor approximation of $\mathcal{L}_{penPL}$ at $\boldsymbol{\theta}$. Then, there exist constants $\delta_{\boldsymbol{\theta}^{(0)}}, \varepsilon_{\boldsymbol{\theta}^{(0)}} > 0$ such that for all,

$$\boldsymbol{\psi}^{(0)} \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}\|_2 < \delta_{\boldsymbol{\theta}^{(0)}}, \boldsymbol{\theta}\in\mathbb{R}^q\}, \tag{25}$$

one step of Algorithm (1) with $\boldsymbol{\psi}^{(0)}$ generates a collection of block-wise updates $\{\tilde{\boldsymbol{\psi}}_j\}_{j\in\mathcal{V}}$ (line 6 of Algorithm 1) that satisfies,

$$L\Big(\boldsymbol{\psi}^{(0)} + \tfrac{1}{p}\sum_{j\in\mathcal{V}}\tilde{\boldsymbol{\psi}}_j;\boldsymbol{\psi}^{(0)}\Big) \geq L(\boldsymbol{\psi}^{(0)};\boldsymbol{\psi}^{(0)}) + \varepsilon_{\boldsymbol{\theta}^{(0)}}. \tag{26}$$

Proof The proof proceeds by showing that, for $\boldsymbol{\psi}^{(0)}$ sufficiently close to $\boldsymbol{\theta}^{(0)}$, Algorithm (1) yields block-wise updates $\{\tilde{\boldsymbol{\psi}}_j\}_{j\in\mathcal{V}}$ and $\{\tilde{\boldsymbol{\theta}}_j\}_{j\in\mathcal{V}}$ for $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ (respectively) for which constants $\varepsilon_0,\varepsilon_1 > 0$ exist such that the following sequence of (in)equalities holds:

$$\begin{aligned} L(\boldsymbol{\psi}^{(0)}+\tilde{\boldsymbol{\psi}}_j;\boldsymbol{\psi}^{(0)}) \geq^{(i)} \;& L(\boldsymbol{\psi}^{(0)}+\tilde{\boldsymbol{\theta}}_j;\boldsymbol{\psi}^{(0)}) && \geq^{(ii)} \; L(\boldsymbol{\theta}^{(0)}+\tilde{\boldsymbol{\theta}}_j;\boldsymbol{\psi}^{(0)}) - \varepsilon_0 \\ \geq^{(iii)} \;& L(\boldsymbol{\theta}^{(0)}+\tilde{\boldsymbol{\theta}}_j;\boldsymbol{\theta}^{(0)}) - 2\varepsilon_0 && =^{(iv)} \; L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)}) - 2\varepsilon_0 + \varepsilon_1 \\ \geq^{(v)} \;& L(\boldsymbol{\psi}^{(0)};\boldsymbol{\psi}^{(0)}) - 3\varepsilon_0 + \varepsilon_1. \end{aligned} \tag{27}$$

This, together with the Jensen's inequality, implies that there is a $\varepsilon_{\boldsymbol{\theta}^{(0)}} > 0$ such that inequality (26) holds for all $\boldsymbol{\psi}^{(0)}$ close to $\boldsymbol{\theta}^{(0)}$. In the remainder the inequalities of display (27) are proven subsequentially.

To substantiate inequality *(i)* of display (27) the block-wise updates are constructed. To this end, define the neighborhood of $\boldsymbol{\theta}^{(0)}$ as,

$$N_{\boldsymbol{\theta}^{(0)}}(\delta) := \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}\|_2 < \delta, \boldsymbol{\theta}\in\mathbb{R}^q\}.$$

Let $\delta_0 > 0$ and suppose that $\boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_0)$. Perform one step of Algorithm (1) with both $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ and let $\{\tilde{\boldsymbol{\psi}}_j\}_{j\in\mathcal{V}}$ and $\{\tilde{\boldsymbol{\theta}}_j\}_{j\in\mathcal{V}}$ be the block-wise updates (line 6 of Algorithm 1) for $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ respectively. Define: $\boldsymbol{\psi}^{(1)} = \boldsymbol{\psi}^{(0)} + \frac{1}{p}\sum_{j\in\mathcal{V}}\tilde{\boldsymbol{\psi}}_j$. As in Lemma (3), application of Jensen's inequality to the strictly concave second-order Taylor approximation of $\mathcal{L}_{\text{penPL}}$ at $\boldsymbol{\psi}^{(0)}$ yields:

$$L(\boldsymbol{\psi}^{(1)};\boldsymbol{\psi}^{(0)}) = L\Big(\boldsymbol{\psi}^{(0)} + \tfrac{1}{p}\sum_{j\in\mathcal{V}}\tilde{\boldsymbol{\psi}}_j;\boldsymbol{\psi}^{(0)}\Big) \;\; \geq \;\; \tfrac{1}{p}\sum_{j\in\mathcal{V}} L(\boldsymbol{\psi}^{(0)}+\tilde{\boldsymbol{\psi}}_j;\boldsymbol{\psi}^{(0)}). \tag{28}$$

Furthermore, as $\tilde{\boldsymbol{\psi}}_j$ and $\tilde{\boldsymbol{\theta}}_j$ have a shared support (defined by those parameters corresponding to the j-th variate) and by the definition of $\tilde{\boldsymbol{\psi}}_j$ (maximizing $\mathcal{L}_{\text{penPL}}$ with respect to the parameters of the j-th variate while keeping all others constant at $\boldsymbol{\psi}^{(0)}$), its holds that:

$$L(\boldsymbol{\psi}^{(0)} + \tilde{\boldsymbol{\psi}}_j; \boldsymbol{\psi}^{(0)}) \geq L(\boldsymbol{\psi}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)}) \qquad \text{for all } j \in \mathcal{V}.$$

Substitution of the preceeding inequality into that of display (28) yields the following lower bound for $L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)})$:

$$L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)}) \geq \tfrac{1}{p} \sum_{j \in \mathcal{V}} L(\boldsymbol{\psi}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)}), \tag{29}$$

which is the first step *(i)* of display (27) towards proving inequality (26).

For inequalities *(ii)* and *(iii)* of display (27) note that the second-order Taylor approximation $L(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)})$ of $\mathcal{L}_{\text{penPL}}(\boldsymbol{\theta})$ is a bicontinuous function in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(0)}$, in the latter variable by virtue of the well-definedness of $P_{\boldsymbol{\Theta}}(\mathbf{Y})$. By the continuity in $\boldsymbol{\theta}$ for every ε_0 there exists δ_1 such that:

$$\|L(\boldsymbol{\psi}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)}) - L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)})\|_2 < \varepsilon_0 \qquad \text{for all } \boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_1) \text{ and } j \in \mathcal{V}. \tag{30}$$

The continuity of the second-order Taylor approximation in $\boldsymbol{\theta}^{(0)}$ implies, for all $\varepsilon_0 > 0$, the existence of $\delta_2 > 0$ such that:

$$\|L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)}) - L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\theta}^{(0)})\|_2 < \varepsilon_0 \qquad \text{for all } \boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_2) \text{ and } j \in \mathcal{V}. \tag{31}$$

Now set $\delta_0 = \min\{\delta_1, \delta_2\}$ and assume that $\boldsymbol{\psi}^0 \in N_{\boldsymbol{\theta}^{(0)}}(\delta_0)$. Then, (30) and (31) together ensure that:

$$L(\boldsymbol{\psi}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)}) \geq L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\psi}^{(0)}) - \varepsilon_0 \;\geq\; L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\theta}^{(0)}) - 2\varepsilon_0 \qquad \text{for all } j \in \mathcal{V}.$$

Substitution of the preceeding inequality into that of display (29) yields:

$$L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)}) \geq \tfrac{1}{p} \sum_{j \in \mathcal{V}} L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\theta}^{(0)}) - 2\varepsilon_0, \tag{32}$$

which is the next step *(iii)* of display (27) towards proving inequality (26).

Lemma (3) is now exploited to find a lower bound for the convex combination $\frac{1}{p}\sum_{j \in \mathcal{V}} L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\theta}^{(0)})$ and thereby proving (in)equality *(iv)* of display (27). To this end average inequality (24) obtained in the proof of Lemma (3) over $j \in \mathcal{V}$ to arrive at:

$$\tfrac{1}{p} \sum_{j \in \mathcal{V}} L(\boldsymbol{\theta}^{(0)} + \tilde{\boldsymbol{\theta}}_j; \boldsymbol{\theta}^{(0)}) = L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)}) - \tfrac{1}{2p} \sum_{j \in \mathcal{V}} \tilde{\boldsymbol{\theta}}_j^\top \mathbf{H} \tilde{\boldsymbol{\theta}}_j. \tag{33}$$

Set ε_{min} equal to the second summand of the preceeding display, which is positive by the negative definiteness of $\mathbf{H}$ and the fact that $\|\tilde{\boldsymbol{\theta}}_j\|_2 > 0$ for at least one $j \in \mathcal{V}$ as $\boldsymbol{\theta}^{(0)}$ is not the maximum of $\mathcal{L}_{\text{penPL}}$. Finally, substitution of (33) and ε_{min} into the bound (32) yields:

$$L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)}) \geq L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)}) - 2\varepsilon_0 + \varepsilon_{min} \qquad \text{for all } \boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_0), \tag{34}$$

which provides a lower bound for $L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)})$ and yields inequality *(iv)* of display (27).

Finally, using continuity of $\mathcal{L}_{\text{penPL}}(\boldsymbol{\theta})$ (proof Theorem 4), a lower bound for $L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)})$ in terms of $L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)})$ and ε_{min} is derived conform inequality *(v)* of display (27). As $\mathcal{L}_{\text{penPL}}$ is continuous in $\boldsymbol{\Theta}$, there exists $\delta_3 > 0$ for which:

$$\|\mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^{(0)})\|_2 < \varepsilon_0 \qquad \text{for all } \boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_3).$$

Set $\delta_0 = \min\{\delta_1, \delta_2, \delta_3\}$ and assume that $\boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_0)$. It follows from the preceeding display that $\mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}^{(0)}) \geq \mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^0) - \varepsilon_0$, which yields, by the definition of the Taylor approximation (22),

$$L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)}) = \mathcal{L}_{\text{penPL}}(\boldsymbol{\theta}^{(0)}) \;\geq\; \mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^{(0)}) - \varepsilon_0 \;=\; L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) - \varepsilon_0. \tag{35}$$

Substitution of (35) into (34) yields the following lower bound for $L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)})$,

$$L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)}) \geq L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) - 3\varepsilon_0 + \varepsilon_{min}.$$

When ε_0 is set equal to $\frac{1}{4}\varepsilon_{min}$, there thus exists a $\delta_{min} > 0$ such that:

$$L(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)}) \geq L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) + \frac{1}{4}\varepsilon_{min} \qquad \text{for all } \boldsymbol{\psi}^{(0)} \in N_{\boldsymbol{\theta}^{(0)}}(\delta_{min}),$$

which proves the statement (26). □

Where the preceeding two lemma's show that a single step of Algorithm (1) improves the second-order Taylor expansion of the penalized pseudo-likelihood in neighborhoods, the next lemma translates this result to the penalized pseudo-likelihood itself.

Lemma 5
Let $\mathbf{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$ *be p-variate 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 penalized pseudo-likelihood estimator* $\widehat{\boldsymbol{\Theta}}(\lambda)$. *Then there exist* $\delta_{\boldsymbol{\theta}^{(0)}} > 0$ *and* $\alpha_{min,\, \boldsymbol{\theta}^{(0)}} \geq p$ *such that for all,*

$$\boldsymbol{\psi}^{(0)} \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}\|_2 < \delta_{\boldsymbol{\theta}^{(0)}}, \boldsymbol{\theta} \in \mathbb{R}^q\} \quad \text{and} \quad \alpha \geq \alpha_{min}, \tag{36}$$

a single step of Algorithm (1) initiated by $\boldsymbol{\psi}^{(0)}$ *yields an updated parameter estimate* $\boldsymbol{\psi}^{(1)}$ *that satisfies for some* $\varepsilon_{\alpha, \boldsymbol{\theta}^{(0)}} > 0$*:*

$$\mathcal{L}_{penPL}(\boldsymbol{\psi}^{(1)}) \geq \mathcal{L}_{penPL}(\boldsymbol{\psi}^{(0)}) + \varepsilon_{\alpha, \boldsymbol{\theta}^{(0)}}. \tag{37}$$

Proof The proof uses inequality (26) of Lemma (4) to show, when α exceeds some lower bound, it also holds for the update $\frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j$. The proof then arrives at inequality (37) through the expression of the second-order Taylor approximation $L(\cdot; \boldsymbol{\psi}^{(0)})$ in terms of $\mathcal{L}_{\text{penPL}}$ with an upper bound on the Lagrange remainder of $L(\cdot; \boldsymbol{\psi}^{(0)})$.

For the first part, suppose that $\boldsymbol{\psi}^{(0)} \in \mathbb{R}^q$ with $||\boldsymbol{\psi}^{(0)} - \boldsymbol{\theta}^{(0)}||_2 < \delta_p$. Then, for $\alpha \geq p$ and by the concavity of the second-order Taylor approximation $L(\boldsymbol{\theta}; \boldsymbol{\psi}^{(0)})$, it holds that:

$$\begin{aligned} L\big(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j; \boldsymbol{\psi}^{(0)}\big) &= L\Big[\Big(1 - \frac{p}{\alpha}\Big)\boldsymbol{\psi}^{(0)} + \frac{p}{\alpha}\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{p}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j\Big); \boldsymbol{\psi}^{(0)}\Big] \\ &\geq \Big(1 - \frac{p}{\alpha}\Big)L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) + \frac{p}{\alpha}L\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{p}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j; \boldsymbol{\psi}^{(0)}\Big). \end{aligned} \tag{38}$$

Substitution of inequality (26) of Lemma (4) into the preceeding display then gives, for some $\delta_p = \delta_{\boldsymbol{\theta}^{(0)}}$ and $\epsilon_{p, \boldsymbol{\theta}^{(0)}} > 0$,

$$\begin{aligned} L\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j; \boldsymbol{\psi}^{(0)}\Big) &\geq \Big(1 - \frac{p}{\alpha}\Big)L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) + \frac{p}{\alpha}L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) + \frac{p}{\alpha}\varepsilon_{p, \boldsymbol{\theta}^{(0)}} \\ &= L(\boldsymbol{\psi}^{(0)}; \boldsymbol{\psi}^{(0)}) + \frac{p}{\alpha}\varepsilon_{p, \boldsymbol{\theta}^{(0)}}. \end{aligned} \tag{39}$$

By the continuity and smoothness of $\mathcal{L}_{\text{penPL}}$, there exists some $C > 0$ such that the Lagrange remainder of $L(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j; \boldsymbol{\psi}^{(0)})$ can be bounded to produce:

$$\mathcal{L}_{\text{penPL}}\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j\Big) \geq L\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j; \boldsymbol{\psi}^{(0)}\Big) - C\Big|\Big|\frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j\Big|\Big|_2^2. \tag{40}$$

Let R be the radius of a sphere that contains the compact parameter space in $\mathbb{R}^q$. The triangle inequality then yields: $\|\frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j\|_2^2 \leq \alpha^{-2}(2pR)^2$. Substitute this and inequality (40) into display (39) to find:

$$\mathcal{L}_{\text{penPL}}\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j\Big) \geq \mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^{(0)}) + \frac{p}{\alpha}\varepsilon_{p, \boldsymbol{\theta}^{(0)}} - \frac{1}{\alpha^2}C(2pR)^2 \tag{41}$$

Finally, $\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j$ must strictly increase $\mathcal{L}_{\text{penPL}}$ for all $\boldsymbol{\psi}^{(0)} \in \mathbb{R}^q$ with $\|\boldsymbol{\psi}^{(0)} - \boldsymbol{\theta}^{(0)}\|_2 < \delta_{\boldsymbol{\theta}^{(0)}}$. Preserving half of $\frac{p}{\alpha}\varepsilon_{p, \boldsymbol{\theta}^{(0)}}$, this requires $\frac{1}{2}p\alpha^{-1}\varepsilon_{p, \boldsymbol{\theta}^{(0)}} \geq \alpha^{-2}C(2pR)^2$. When combined with the assumption that $\alpha \geq p$ this yields: $\alpha_{\min,\, \boldsymbol{\theta}^{(0)}} \geq \max\{p, 8pCR^2\varepsilon_{p, \boldsymbol{\theta}^{(0)}}^{-1}\}$. Then, for all $\boldsymbol{\psi}^{(0)} \in \mathbb{R}^q$ with $\|\boldsymbol{\psi}^{(0)} - \boldsymbol{\theta}^{(0)}\|_2 < \delta_{\boldsymbol{\theta}^{(0)}}$ and $\alpha \geq \alpha_{\min,\, \boldsymbol{\theta}^{(0)}}$, a single step of Algorithm (1) initiated with $\boldsymbol{\psi}^{(0)}$ generates a collection block-wise updates $\{\tilde{\boldsymbol{\psi}}_j\}_{j\in\mathcal{V}}$ that satisfy:

$$\mathcal{L}_{\text{penPL}}\Big(\boldsymbol{\psi}^{(0)} + \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\psi}}_j\Big) \geq \mathcal{L}_{\text{penPL}}(\boldsymbol{\psi}^{(0)}) + \frac{p}{2\alpha}\varepsilon_{p, \boldsymbol{\theta}^{(0)}}, \tag{42}$$

which concludes the proof. □

Supplementary Material G: Estimating the variance of conditionally Gaussian variates

The pairwise MRF distribution assumes that all Gaussian variates have known variance (being equal to one). We here propose a modification of the algorithm to estimate the Gaussian variances for two reasons: 1) to allow the variance of the Gaussian variates to be unknown (instead of being known), and 2) such that the algorithm estimates the precision matrix when all variates are conditionally Gaussian for convenience of the user. The precision matrix of a multivariate normal distribution has the reciprocal of the conditional variance on its diagonal. So far for the pairwise MRF distribution, the diagonal of $\boldsymbol{\Theta}$ does not represent the reciprocal of the conditional variance. Instead, the proposed algorithm estimates the parameter $\boldsymbol{\Theta}$ of the pairwise MRF distribution where the diagonal represents one of the elements in the sum of the link function ($\boldsymbol{\Theta}_{j,j}$ in the link function $\eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j}) = \boldsymbol{\Theta}_{j,j} + \sum_{j'\in\mathcal{V},j'\neq j}\boldsymbol{\Theta}_{j,j'}T_{j'}(Y_{j'})$ for variate $j \in \mathcal{V}$). The algorithm thus only estimates the conditional mean of all Gaussian variates, represented by the link function. To additionally estimate the reciprocal conditional variance of the Gaussian variables – the diagonal of the precision matrix – we introduce an additional step to the algorithm (line 9).

Let $\boldsymbol{\Omega}_j$ for $j \in \mathcal{V}_G$ be the reciprocal conditional variance of the conditionally Gaussian variates $\mathcal{V}_G$ and recall the link function of the conditional distributions of the pairwise MRF distribution,

$$\eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j}) = \boldsymbol{\Theta}_{j,j} + \sum\nolimits_{j'\in\mathcal{V},j'\neq j}\boldsymbol{\Theta}_{j,j'}T_{j'}(Y_{j'}). \tag{43}$$

The conditional distribution of Gaussian variates in the pairwise MRF distribution must be a normal distribution,

$$Y_j \,|\, \mathbf{Y}_{\setminus j} \sim \mathcal{N}[(\boldsymbol{\Omega}_j)^{-1}\eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j}),(\boldsymbol{\Omega}_j)^{-1}],$$

where $\text{Var}(Y_j \,|\, \mathbf{Y}_{\setminus j}) = (\boldsymbol{\Omega}_j)^{-1}$ with $\boldsymbol{\Omega}_j$ the reciprocal of the conditional variance (being the diagonal of the precision matrix). Then:

$$P(Y_j \,|\, \mathbf{Y}_{\setminus j}) \propto (\boldsymbol{\Omega}_j)^{1/2}\exp\{-\tfrac{1}{2}[Y_j - (\boldsymbol{\Omega}_j)^{-1}\eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]^2\boldsymbol{\Omega}_j\}.$$

Let us now consider the pseudo-loglikelihood using the preceeding display, giving (temporarily ignoring the sum over the samples):

$$\log[P(Y_j \,|\, \mathbf{Y}_{\setminus j})] \propto \tfrac{1}{2}\log(\boldsymbol{\Omega}_j) - \tfrac{1}{2}[Y_j\boldsymbol{\Omega}_j - \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]^2(\boldsymbol{\Omega}_j)^{-1}.$$

We then maximize the pseudo-likelihood in $\boldsymbol{\Omega}_j$:

$$\begin{aligned}\frac{d}{d\boldsymbol{\Omega}_j}\sum\nolimits_{j=1}^{p}\log[P(Y_j \,|\, \mathbf{Y}_{\setminus j})] &= \frac{d}{d\boldsymbol{\Omega}_j}\Big(\tfrac{1}{2}\log(\boldsymbol{\Omega}_j) - \tfrac{1}{2}[Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]^2(\boldsymbol{\Omega}_j)^{-1}\Big) \\ &= \frac{1}{2\boldsymbol{\Omega}_j} + \tfrac{1}{2}[Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]^2(\boldsymbol{\Omega}_j)^{-2} - [Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]\frac{Y_j}{\boldsymbol{\Omega}_j}\end{aligned} \tag{44}$$

Finally, set the above equation to zero and solve for $\boldsymbol{\Omega}_j$,

$$\begin{aligned}\frac{1}{2\boldsymbol{\Omega}_j} + \tfrac{1}{2}[Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]^2(\boldsymbol{\Omega}_j)^{-2} - [Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]\frac{Y_j}{\boldsymbol{\Omega}_j} &= 0 \\ \boldsymbol{\Omega}_j + [Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})]^2 - 2[Y_j\boldsymbol{\Omega}_j + \eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})](Y_j\boldsymbol{\Omega}_j) &= 0 \\ \boldsymbol{\Omega}_j + \big(\eta_j(\boldsymbol{\Theta}_{j,\setminus j};\mathbf{Y}_{\setminus j})\big)^2 - \boldsymbol{\Omega}_j^2Y_j^2 &= 0\end{aligned} \tag{45}$$

This is a simple quadratic equation, and straightforwardly solved. We then modify the algorithm to analytically find the $\{\boldsymbol{\Omega}_j\}_{j\in\mathcal{V}}$ at the end of each step (after line 8, Algorithm 1).

Supplementary Material H: Optimizing the step-size via the multiplier α

Finally, it is shown that the choice of the block updates multiplier in the parallelized algorithm of the ridge pseudo-likelihood estimator can be relaxed, which enhances convergence. Lemma (7) states the result, which uses Lemma (6) as a prerequisite.

Lemma 6
Let $\mathbf{H}$ *be the Hessian of the ridge-penalized pseudo-likelihood* $\mathcal{L}_{penPL}$ *and its submatrices* $\mathbf{H}_j$ *obtained by restriction to the elements related to the* j*-th variate. Define the* $q \times q$ *diagonal matrix* $\mathbf{G}$ *with elements:*

$$\mathbf{G}_{q,q} = \begin{cases} \mathbf{H}_{q,q} & \text{if } \boldsymbol{q} = (j,j') \in \mathcal{Q} \text{ such that } j \neq j', \\ 0 & \text{if } \boldsymbol{q} = (j,j') \in \mathcal{Q} \text{ such that } j = j'. \end{cases} \tag{46}$$

Consider a vector $\mathbf{r} \in \mathbb{R}^q$ *and define the (overlapping) subvectors* $\{\mathbf{r}_j\}_{j \in \mathcal{V}}$ *which are obtained through subsetting* $\mathbf{r}$ *to the elements corresponding to the* j*-th variate. Then:*

$$\mathbf{r}^\top \mathbf{H} \mathbf{r} = \sum_{j \in \mathcal{V}} \mathbf{r}_j^\top \mathbf{H}_j \mathbf{r}_j - \mathbf{r}^\top \mathbf{G} \mathbf{r}. \tag{47}$$

Moreover, $\mathbf{G}$ *is negative semi-definite and* $\mathbf{H}_j$ *is negative definite for all* $j \in \mathcal{V}$.

Proof The negative (semi-)definiteness of the principal submatrix $\mathbf{H}_j$ and $\mathbf{G}$ are immediate from their definition and $\mathbf{H} \prec 0$. Rests to show identity (47). To this end observe that the structure of the pseudo-likelihood and that of the ridge penalty imply that $(\mathbf{H})_{\boldsymbol{q}_1, \boldsymbol{q}_2} = 0$ for $\boldsymbol{q}_1 = (j_1', j_1'), \boldsymbol{q}_2 = (j_2, j_2') \in \mathcal{Q}$ such that $\{j_1, j_1'\} \cap \{j_2, j_2'\} = \emptyset$. This simplifies $\mathbf{r}^\top \mathbf{H} \mathbf{r}$ to:

$$\mathbf{r}^\top \mathbf{H} \mathbf{r} = \sum_{\substack{\{\boldsymbol{q}_1 = (j_1, j_1'), \boldsymbol{q}_2 = (j_2, j_2') \in \mathcal{Q}: \\ \{j_1, j_1'\} \cap \{j_2, j_2'\} \neq \emptyset\}}} \mathbf{r}_{\boldsymbol{q}_1} (\mathbf{H})_{\boldsymbol{q}_1, \boldsymbol{q}_2} \mathbf{r}_{\boldsymbol{q}_2}.$$

On the other hand,

$$\sum_{j \in \mathcal{V}} \mathbf{r}_j^\top \mathbf{H}_j \mathbf{r}_j = \sum_{\{\boldsymbol{q} = (j,j') \in \mathcal{Q}, j \neq j'\}} \mathbf{r}_{\boldsymbol{q}} (\mathbf{H})_{\boldsymbol{q},\boldsymbol{q}} \mathbf{r}_{\boldsymbol{q}} + \sum_{\substack{\{\boldsymbol{q}_1 = (j_1, j_1'), \boldsymbol{q}_2 = (j_2, j_2') \in \mathcal{Q}: \\ \{j_1, j_1'\} \cap \{j_2, j_2'\} \neq \emptyset\}}} \mathbf{r}_{\boldsymbol{q}_1} (\mathbf{H})_{\boldsymbol{q}_1, \boldsymbol{q}_2} \mathbf{r}_{\boldsymbol{q}_2}.$$

Combine the latter two identities to find:

$$\sum_{j \in \mathcal{V}} \mathbf{r}_j^\top \mathbf{H}_j \mathbf{r}_j - \mathbf{r}^\top \mathbf{H} \mathbf{r} = \sum_{\{\boldsymbol{q} = (j,j') \in \mathcal{Q}, j \neq j'\}} \mathbf{r}_{\boldsymbol{q}} (\mathbf{H})_{\boldsymbol{q},\boldsymbol{q}} \mathbf{r}_{\boldsymbol{q}} \;=\; \mathbf{r}^\top \mathbf{G} \mathbf{r},$$

which proves identity (47). □

Lemma 7
Let $\mathbf{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$ *be p-variate 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 ridge pseudo-likelihood estimator* $\widehat{\boldsymbol{\Theta}}^{ridge}(\lambda)$. *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:*

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \tfrac{1}{\alpha} \sum_{j \in \mathcal{V}} \tilde{\boldsymbol{\theta}}_j.$$

Next, define the p-dimensional difference vectors $\{\boldsymbol{\delta}_j\}_{j \in \mathcal{V}}$ *with elements:*

$$(\boldsymbol{\delta}_j)_{j'} = \begin{cases} (\boldsymbol{\theta}_{j'})_j - (\boldsymbol{\theta}_j)_{j'} & \text{if } j \neq j' \\ -(\boldsymbol{\theta}_j)_{j'} & \text{if } j = j' \end{cases} \qquad \text{for all } j, j' \in \mathcal{V}. \tag{48}$$

Let $L(\cdot; \boldsymbol{\theta}^{(0)})$ *be the second-order Taylor approximation of* $\mathcal{L}_{penPL}$ *at* $\boldsymbol{\theta}^{(0)}$. *Then,*

$$L(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) > L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)}), \tag{49}$$

if,

$$\alpha \geq 3 + \frac{3}{2} \cdot \frac{\sum_{j \in \mathcal{V}} \boldsymbol{\delta}_j^\top \mathbf{H}_j \boldsymbol{\delta}_j}{\sum_{j \in \mathcal{V}} \boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j}. \tag{50}$$

Proof The proof reformulates (59) as

$$L(\boldsymbol{\theta}^{(1)};\boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)}) = \boldsymbol{\Delta}^\top \frac{d\mathcal{L}_{\text{penPL}}}{d\boldsymbol{\theta}} + \frac{1}{2}\boldsymbol{\Delta}^\top \mathbf{H}\boldsymbol{\Delta}, \tag{51}$$

where $\boldsymbol{\Delta} = \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\theta}}_j$. It then proceeds in a manner similar to the proof of Lemma (3) by expressing the right-hand side of (51) in terms of the $\{\mathbf{H}_j\}_{j\in\mathcal{V}}$, followed by the derivation of condition (50) on the multiplier α for the difference $L(\boldsymbol{\theta}^{(1)};\boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)})$ to be positive.

First, we derive a lower bound for the quadratic form $\boldsymbol{\Delta}^\top\mathbf{H}\boldsymbol{\Delta}$ of the second-order Taylor approximation difference (51). It crucially relies on the key identity (Lemma 6):

$$\boldsymbol{\Delta}^\top\mathbf{H}\boldsymbol{\Delta} = \sum_{j\in\mathcal{V}} \boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j - \boldsymbol{\Delta}^\top\mathbf{G}\boldsymbol{\Delta} \;\geq\; \sum_{j\in\mathcal{V}} \boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j, \tag{52}$$

where $\{\boldsymbol{\Delta}_j\}_{j\in\mathcal{V}}$ are subvectors of $\boldsymbol{\Delta}$ obtained through restriction of the latter to the elements relating to the j-th variate, and $\mathbf{G}$ is a negative semi-definite matrix as defined in Lemma (6).

Next, we seek a lower bound for $\boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j$ in terms of $\boldsymbol{\theta}_j$ and $\boldsymbol{\delta}_j$. To this end note that the latter three are related by the identity: $\boldsymbol{\Delta}_j = \frac{1}{\alpha}(2\boldsymbol{\theta}_j + \boldsymbol{\delta}_j)$ for $j \in \mathcal{V}$. Substitute this identity in $\boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j$ and obtain:

$$\boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j = \frac{1}{\alpha^2}(4\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j + 4\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j + \boldsymbol{\delta}^\top\mathbf{H}_j\boldsymbol{\delta}_j). \tag{53}$$

Being a principal submatrix of the negative definite $\mathbf{H}$, $\mathbf{H}_j$ is also negative definite. This warrants:

$$4\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j > 4\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j + 2(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j)^\top\mathbf{H}_j(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j) \;=\; 2\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j + 2\boldsymbol{\delta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j.$$

Combine the result of the preceeding display and that of (53) to arrive at a lower bound for $\boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j$:

$$\boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j > \frac{1}{\alpha^2}(6\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j + 3\boldsymbol{\delta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j). \tag{54}$$

Finally, use this in (52) to obtain a lower bound for $\boldsymbol{\Delta}^\top\mathbf{H}\boldsymbol{\Delta}$ in terms of the $\boldsymbol{\theta}_j$ and $\boldsymbol{\delta}_j$,

$$\boldsymbol{\Delta}^\top\mathbf{H}\boldsymbol{\Delta} \geq \sum_{j\in\mathcal{V}} \boldsymbol{\Delta}_j^\top\mathbf{H}_j\boldsymbol{\Delta}_j \;>\; \frac{1}{\alpha^2}\sum_{j\in\mathcal{V}}(6\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j + 3\boldsymbol{\delta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j). \tag{55}$$

Now express $\boldsymbol{\Delta}^\top(\partial\mathcal{L}_{\text{penPL}}/\partial\boldsymbol{\theta})$ in terms of the $\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j$. To this end notice that,

$$\begin{aligned}\sum_{j\in\mathcal{V}} \boldsymbol{\theta}_j^\top \frac{\partial\mathcal{L}_{\text{penPL}}}{\partial\boldsymbol{\theta}_j} &= \sum_{j\in\mathcal{V}} (\boldsymbol{\theta}_j)_j \frac{\partial\mathcal{L}_{\text{penPL}}}{\partial(\boldsymbol{\theta}_j)_j} + \sum_{\substack{(j,j')\in\mathcal{Q},\\ j\neq j'}} [(\boldsymbol{\theta}_j)_{j'} + (\boldsymbol{\theta}_{j'})_j]\frac{\partial\mathcal{L}_{\text{penPL}}}{\partial(\boldsymbol{\theta}_j)_{j'}} \\ &= \alpha\sum_{q\in\mathcal{Q}} \boldsymbol{\Delta}_q \frac{\partial\mathcal{L}_{\text{penPL}}}{\partial\boldsymbol{\theta}_q} \;=\; \alpha\,\boldsymbol{\Delta}^\top\frac{\partial\mathcal{L}_{\text{penPL}}}{\partial\boldsymbol{\theta}}.\end{aligned} \tag{56}$$

Substitution of the equality $-(\partial\mathcal{L}_{\text{penPL}}/\partial\boldsymbol{\theta}_j) = \mathbf{H}_j\boldsymbol{\theta}_j$ into (56) yields:

$$\boldsymbol{\Delta}^\top\frac{\partial\mathcal{L}_{\text{penPL}}}{\partial\boldsymbol{\theta}} = -\frac{1}{\alpha}\sum_{j\in\mathcal{V}} \boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j, \tag{57}$$

with which now all summands of the right-hand side of (51) have been expressed in terms of the $\mathbf{H}j$.

When (in)equality's (55) and (57) are used together in equation (51) followed by some grouping of terms, one arrives at:

$$L(\boldsymbol{\theta}^{(1)};\boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)}) > \frac{1}{\alpha^2}\sum_{j\in\mathcal{V}}\big[(3-\alpha)\boldsymbol{\theta}_j^\top\mathbf{H}_j\boldsymbol{\theta}_j + \frac{3}{2}\boldsymbol{\delta}_j^\top\mathbf{H}_j\boldsymbol{\delta}_j\big].$$

From the above it follows that $L(\boldsymbol{\theta}^{(1)};\boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(0)};\boldsymbol{\theta}^{(0)})$ is positive when α is chosen such that:

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

which is well-defined by the negative definiteness of the $\mathbf{H}_j$. □

The following corollary generalizes the previous result by updating the diagonal elements of the estimator more than once. This accounts for the intuition that all elements of the parameter should receive an update having the same size. So far, all off-diagonal elements $\boldsymbol{\Theta}_{j,j'}$ of the estimator get updated with two estimates $(\boldsymbol{\theta}_j)_{j'}$ and $(\boldsymbol{\theta}_{j'})_j$ coming from the two different coordinate blocks j and j'. In contrast, the diagonal elements $\boldsymbol{\Theta}_{j,j}$ of the estimator get a single update $(\boldsymbol{\theta}_j)_j$ coming from coordinate block j. Ideally, both updates $(\boldsymbol{\theta}_j)_{j'}$ and $(\boldsymbol{\theta}_{j'})_j$ for $j \neq j'$ align – at least to some degree, this makes the parallel approach work – resulting in an update $2(\boldsymbol{\theta}_{j'})_j$ for the off-diagonal elements. In contrast, the diagonal elements are always updated with $(\boldsymbol{\theta}_j)_j$. Then the off-diagonal elements are updated with double the step size of the diagonal elements. However, the block-wise Newton-Raphson solution tells us to update the off-diagonal with $(\boldsymbol{\theta}_j)_{j'}$ and the diagonal with $(\boldsymbol{\theta}_j)_j$. Thus, intuitively – and when the block-wise updates align – we could update the diagonal elements twice as well. The following corollary formalizes this idea by generalizing Lemma (7).

Corollary 5
Let $\mathbf{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$ be p-variate 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 ridge pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}^{ridge}(\lambda)$. 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:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \tfrac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\theta}}_j + \tfrac{\beta}{\alpha}\boldsymbol{\theta}_D,$$

where $\beta \in [0, 1]$ and $(\boldsymbol{\theta}_D)_{(j,j)} = (\boldsymbol{\theta}_j)_j$ for $j \in \mathcal{V}$ (the indices of the diagonal elements) and $(\boldsymbol{\theta}_D)_{(j,j')} = 0$ for $j \neq j' \in \mathcal{V}$ elsewhere. Next, define the p-dimensional difference vectors $\{\boldsymbol{\delta}_j\}_{j\in\mathcal{V}}$ with elements:

$$(\boldsymbol{\delta}_j)_{j'} = \begin{cases} (\boldsymbol{\theta}_{j'})_j - (\boldsymbol{\theta}_j)_{j'} & \text{if } j \neq j' \\ -(1-\beta)\cdot(\boldsymbol{\theta}_j)_{j'} & \text{if } j = j' \end{cases} \qquad \text{for all } j, j' \in \mathcal{V}. \tag{58}$$

Let $L(\cdot; \boldsymbol{\theta}^{(0)})$ be the second-order Taylor approximation of $\mathcal{L}_{penPL}$ at $\boldsymbol{\theta}^{(0)}$. Then,

$$L(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) > L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)}), \tag{59}$$

if,

$$\alpha \geq 3 + \frac{\sum_{j\in\mathcal{V}} \left[\beta\cdot(\boldsymbol{\theta}_j)_j \frac{\partial \mathcal{L}_{penPL}}{\partial(\boldsymbol{\theta}_j)_j} + \frac{3}{2}\boldsymbol{\delta}_j^\top \mathbf{H}_j \boldsymbol{\delta}_j\right]}{\sum_{j\in\mathcal{V}} \boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j}. \tag{60}$$

Proof The proof slightly modifies the proof of (7). Notice that now $\boldsymbol{\Delta} = \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = \frac{1}{\alpha}\sum_{j\in\mathcal{V}} \tilde{\boldsymbol{\theta}}_j + \frac{\beta}{\alpha}\boldsymbol{\theta}_D$ such that the identity $\boldsymbol{\Delta}_j = \frac{1}{\alpha}(2\boldsymbol{\theta}_j + \boldsymbol{\delta}_j)$ for $j \in \mathcal{V}$ remains valid. Then the lower bound (55) also holds in this case,

$$\boldsymbol{\Delta}^\top \mathbf{H} \boldsymbol{\Delta} \geq \sum_{j\in\mathcal{V}} \boldsymbol{\Delta}_j^\top \mathbf{H}_j \boldsymbol{\Delta}_j \;\; > \;\; \frac{1}{\alpha^2}\sum_{j\in\mathcal{V}} (6\boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j + 3\boldsymbol{\delta}_j^\top \mathbf{H}_j \boldsymbol{\delta}_j). \tag{61}$$

Analogous to (57) we get,

$$\boldsymbol{\Delta}^\top \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial \boldsymbol{\theta}} = -\frac{1}{\alpha}\sum_{j\in\mathcal{V}} \boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j + \frac{\beta}{\alpha}\sum_{j\in\mathcal{V}} (\boldsymbol{\theta}_j)_j \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial(\boldsymbol{\theta}_j)_j}, \tag{62}$$

and in identical fashion to the proof of Lemma (7) one arrives at,

$$L(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)}) > \frac{1}{\alpha^2}\sum_{j\in\mathcal{V}} \left[(3-\alpha)\boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j + \beta\cdot(\boldsymbol{\theta}_j)_j \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial(\boldsymbol{\theta}_j)_j} + \frac{3}{2}\boldsymbol{\delta}_j^\top \mathbf{H}_j \boldsymbol{\delta}_j\right].$$

From the above it follows that $L(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(0)}; \boldsymbol{\theta}^{(0)})$ is positive when α satisfies,

$$\alpha \geq 3 + \frac{\sum_{j\in\mathcal{V}} \left[\beta\cdot(\boldsymbol{\theta}_j)_j \frac{\partial \mathcal{L}_{\text{penPL}}}{\partial(\boldsymbol{\theta}_j)_j} + \frac{3}{2}\boldsymbol{\delta}_j^\top \mathbf{H}_j \boldsymbol{\delta}_j\right]}{\sum_{j\in\mathcal{V}} \boldsymbol{\theta}_j^\top \mathbf{H}_j \boldsymbol{\theta}_j}.$$

□

Notice that this choice for $\boldsymbol{\theta}^{(1)}$ and the lower bound for α simplify to the ones derived in Lemma (7) for $\beta = 0$. Similarly, we update the diagonal of the parameter of the pairwise MRF distribution twice if $\beta = 1$.

Supplementary Material I: Ensuring well-definedness of estimator via parameter constraints

j \ j'	Bernoulli	Poisson	Gaussian	Exponential
Bernoulli	$\boldsymbol{\Theta}_{j,j'} \in \mathbb{R}$	$\boldsymbol{\Theta}_{j,j'} \in \mathbb{R}$	$\boldsymbol{\Theta}_{j,j'} \in \mathbb{R}$	$\sum_{j \in \mathcal{V}_B} \max\{\boldsymbol{\Theta}_{j,j'}, 0\} < -\boldsymbol{\Theta}_{j,j'}$
Poisson		$\boldsymbol{\Theta}_{j,j'} \leq 0$	$\boldsymbol{\Theta}_{j,j'} = 0$	$\boldsymbol{\Theta}_{j,j'} \leq 0$
Gaussian			$\boldsymbol{\Theta}_G \prec 0$	$\boldsymbol{\Theta}_{j,j'} = 0$
Exponential				$\boldsymbol{\Theta}_{j,j'} \leq 0$

Table 2 Parameters constraints for a well-defined pairwise MRF distribution $P_{\boldsymbol{\Theta}}(\mathbf{Y})$ (Chen et al., 2015). Here, $\boldsymbol{\Theta}_G$ refers to the submatrix of $\boldsymbol{\Theta}$ that correspond to variates that are conditionally Gaussian distributed, while $\mathcal{V}_B \subset \mathcal{V}$ refers the subset of variates with a conditional Bernoulli distribution. In additional to the constraints in the table the rate parameters of variates with a conditional exponential distribution are required to be negative: $\boldsymbol{\Theta}_{j,j} < 0$.

Recall the parameter constraints of the pairwise MRF distribution parameter (Table 1, repeated here for convenience). The algorithm finds an estimator satisfying all parameter constraints – and thus a well-defined distribution – in several stages.

1. First the pseudo-likelihood is maximized without considering pairwise parameter constraints. As exceptions, the parameter elements between a variate $j \in \mathcal{V}$ that is conditionally Gaussian distributed and a variate $j' \in \mathcal{V}$ that is Poisson or Exponentially distributed are set to zero, as these cannot be nonzero ($\boldsymbol{\Theta}_{j,j'} = 0$). Also the requirement that $\boldsymbol{\Theta}_{j,j} < 0$ for variates $j \in \mathcal{V}$ that are conditionally exponentially distributed is ensured by introducing an border function in the form of an additional penalty to the pseudo-likelihood. This border function is given by $\kappa / \boldsymbol{\Theta}_{j,j}^2$ for small enough κ. This penalty diverges as $\boldsymbol{\Theta}_{j,j}$ approaches zero, ensuring that $\boldsymbol{\Theta}_{j,j}$ cannot be non-negative.
2. **Poisson-Exponential & Poisson-Exponential**. Next, the algorithm ensures that the estimator satisfies the parameter constraints between variates $j \neq j' \in \mathcal{V}$ that are conditionally Poisson or Exponentially distributed. To ensure that $\boldsymbol{\Theta}_{j,j'} \leq 0$, we set all parameters that violate this constraint to zero ($\boldsymbol{\Theta}_{j,j'} > 0 \rightarrow \boldsymbol{\Theta}_{j,j'} = 0$). This constraint is relaxed in the sense that $\boldsymbol{\Theta}_{j,j'}$ is free to become negative again at any time (in that case we know that $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}_{j,j'}} < 0$).
3. **Exponential & Bernoulli**. At this stage the estimator maximizes the Pseudo-likelihood under all the previously described constraints. Next, the algorithm ensures that the parameter satisfies the parameter constraints between variates that are conditionally Bernoulli and Exponentially distributed. For a variate $j \in \mathcal{V}$ that is conditionally Exponentially distributed and the set $\mathcal{V}_B \subset \mathcal{V}$ of conditional Bernoulli variates the parameter constraint requires $\sum_{j' \in \mathcal{V}_B} \max\{\boldsymbol{\Theta}_{j,j'}, 0\} < -\boldsymbol{\Theta}_{j,j}$. When this constraint is violated, we penalize all parameters $\boldsymbol{\Theta}_{j,j'} > 0$ with $j' \in \mathcal{V}_B$ using a ridge penalty $\kappa_{j,j'} \boldsymbol{\Theta}_{j,j'}^2$ for small $\kappa_{j,j'}$, individually increasing $\kappa_{j,j'}$ untill the constraint is satisfied. These constraints are not relaxed, and overshoot is limited by taking small increments of κ.
4. **Gaussian-Gaussian**. The final constraint the estimator has to satisfy is the negative definiteness constraint on the precision matrix $\boldsymbol{\Theta}_G$ of the conditionally Gaussian variates $\mathcal{V}_G \subset \mathcal{V}$. Simple sufficient (but not necessary) conditions to enforce negative definiteness are ensuring that $(\boldsymbol{\Theta}_G)_{j,j'} < 0$ for $j \neq j' \in \mathcal{V}_G$ or $\sum_{j' \in \mathcal{V}_G, j' \neq j} |(\boldsymbol{\Theta}_G)_{j,j'}| < |(\boldsymbol{\Theta}_G)_{j,j}|$ for $j \in \mathcal{V}_G$ (Koller and Friedman, 2009). When $\boldsymbol{\Theta}_G$ is not negative definite, we penalize the off-diagonal elements of $\boldsymbol{\Theta}_G$. We do this by introducing a ridge penalty that forces the off-diagonal elements to satisfy the sufficient constraints, similarly to the penalty described for the Exponential & Bernoulli constraint.

Supplementary Material J: Pseudo-code for Gibbs sampling

Data are drawn from the pairwise MRF distribution by means of a Gibbs sampler, which exploits the fact that the variates' conditional distributions are known univariate exponential family members. Details of the resulting Gibbs sampler are immediate from its pseudo-code (Algorithm S1). This algorithm is applied throughout to obtain synthetic data, using a burn-in period of length 5.000 and a thinning factor smaller than 1/500, such that samples are selected from the chain at least 500 iterations apart ensuring independence among samples (Chen et al., 2015).

input : n sample size;
p exponential family member;
parameter $\boldsymbol{\theta}$;
output: $n \times p$ dimensional data matrix $\mathbf{Y}$.

initialize $\mathbf{Y}^{(0)}$.
for $i = 1$ *to* $n_{burn\text{-}in} + nf_{thinning}$ **do**
  for $j \in \mathcal{V}$ **do**
    draw from $P(Y_j \,|\, \mathbf{Y}_{\setminus j})$ (an exponential family member).
  end
end
Remove the first $n_{\text{burn-in}}$ draws (representing the burn-in phase).
Select every f_{thinning}-th sample (thinning).

Algorithm S1: Pseudocode of the Gibbs sampler of the pairwise MRF distribution.

Supplementary Material K: Cross-validation

The penalty parameter λ of the ridge pseudo-likelihood estimator is selected using k-fold cross-validation. This amounts to dividing the n samples over k exhaustive and mutually exclusive groups. The samples from all-but-one of these groups are used to compute the estimator for a given choice of the penalty parameter. The performance of the obtained estimator is evaluated on the samples of the left-out group. This process is repeated k times, leaving each group out once, but using the same penalty parameter value throughout. The resulting k performances are averaged and this average is the estimated performance of the estimator for the employed penalty parameter value. The performance is assessed for a grid of penalty parameters. The value that yields the best estimated performance is considered optimal and used to obtain the final, optimal estimator. In the remainder the performance is evaluated with the pseudo-loglikelihood and $k = 10$ unless stated otherwise. The choice for $k = 10$ follows the recommendation provided in Hastie et al. [2001], where it is suggested to be a reasonable compromise between the resulting bias and variance of the cross-validated prediction error.

The simplicity and generality of cross-validation makes it an obvious choice for penalty parameter selection. Moreover, as cross-validation chooses a model with good prediction performance, it is also a natural choice for ridge-type estimators that do not induce sparsity. Alternatively, one may employ an information criterion, that balances model fit and complexity, for penalty parameter selection. The extended Bayesian information criterion (eBIC) has been proposed for model selection with high-dimensional data (Chen and Chen, 2008). Gao and Song [2010] generalized the eBIC for the sparse maximum pseudo-likelihood estimators. Here the application of eBIC for penalty parameter selection is currently hampered by the lack of a proper definition of model complexity for non-sparse ridge-type estimators.

Supplementary Material L: Sparsification

The maximum ridge pseudo-likelihood estimate of $\boldsymbol{\Theta}$ does not contain any zero's. Zero's can be inferred by a post-estimation sparsification procedure. Hereto we propose to use the off-the-shelf empirical Bayes methodology of Efron [2004]. It assumes that the absolute values of the off-diagonal elements of the standardized version of $\widehat{\boldsymbol{\Theta}}$ are i.i.d. following a two-component mixture: $f(\cdot) = \pi_0 f_0(\cdot) + (1 - \pi_0) f_1(\cdot)$ with mixing proportion π_0 and components $f_0(\cdot)$ and $f_1(\cdot)$. The first component, $f_0(\cdot)$, represents the distribution of the off-diagonal elements of the standardized $\widehat{\boldsymbol{\Theta}}$ corresponding to edges absent in the underlying conditional independence graph, while the second component, $f_1(\cdot)$, relates to its present edges. The parameter π_0, and densities $f(\cdot)$ and $f_0(\cdot)$ are all estimated, employing some convenient assumptions, from the histogram of the off-diagonal elements of the standardized version of $\widehat{\boldsymbol{\Theta}}$. With these estimates at hand the posterior probability of an edge in the underlying graph given the observed element of the standardized $\widehat{\boldsymbol{\Theta}}$, i.e. $P((j, j') \in \mathcal{E} \mid \widehat{\boldsymbol{\Theta}}_{j,j'})$, is calculated. This probability is used as the basis for sparsification: a high probability infers an edge, a low probability infers none. Moreover, this probability can be endowed with a local false discovery rate interpretation (see Efron, 2004).

Supplementary Material M: Supplementary figures

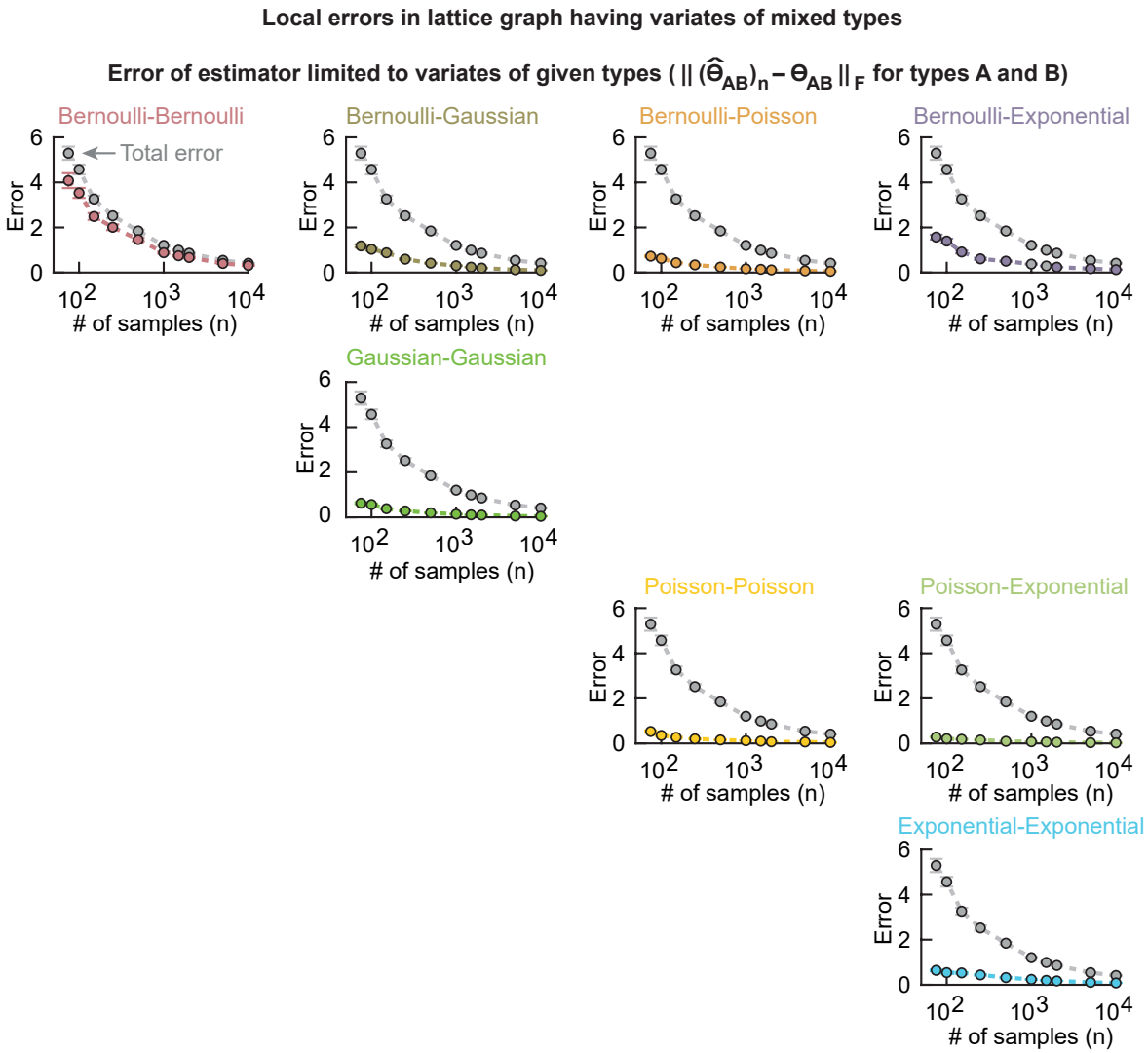

Figure S1. Data type specific error of estimator of lattice graph (Relates to Figure 2). Frobenius error of submatrix of estimator when restricted to variates of given types. We used the same data as in the lattice graph example in the main text (having all possible interactions of data types, Figure 2). Shown is the error ($||\widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}||_F$) of the unpenalized pseudo-likelihood estimator (grey, identical for all plots), together with the error of the submatrix corresponding to variates of two given types ($||\widehat{\boldsymbol{\Theta}}_{A,B} - \boldsymbol{\Theta}_{A,B}||_F$ with $\boldsymbol{\Theta}_{A,B}$ being the parameter $\boldsymbol{\Theta}$ restricted to the interactions between variates of type A and B, for example, all parameters that involve two Bernoulli variates). The error of the submatrix corresponding to parameters involving conditionally Bernoulli variates dominate the error. All curves show mean ± standard error of the mean (20 replicates per condition).

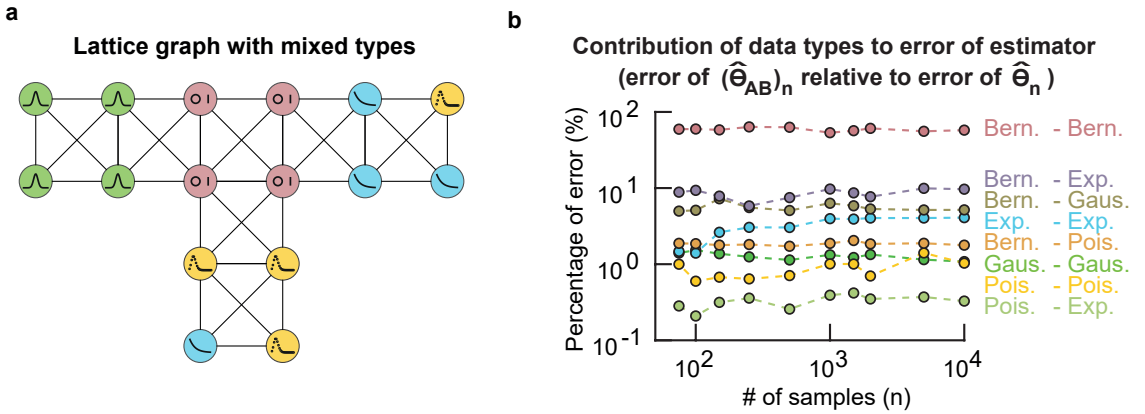

Figure S2. Bernoulli variates dominate error of estimator in lattice graph (Relates to Figure 2). (**a**) Recall the lattice graph network having all possible interactions between data types. (**b**) Summary of the data displayed in Figure S1. Contribution of interactions between given pairs of data types to the total error of the estimator. This is described by the error of the parameter restricted to given types relative to the error of the whole parameter (computed as the fraction of their squared Frobenius norms, $||\widehat{\mathbf{\Theta}}_{A,B} - \mathbf{\Theta}_{A,B}||_F^2 / ||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}||_F^2$ with $\mathbf{\Theta}_{A,B}$ being the parameter $\mathbf{\Theta}$ restricted to the interactions between variates of type A and B). The contributions of given interactions to the total error of the estimator seem independent of the given number of samples. Interactions involving variates that are conditionally Bernoulli constitute the far majority of the error, followed by variates that are conditionally Exponential. All curves show mean $\pm$ standard error of the mean (20 replicates per condition).

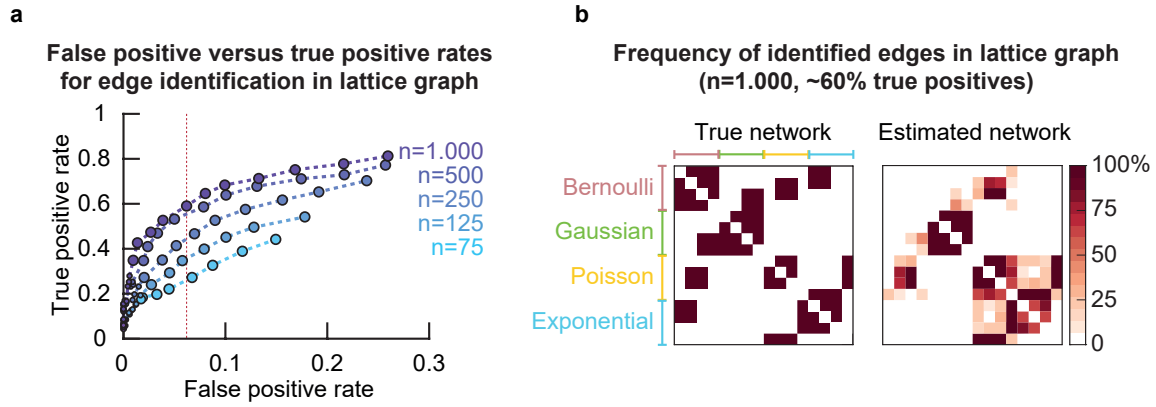

Figure S3. Bernoulli variates dominate misidentified edges in lattice graph with lasso estimator (Relates to Figure 2). We used the same data as in the lattice graph example in the main text (Figure 2) and $\tau = 10^{-10}$ as threshold for convergence. The lasso penalized pseudo-likelihood estimator was fitted to the data. The lasso penalty was approximated by a weighed ridge penalty with parameter λ, which was iteratively reweighed by re-evaluating the estimator $\hat{\boldsymbol{\theta}}^k$ from step k using an adjusted penalty $\lambda^{k+1} = \frac{\lambda^k}{|\hat{\boldsymbol{\theta}}^k_{j,j'}|}$ to find $\hat{\boldsymbol{\theta}}^{k+1}_{j,j'}$ at step $k+1$. The approximation then converges to the lasso estimator (we iterated till $||\hat{\boldsymbol{\theta}}^{k+1} - \hat{\boldsymbol{\theta}}^k||_F < 10^{-6}$) (Fan and Li [2001]). (**a**) False positive versus true positive rates (ROC curves) for edge identification in the lattice graph as function of the number of samples n. The false positive rate increases as the number of samples decreases. All curves show mean $\pm$ standard error of the mean (20 replicates per condition). (**b**) Here we studied the edge selection frequency in the elbow of the ROC curve (red dotted line in (**a**)) for $n = 1.000$ samples (penalty $\lambda \approx 0.126$). Shown are the true network adjacency matrix (left) and the average estimated adjacency matrix (right). The estimator poorly identifies edges of Bernoulli nodes. Moreover, the estimator incorrectly identifies edges that do not exist in the true network for mostly Poisson and Exponential nodes. In the former this is probably because the lasso penalty is too large, in the latter because the lasso penalty is too small. One way to remedy this unbalance in edge selection would be to use a types pair-specific penalty parameter. The differences in quality of edge selection between different data types is most likely due to differences in quality of the parameter estimates for different types as observed before (Chen et al., 2015). This matches our observation on the ridge estimator where, for example, the Bernoulli variates dominate the error of the estimator (Figure 2, Figure S2).

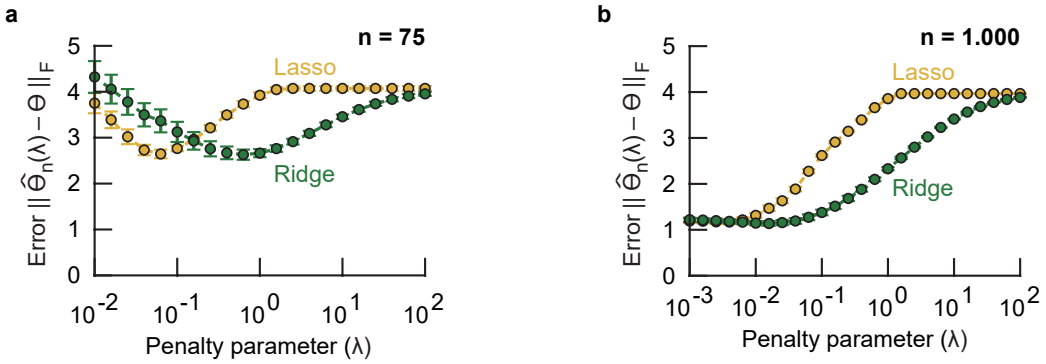

Figure S4. Comparison of the error of lasso and ridge estimators for the lattice graph (Relates to Figure 2). Scaling of the error of the penalized estimators with their penalty parameter λ. Here, the error is defined as the Frobenius norm of the difference between the parameter and its estimate $\|\widehat{\boldsymbol{\Theta}}(\lambda_{\text{opt}}) - \boldsymbol{\Theta}\|_F$. We used the same data as the examples in the main text (Figure 2), with $\tau = 10^{-10}$ as threshold for convergence. Shown is the error for the lasso (yellow) and ridge (green) pseudo-likelihood estimator for $n = 75$ (**a**) and $n = 1.000$ (**b**) samples. For both estimators and sample sizes, certain optimal λ minimizes the error. The optimal penalty minimizing the error yields estimators whose quality are indistinguishable (in terms of difference in error to the true parameter). The ($k = 10$) cross-validated ridge estimator selected $\lambda \approx 1$ ($n = 75$) or $\lambda = 0.063$ ($n = 1.000$). There is no obvious choice for an optimal lasso penalty in the (dense) lattice graph, as the error is large for the preferably low number of edges that would be selected. All curves show mean $\pm$ standard error of the mean (10 replicates per condition).

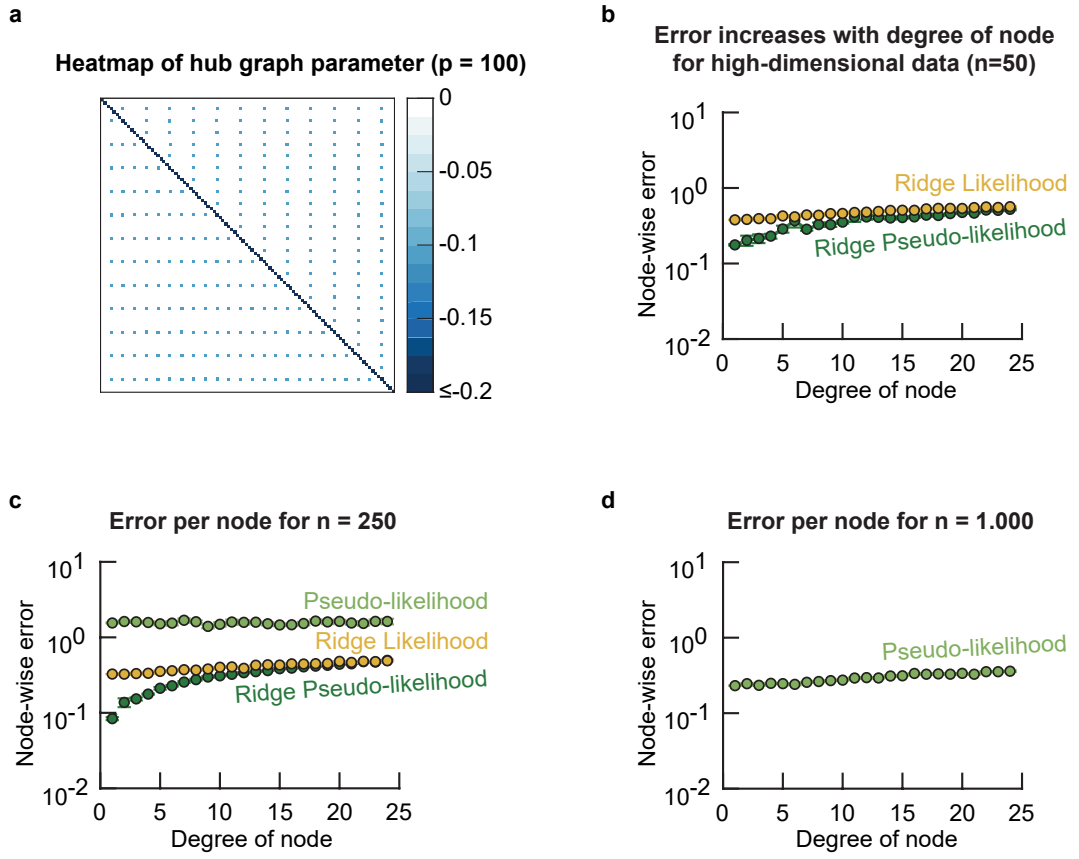

Figure S5. Node-wise error of the estimator increases with degree of a node (Relates to Figure 3). (**a**) We constructed a model having a hub-like structure, defined by the conditional independence graph as described by the non-zero off-diagonal elements in the pairwise MRF distribution parameter. Shown is the heatmap of the distribution parameter with $p = 100$ variates. Variates follow the multivariate normal distribution with unit-diagonal precision matrix $\mathbf{\Omega}$; Every eight variate (row-wise) is connected to every fourth variate (column-wise) by setting the elements to 0.1 ($\Omega_{j,j'} = -0.1 = \Omega_{j',j}$ for $j \mod 8 = 0$, $j' < j$ and $j' \mod 4 = 0$). All other entries of precision matrix are set equal to zero. The resulting graph contains hub nodes that are highly connected to their own set of sparsely connected nodes. (**b-d**) As in the comparison study of the main study, both the (ridge) precision estimator and the (ridge) pseudo-likelihood estimator were evaluated for each data set. The node-wise error is defined as the Frobenius norm of the diffence between the estimated and true parameter restricted to each variate, $\|\widehat{\mathbf{\Omega}}_{j,*}(\lambda_{\text{opt}}) - \mathbf{\Omega}_{j,*}\|_F$. Each configuration was run with 10 replicates, using $\tau = 10^{-10}$ as threshold for convergence. (**b-c**) The node-wise error for the hub-graph with (**b**) $n = 50$ samples (high-dimensional, penalized) and (**c**) $n = 250$ samples. The node-wise error increases with the node-degree in the underlying conditional independence graph for both the ridge likelihood and ridge pseudo-likelihood estimators. Thus, the total node-wise error is on average larger for more highly connected nodes. (**d**) node-wise error for the hub-graph for $n = 1000$ samples (low-dimensional, unpenalized). *Note:* The node-wise errors of the estimators are not independent. An error in a single parameter element contributes to the node-wise error of two nodes, e.g. a hub node and a sparsely connected node. Consequently, the fact that the node-wise error is dominated by the hub nodes is mostly an artefact: the node-wise error is dominated by a large error in a few elements of the parameter (those connecting a hub-node and a sparsely connected node). As a result, the ridge pseudo-likelihood seems better at estimating poorly connected nodes than the ridge likelihood, probably due to differences of the estimators.

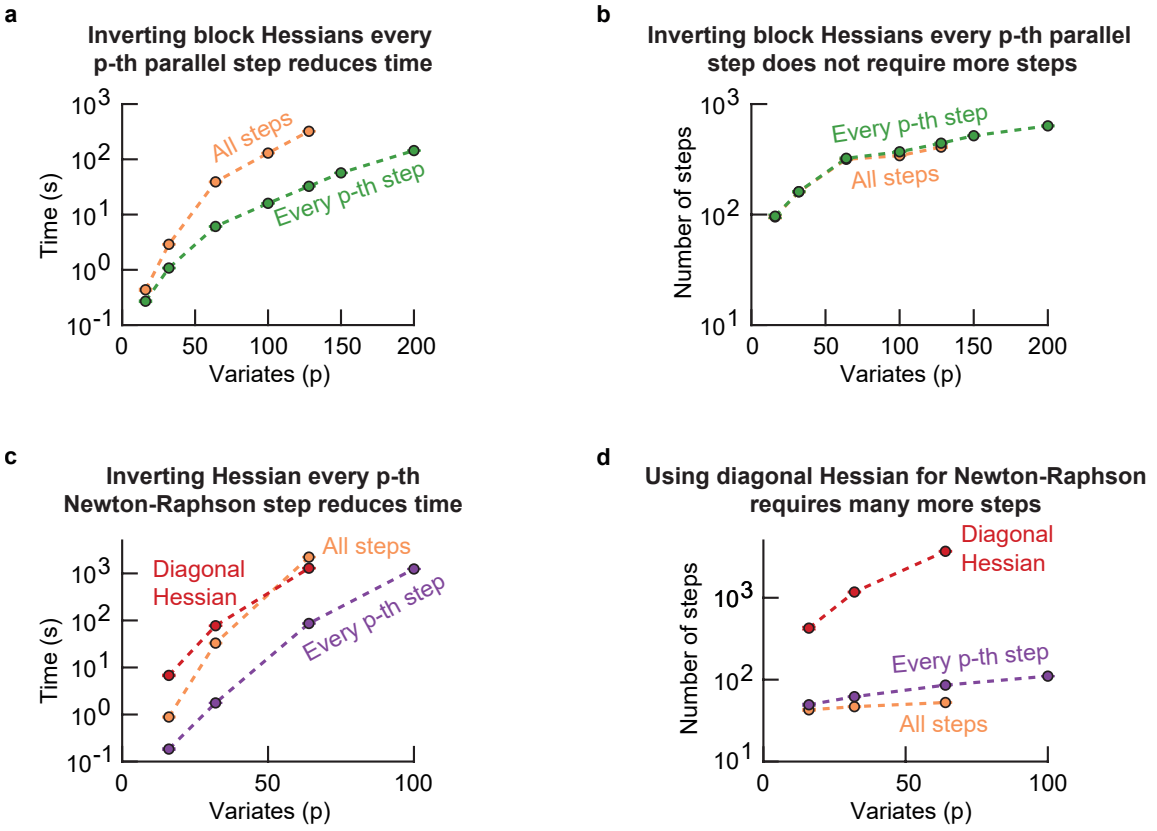

Figure S6. Reusing inverse Hessian for several steps speeds up algorithm (Relates to Figure 4). Justification of the approximations used in the benchmark study of our parallel algorithm (**a-b**) and the naive Newton-Raphson approach (**c-d**). We used the same data as in the benchmarking study in the main text (a Bernoulli-Gaussian pairwise MRF distribution, Figure 4), with $\tau = 10^{-10}$ as threshold for convergence. For the proposed parallel algorithm (**a-b**), the block Hessian matrices (each a $p \times p$ matrix) were inverted either at every step of the algorithm (orange) or every $k_0 = p$ steps (green). The simulation time (**a**) and required number of steps (**b**) for the parallel algorithm to converge are plotted against the number of variates. The runtime of the algorithm to find the pseudo-likelihood estimator reduces by about an order of magnitude when reusing the inverse block Hessian matrices. In contrast, the number of steps to find the pseudo-likelihood estimator barely increases when reusing the Hessians. Thus, by reusing the inverted block Hessians for consecutive steps significantly reduces computation time without hampering convergence. This approximation is used throughout all simulations. We considered the following approaches for the comparison with naive methods (**c-d**). The naive Newton-Raphson algorithm that inverts the full Hessian (a $\sim \frac{1}{2}p^2 \times \frac{1}{2}p^2$ matrix) at every step (orange) can be approximated by using a diagonal Hessian (red) or inverting the Hessian only every $k_0 = p$ steps (purple). The simulation time (**c**) and required number of steps (**d**) for the naive algorithm to converge are plotted against the number of variates. The diagonal Hessian is computationally cheap to compute but requires many steps for convergence. The diagonal Hessian approach is therefore unfavorable for even small problem sizes ($p > 32$). Inverting the full Hessian every step of the algorithm results in quadratic convergence, and is the preferred algorithm for small – in p – problem sizes. However, the computational complexity of inverting a $\sim \frac{1}{2}p^2 \times \frac{1}{2}p^2$ Hessian matrix dominates the computation time for a larger number of variates (e.g. $p \geq 100$). When using a naive approach, the preferred (and fastest) method is thus to invert the full Hessian matrix only every few steps of the algorithm and reuse that inverse for the remaining steps. In practice, this amounts to computing the inverse Hessian only once, as the algorithm then generally converges within p steps. Thus, we always use a Newton-Raphson algorithm that reuses the inverse Hessian for Benchmarking the proposed algorithm. Alternatively, we compared to the block-wise approach in which the parameter is updated sequentially in a block-wise fashion (Figure 4). All curves show mean ± standard error of the mean (5 replicates per condition, using data with $n = 1.000$ samples). Processor cores were artificially limited to 2 GHz for a fair comparison (SM N).

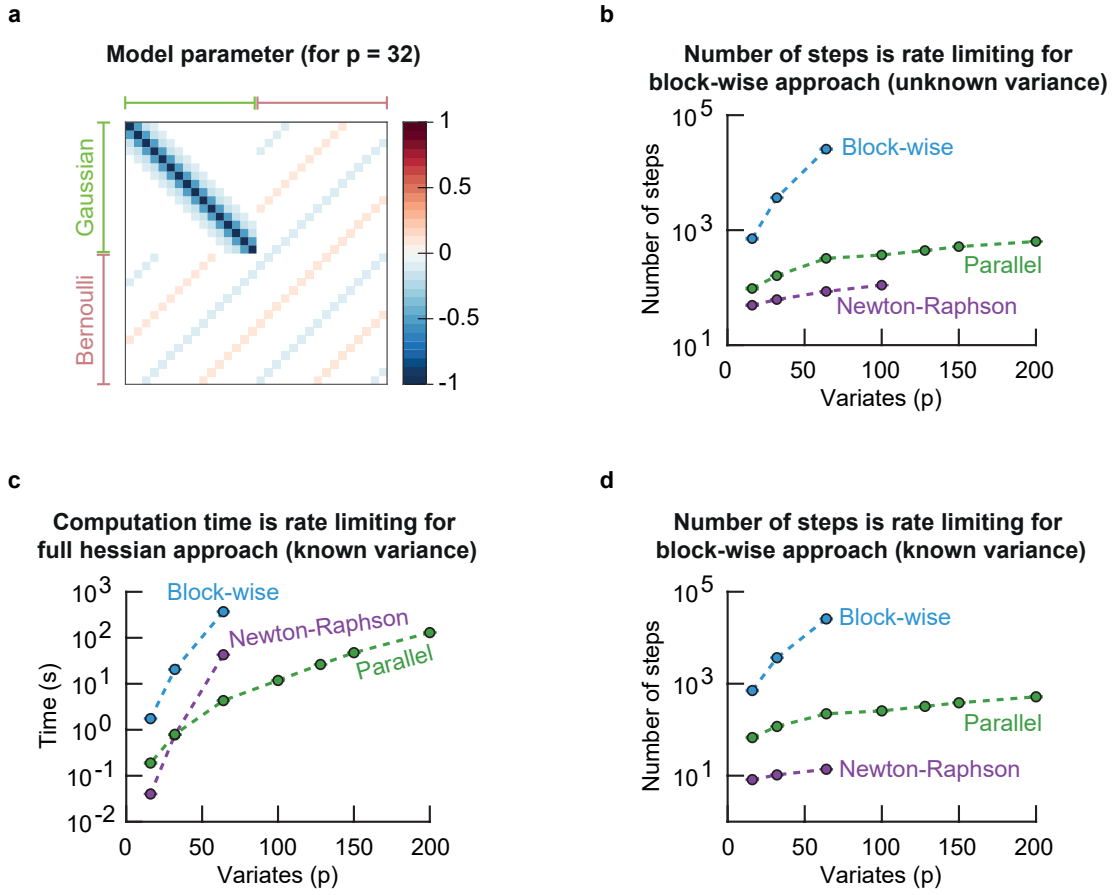

Figure S7. Proposed algorithm is faster for both known and unknown Gaussian variance (Relates to Figure 4). Further benchmarking of the proposed algorithm comparing known and unknown Gaussian variance. (**a**) Model parameter of the Bernoulli-Gaussian pairwise MRF distribution for $p = 32$ variates (16 Gaussian, 16 Bernoulli), analogously defined for other dimensions. The Gaussian variates have a three banded precision matrix with unit diagonal ($\mathbf{\Theta}_{j,j+1} = 0.5 = \mathbf{\Theta}_{j+1,j}$ for $j = 1, ..., p-1$, $\mathbf{\Theta}_{j,j+2} = 0.2 = \mathbf{\Theta}_{j+2,j}$ for $j = 1, ..., p-2$, $\mathbf{\Theta}_{j,j+3} = 0.1 = \mathbf{\Theta}_{j+3,j}$ for $j = 1, ..., p-4$ and all other entries equal to zero). Each Bernoulli variate has an interaction with every $\sqrt{p}$-th other variate, with the corresponding parameters set to alternating ± 0.1. We used the same data as in the benchmarking study in the main text (Figure 4), with $n = 1.000$ samples and $\tau = 10^{-10}$ as threshold for convergence. (**b**) Supplements to Figure 4c in the main text. The number of steps required to find the pseudo-likelihood estimator for the naive Newton-Raphson algorithm (purple), block-wise sequential updating (blue) and our proposed parallel algorithm (green). The variance of the Gaussian variates is unknown and also estimated. The number of steps is rate-limiting when sequentially updating the parameter block-wise, consistent with previous findings (Chen et al., 2015). (**c-d**) Same simulations as in Figure 4c. The simulation time (**c**) and required number of steps (**d**) as function of the number of variates for Newton-Raphson algorithm (purple), block-wise updating (blue) and parallel algorithm (green). Here the variance of the Gaussian variates is assumed to be known (assuming a unit diagonal for the precision matrix). Simulation time and number of steps are lower than for unknown variance, but differences between approaches are preserved. All curves show mean ± standard error of the mean (5 replicates per condition, using data with $n = 1.000$ samples). Processor cores were artificially limited to 2 GHz for a fair comparison (SM N).

Supplementary Material N: Computer components used in simulation study

All simulations were run on a laptop with the following specifications:

1. **processor** Intel(R) Core(TM) i7-8650U CPU (that we fixed to 2 GHz for comparable results)
2. **ram** 8.0 GB
3. **storage** SK Hynix sc311 sata 512 gb
4. **operating system** Windows 10 Enterprise, 64-bit
5. **RStudio** Version 1.2.1335 (R version 3.6.0)

Supplementary Material O: Real-world application to -omics data

We illustrate an application of the proposed methodology by means of learning the pairwise MRF distribution of (part of the) cellular regulatory network from omics data. For this we used the publicly available non-silent somatic mutation and gene expression data from the invasive breast carcinoma study of the Cancer Genome Atlas (TCGA) project (The Cancer Genome Atlas Network, 2012). The gene expression data comprises the RNA sequencing profiles of 445 patients with invasive breast carcinomas (BRCA) and are obtained via the `brcadat` dataset included in the R-package `XMRF` (Wan et al., 2016a). The thus obtained data set includes the expression levels (normalized mRNA read counts) of 353 genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2015). Subsequently, the RNA sequencing data are preprocessed as described in Allen and Liu, 2013 using the `processSeq` function from the `XMRF` package. This results in expression levels that – presumably – can be modeled with a Poisson distribution. Then, the top 15% genes with largest variance in expression levels across patients are retained. Next, the mutation data of 977 BRCA patients are obtained from the TCGA project and filtered with the `getTCGA`-function from the R-package `TCGA2STAT` (Wan et al., 2016b). The binary data indicate the presence of a mutation in the coding region of a gene. Furthermore, only mutation data of genes with mutations present in at least 5% of the patients are included. Finally, the mutation and expression data sets are merged at the gene level and for common patients using the `OMICSBind` function from the `TCGA2STAT`-package. The resulting final data matrix contains $p = 63$ variates comprising the non-silent somatic mutations of 11 genes (binary) and the expression level of 52 genes (counts) from $n = 433$ BRCA patients. The genes *CDH1* and *GATA3* have both mutation and expression data measured.

We can model BRCA data set with the pairwise MRF distribution having Bernoulli and Poisson distributions conditionally for mutation and expression variables respectively. Using the BRCA data set we computed the ridge pseudo-likelihood estimator with Algorithm 1. The ridge penalty parameter was selected with 5-fold cross-validation using a grid search over a range of penalty parameters maximizing the log-likelihood. The resulting estimator is visualized by means of a heatmap (Figure S7). Interaction parameters between genes' expression levels are always negative, due to the constraints for well-definedness. Moreover, the majority of strongest interactions (in absolute sense) relate gene expression levels to somatic mutations. Noteworthy are *TP53*, *PIK3CA* and *CDH1*. Non-silent somatic mutations in *TP53* – a well-known tumor suppressor gene (Levine et al., 1991) – or *PIK3CA* – a known oncogene that is often mutated in breast cancers (Cizkova et al., 2012) – seem strongly related to a change in the expression of genes that have been causally implicated with cancer (Forbes et al., 2015). This for example confirms the role of *TP53* as guardian of the genome (Lane, 1992). Finally, *CDH1* is included in the BRCA data with both its mutation and expression levels. The non-silent somatic mutation of *CDH1* is strongly related to its expression level, unsurprisingly indicating that its expression level is related to a mutation in the DNA template. We additionally estimated a network by sparsifying the ridge estimator via local FDR in a comparison with model selection by the lasso via BIC (Figure S8). We notice that the model complexity is very similar for the FDR sparsified (ridge) and BIC (lasso) networks. The ridge and lasso networks are completely disjoint, perhaps as lasso selects edges involving Poisson variates before Bernoulli variates (Figure S3).

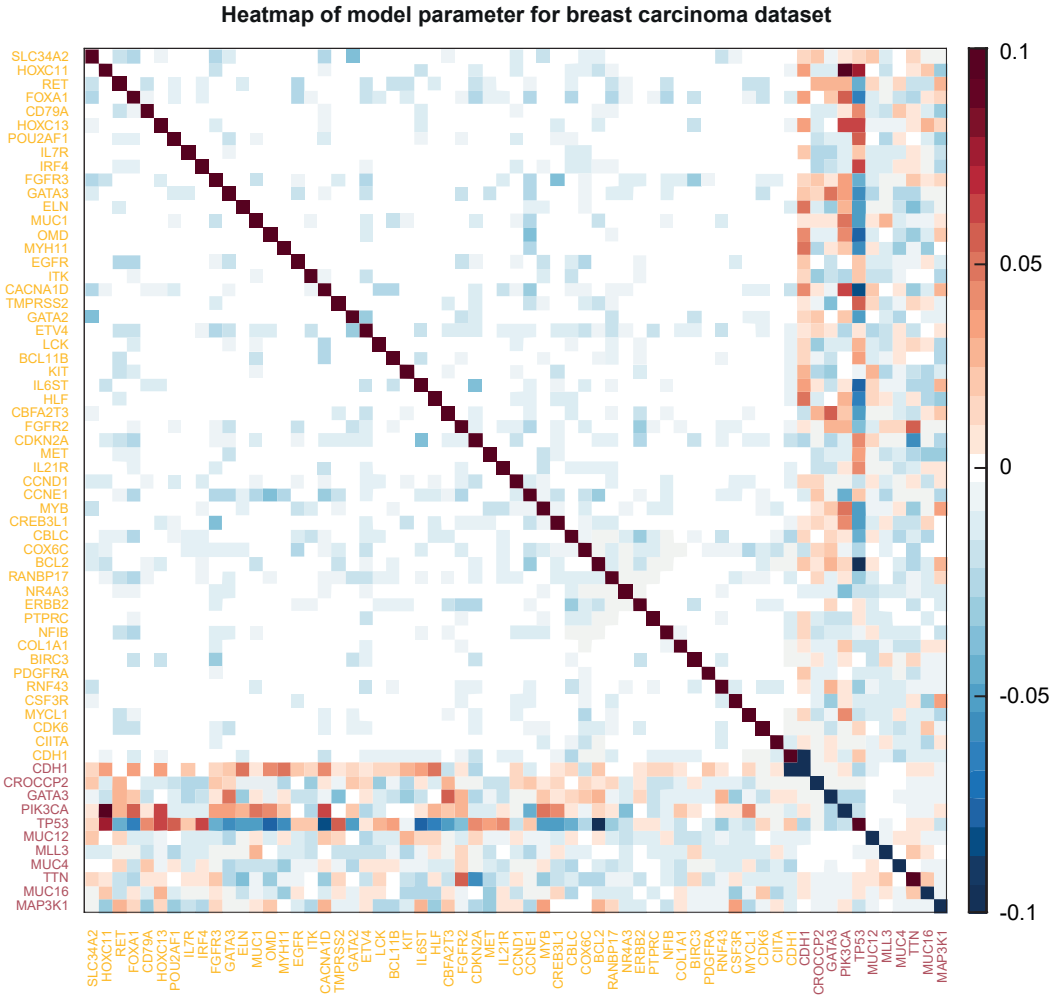

Figure S8. Heatmap of the ridge pseudo-likelihood estimator of the invasive breast carcinoma gene network. Non-silent somatic mutations in genes are labeled red, gene expression levels are labeled yellow. Positive interactions between genes are shaded red, negative interactions are shaded blue. Algorithm 1 was run using threshold $\tau = 10^{-6}$ (for sparsification, see Figure S9).

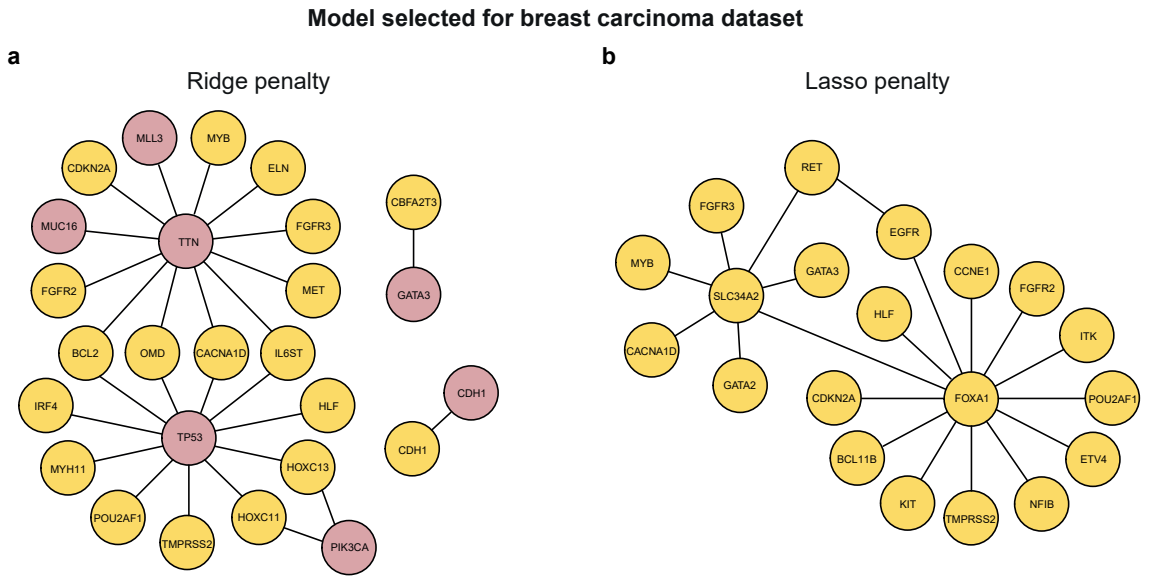

Figure S9. Models selected from ridge and lasso estimators (Relates to Figure 5). Selected networks of the ridge and lasso pseudo-likelihood estimators for the invasive breast carcinoma gene network. We used the same data as in the BRCA example (Figure S7), with $\tau = 10^{-6}$ as threshold for convergence. Nodes represent non-silent somatic mutations in genes (red) or gene expression levels (yellow). We fitted the pairwise MRF distribution parameter to the BRCA data to find the penalized pseudo-likelihood estimator. (**a**) Network for the ridge penalty. The ridge penalty $\lambda = 0.6832995$ was selected via 10-fold cross-validation (the final estimator is identical to the one represented by the heatmap in Figure S7). Model was selected by standardizing the estimator, after which the estimator was sparsified via local False Discovery Rate (a mixture model was fitted to the nonredundant partial correlations, and elements with a posterior probability larger than 0.999 were retained, see e.g. Schäfer and Strimmer, 2005). All of the selected edges involve at least one Bernoulli node, and many of the selected edges are the ones with highest edge weight (in absolute sense) in the heatmap (Figure 5). (**b**) Network for the lasso penalty. The lasso penalty was approximated with an iteratively reweighed ridge penalty (described in Figure S3). The penalty parameter ($\lambda = 1.5$) was chosen via the BIC for the pairwise MRF distribution (following Chen et al., 2015, $BIC_{\lambda,n} = -2n\mathcal{L}(\boldsymbol{\Theta}, \mathbf{Y})) + \log(n)|\boldsymbol{\Theta}|_0$ where $|\boldsymbol{\Theta}|_0$ is the number of selected edges). The selected edges only involve Poisson nodes.

References

G I Allen and Z Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on Nanobioscience*, 12(3):189–198, 2013.

A Anandkumar, V Y F Tan, F Huang, and A S Willsky. High-dimensional structure estimation in Ising models: Local separation criterion. *Annals of Statistics*, 40(3):1346–1375, 2012.

O Banerjee, L El Ghaoui, and A D'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

J Bento and A Montanari. Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*, pages 1303–1311, 2009.

J Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

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

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

J Chen and Z Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

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

M Cizkova, A Susini, S Vacher, G Cizeron-Clairac, C Andrieu, K Driouch, E Fourme, R Lidereau, and I Bieche. PIK3CA mutation impact on survival in breast cancer patients and in ERa, PR and ERBB2-based subgroups. *Breast Cancer Research*, 14(1), 2012.

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

B Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

J Fan and R Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

S A Forbes, D Beare, P Gunasekaran, K Leung, N Bindal, H Boutselakis, M Ding, S Bamford, C Cole, S Ward, C Y Kok, M Jia, T De, J W Teague, M R Stratton, U McDermott, and P J Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43:805–811, 2015.

J Friedman, T Hastie, and R Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

X Gao and P X-K Song. Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540, 2010.

T Hastie, R Tibshirani, and J Friedman. *The elements of Statistical Learning, second edition.* Springer, 2001.

H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10:883–906, 2009.

O Holder. Uber einen Mittelwertsatz. *Gottingen Nachrichten*, 2:38–47, 1889.

A Jalali, P Ravikumar, V Vasuki, and S Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 378–387, 2011.

R Jennrich. Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, 40(2):633–643, 1969.

J Jensen. Sur les fonctions convexes et les inegalites entre les valeurs moyennes. *Acta Mathematica*, 30: 175–193, 1906.

D Koller and N Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, Massachusetts, 2009.

D P Lane. Cancer. p53, guardian of the genome. *Nature*, 358:15–16, 1992.

S Lauritzen. *Graphical Models.* Oxford University Press, 1996.

S L Lauritzen and N Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17(1):31–57, 1989.

J Lee and T Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2013.

J D Lee, Y Sun, and J Taylor. On model selection consistency of m-estimators. *Electronic Journal of Statistics*, 9, 2015.

S-I Lee, H Lee, P Abbeel, and A Y Ng. Efficient L1-regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

S-I Lee, V Ganapathi, and D Koller. Efficient structure learning of markov networks using l1-regularization. In *Advances in Neural Information processing systems*, pages 817–824, 2007.

A J Levine, J Momand, and C A Finlay. The p53 tumour suppressor gene. *Nature*, 351(6326):453–6, 1991.

N Meinshausen and P Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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

P Ravikumar, M Wainwright, and J Lafferty. High-dimensional Ising model selection using L1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

P Ravikumar, M J Wainwright, G Raskutti, and B Yu. High-dimensional covariance estimation by minimizing L1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

J Schäfer and K Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

The Cancer Genome Atlas Network. Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

A Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998.

W N van Wieringen. The generalized ridge estimator of the inverse covariance matrix. *Journal of Computational and Graphical Statistics*, 28(4):932–942, 2019.

W N van Wieringen and C F W Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics & Data Analysis*, 103:284–303, 2016.

M J Wainwright, J D Lafferty, and P K Ravikumar. High-dimensional graphical model selection using l1-regularized logistic regression. In *Advances in Neural Information Processing Systems*, pages 1465–1472, 2007.

Y Wan, G I Allen, Y Baker, E Yang, P Ravikumar, M L Anderson, and Z Liu. XMRF: an R package to fit Markov networks to high-throughput genetics data. *BMC Systems Biology*, 10(69), 2016a.

Y Wan, G I Allen, and Z Liu. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, 32(6):952–954, 2016b.

E Yang, G Allen, Z Liu, and P K Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

E Yang, Y Baker, P Ravikumar, G Allen, and Z Liu. Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2014.