GRAPHICAL MODELS FOR DISCRETE AND CONTINUOUS DATA

By Johannes Lederer

Departments of Statistics and Biostatistics, University of Washington

Abstract We introduce a general framework for undirected graphical models. It generalizes Gaussian graphical models to a wide range of continuous, discrete, and combinations of different types of data. We also show that the models in the framework, called exponential trace models, are amenable to efficient estimation and inference based on maximum likelihood. As a consequence, we expect applications to a large variety of multivariate data that have correlated coordinates.

1. Introduction. Gaussian graphical models [7, 23, 35] describe the dependence structures in normally distributed random vectors. These models have become increasingly popular in the sciences, because their representation of the dependencies is lucid and can be readily estimated. For a brief overview, consider a random vector $X \in \mathbb{R}^p$ that follows a centered normal distribution with density

$$f_\Sigma(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}} e^{-x^\top \Sigma^{-1} x/2}$$

with respect to the Lebesgue measure, where the population covariance $\Sigma \in \mathbb{R}^{p \times p}$ is a symmetric and positive definite matrix. Gaussian graphical models associate these densities with a graph $(V, E)$ that has vertex set $V := \{1, \ldots, p\}$ and edge set $E := \{(i, j) : i, j \in \{1, \ldots, p\}, i \neq j, \Sigma^{-1}_{ij} \neq 0\}$. The graph encodes the dependence structure of $X$ in the sense that any two entries $X_i, X_j, i \neq j$, are conditionally independent given all other entries if and only if $(i, j) \notin E$. A natural and straightforward estimator of $\Sigma^{-1}$, and thus of $E$, is the inverse of the empirical covariance, which is also the maximum likelihood estimator. Inference can then be approached via the quantiles of the normal distribution.

The problem with Gaussian graphical models is that in practice, data often violate the normality assumption. Data can be discrete, heavy-tailed, restricted to positive values, or deviate from normality in other ways. Graphical models for some types of non-Gaussian data have been developed, including copula-based models [16, 27, 28, 39], Ising models [3, 24], and multinomial extensions of the Ising models [29]. However, there is no general
framework that comprises different data types including finite and infinite count data, potentially heavy-tailed continuous data, and combinations of discrete and continuous data and at the same time, ensures a rigid theoretical structure.

We make two main contributions in this paper:

- We formulate a general framework for undirected graphical models that both encompasses previously studied and new models for continuous, discrete, and combined data types.
- We show that maximum likelihood is based on a convex and smooth optimization function and provides consistent estimation and inference for the model parameters in the framework.

Let us have a glance at the framework. For this, we start with the Gaussian densities (1). Our first observation is that \(-\mathbf{x}^\top \Sigma^{-1} \mathbf{x}/2 = -\langle \Sigma^{-1}, \mathbf{x} \mathbf{x}^\top / 2 \rangle_{tr}\), where \(\langle \cdot, \cdot \rangle_{tr}\) is the trace inner product. This formulation looks “somewhat less revealing” [10, Page 125] on first sight, but it has two conceptual advantages. First, we argue that writing the parameters and the data as algebraic duals of each other makes their relationship more symmetric. Second, we argue that it is a good starting point for generalizations. For this, we take the viewpoint that the exponents in the densities are linear functions of the matrix \(\mathbf{x} \mathbf{x}^\top / 2\), and then replace this matrix by a general matrix-valued function \(T\) of \(\mathbf{x}\). Our second observation is that the fundamental quantity in the family of Gaussian graphical models is the inverse covariance matrix \(\Sigma^{-1}\) rather than the covariance matrix \(\Sigma\) itself. This suggests a reparametrization of the model using the matrix \(\Sigma^{-1}\), which is then replaced by a general matrix \(M\). This subtlety is important: as we will see later, the matrix \(M\) contains all information about the dependence structure of \(X\), while the equality of \(M^{-1}\) and the covariance matrix is a mere coincidence in the Gaussian case. With these two observations in mind, and denoting the log-normalization by \(\gamma(M)\), with \(\gamma(M) = \log((2\pi)^p|M^{-1}|)/2\) in the Gaussian case, we can then generalize the densities (1) to

\[ f_M(\mathbf{x}) = e^{-\langle M, T(\mathbf{x}) \rangle_{tr} - \gamma(M)} \]

with respect to an arbitrary \(\sigma\)-finite measure \(\nu\), and with \(M, T \in \mathbb{R}^{q \times q}\) a matrix-valued parameter and data function, respectively. While additional data terms can be absorbed in the measure \(\nu\), it is sometimes illustrative to write them explicitly. We thus consider distributions with densities of the form

\[ f_M(\mathbf{x}) = e^{-\langle M, T(\mathbf{x}) \rangle_{tr} + \xi(\mathbf{x}) - \gamma(M)} , \]
where $\xi(x)$ depends on $x$ only. These densities form an exponential family indexed by $M$ and are called exponential trace models in the following for convenience.

We recall that the well-known pairwise interaction models can also be written in exponential form, but there are important differences to the above formulation. First, pairwise interaction models and exponential trace models are not sub-classes of one another: exponential trace models are not limited to pairwise interactions, while pairwise interaction models are not limited to canonical parameterizations. Yet, importantly, exponential trace models generalize pairwise interaction models in the sense that all generic examples of pairwise interaction models are encompassed. We also highlight that in contrast to pairwise interaction models, we allow for $q \neq p$, which helps for concise formulations of mixed graphical models, for example. In general, we argue that the exponential trace framework is a practical starting point for general studies of graphical models, because it comprises a very large variety of examples and still ensures a firm theoretical structure.

We carefully specify and study the described distributions in the following sections. Section 2 contains the proposed framework: we define the densities in Section 2.1, and we discuss a variety of examples in Sections 2.2 and 2.3. Section 3 is focused on estimation and inference: we discuss the maximum likelihood estimator for the model parameters in Section 3.1, and we show its asymptotic normality in Section 3.2. We conclude the paper with a discussion in Section 4. The proofs are deferred to the Appendix.

Notation. For matrices $A, B \in \mathbb{R}^{s \times t}$, $s, t \in \{1, 2, \ldots \}$, we denote the trace inner product (or Frobenius inner product) by

$$\langle A, B \rangle_{tr} := \text{tr}(A^\top B) = \sum_{i=1}^{s} \sum_{j=1}^{t} A_{ij} B_{ij}$$

and the corresponding norm by

$$|A|_{tr} := \sqrt{\langle A, A \rangle_{tr}} = \sqrt{\sum_{i=1}^{s} \sum_{j=1}^{t} A_{ij}^2}.$$

We consider random vectors $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$, denoting random vectors and their realizations by upper case letters such as $X$ and arguments of functions by lower case, boldface letters such as $x$. Given a set $S \subset \{1, \ldots, p\}$, we denote by $X_S \in \mathbb{R}^{|S|}$ the vector that consists of the coordinates of $X$ with index in $S$, and we set $X_{-S} := X_{\bar{S}} \in \mathbb{R}^{p-|S|}$. Independence of two elements $X_i$ and $X_j$, $i \neq j$, is denoted by $X_i \perp X_j$; conditional independence of $X_i$ and $X_j$ given all other elements is denoted by $X_i \perp X_j | X_{-\{i,j\}}$. 
2. Framework. We first discuss our framework. In Section 2.1, we formulate the densities. In Section 2.2, we show that these densities apply to standard examples of graphical models. In Section 2.3, we study additional examples.

2.1. Exponential Trace Models. In this section, we formulate probabilistic models for vector-valued observations that have dependent coordinates. Specifically, we consider arbitrary (non-empty) finite or continuous domains \( D \subset \mathbb{R}^p \) and random vectors \( X \in D \) that have densities of the form

\[
f_M(x) := e^{-\langle M, T(x) \rangle + \xi(x) - \gamma(M)}
\]

with respect to some \( \sigma \)-finite measure \( \nu \) on \( D \). For reference, we call these models exponential trace models.

Let us specify the different components. The densities are indexed by \( M \in \mathcal{M} \), where \( \mathcal{M} \) is a subset of \( \mathcal{M}^* := \text{interior}\{M \in \mathbb{R}^{q \times q} : \gamma(M) < \infty\} \).

Unlike conventional frameworks, we do not require \( q = p \), with the advantage of concise formulations of mixture models, for example. In generic applications, \( M \) comprises the dependence structure of \( X \) and determines if two coordinates \( X_i \) and \( X_j \) are positively or negatively correlated. We will discuss these aspects in the next sections. The arguably most important note here is that the integrability condition \( \gamma(M) < \infty \) is feasible. In particular, our framework provides natural formulations of models that avoid unreasonable restrictions on the parameter space. It is best to see this in specific examples, so that we refer to later.

Next, the data enters the model via a matrix-valued function

\[
T : D \rightarrow \mathbb{R}^{q \times q} \\
x \mapsto T(x)
\]

and a real-valued function

\[
\xi : D \rightarrow \mathbb{R} \\
x \mapsto \xi(x),
\]

and \( \gamma(M) \) is the normalization defined as

\[
\gamma(M) := \log \int_D e^{-\langle M, T(x) \rangle + \xi(x)} d\nu.
\]
We finally have to impose two technical assumptions on the parameter space. Our first assumption is that the function $M \mapsto f_M$ of $M$ to the densities with respect to the measure $\nu$ is bijective. Sufficient conditions for this are provided in [1, Page 199] and [20, Definition 1.3]; we stress, however, that the bijection is required here only on $\mathcal{M}$ rather than on the full set $\mathcal{M}^*$. Our second assumption is that $\mathcal{M}$ is convex and that $\mathcal{M}$ is open with respect to an affine subspace of $\mathbb{R}^{q \times q}$. The two assumptions ensure especially that the parameter $M$ is identifiable and has a compact and “full-dimensional” neighborhood in an affine subspace of $\mathbb{R}^{q \times q}$. Importantly, however, the assumptions are mild enough to allow for overparametrizations in the sense that $\mathcal{M}$ does not have to be open in $\mathbb{R}^{q \times q}$; a typical example is $\mathcal{M}$ equal to the set of symmetric matrices, see the examples section.

The exponential family formulation equips the framework almost automatically with the necessary structure. For example, we can derive the following.

**Lemma 2.1.** The following two properties are satisfied.

1. The set $\mathcal{M}^*$ is convex;
2. for any $M \in \mathcal{M}^*$, the coordinates of $T(X)$ have moments of all orders with respect to $f_M$.

Property 1 ensures that $\mathcal{M} = \mathcal{M}^*$ works and Property 2 ensures concentration of the maximum likelihood estimator discussed later.

### 2.2. Standard Graphical Models

The goal of this section is to demonstrate that standard graphical models, such as Ising models with binary and $m$-ary responses and Gaussian and non-paranormal graphical models, fit the exponential trace framework. In combination with the results in Section 3, this shows in particular that standard graphical models are automatically equipped with more structure than suggested by common pairwise interaction formulations.

For the standard models, it is sufficient to consider $q = p$, $T_{ij}(x) \equiv T_{ij}(x_i, x_j)$, and $\xi(x) = \sum_{j=1}^p \xi_j(x_j)$. Given $M$, we then define a graph $G := (V, E)$, where $V := \{1, \ldots, p\}$ is the vertex (or node) set and $E := \{(i, j) : i, j \in V, i \neq j, M_{ij} \neq 0\}$ the edge set. Two matrices $M, M' \in \mathcal{M} \subset \mathbb{R}^{p \times p}$ correspond to the same graph if and only if their non-zero patterns are the same. In view of the examples below, we are particularly interested in symmetric dependence structures, that is, we consider symmetric matrices $M$ in what follows. Then, also the edge set $E$ is symmetric, that is, $(i, j) \in E$ if and only if $(j, i) \in E$, and the graph $G$ is called undirected.
The corresponding models $f_M$ are a special case of pairwise interaction models (pairwise Markov networks). Assuming that $\nu$ is a product measure, a density $h$ with respect to $\nu$ is a pairwise interaction model if it can be written in the form
\[
h(x) = \prod_{i,j=1}^{p} h_{ij}(x_i, x_j)
\]
with positive functions $h_{ij}$. This means that the densities of pairwise interaction models can be written as products of terms that depend on maximally two coordinates.

The graph $G$ now encodes the conditional dependence structure, as one can show by applying the Hammersley-Clifford theorem [2, 15]. The theorem implies that for a strictly positive density $h(x)$ with respect to a product measure, two elements $X_i, X_j$ are conditionally independent given all other coordinates if and only if we can write $h(x) = h^1(x_{-i})h^2(x_{-j})$, where $h^1, h^2$ are positive functions. For the described densities in our framework, we can write $f_M(x) = f^1_M(x_{-i})f^2_M(x_{-j})$ with positive functions $f^1_M, f^2_M$ if and only if $M_{ij} = 0$. By the above definition of the graph associated with $M$, the latter is equivalent to $(i, j) \notin E$. We thus find
\[
X_i \perp\!\!\!\!\perp X_j | X_{-\{i,j\}} \quad \text{if and only if} \quad (i, j) \notin E,
\]
meaning that the conditional dependence structure of $X$ is represented by the edge set $E$.

The graph $G$ also determines the unconditional dependence structure. To illustrate this, we define $\overline{E}$ as the set of all connected components in $M$, that is, $(i, j) \in \overline{E}$ if and only if there is a path $(i, i_1), (i_1, i_2), \ldots, (i_q, j) \in E$ that connects $i$ and $j$ (in particular, $E \subset \overline{E}$). For pairwise densities in our framework, it is easy to check that
\[
X_i \perp X_j \quad \text{if and only if} \quad (i, j) \notin \overline{E}.
\]
Hence, the dependence structure of $X$ is captured by $\overline{E}$. We give an illustration of these properties in Figure 1.

We can now turn to the examples.

2.2.1. Counting Measure. We first describe two cases where the base measure $\nu$ is the counting measure on $\{0, 1, \ldots \}^p$. Specifically, we show that the well-known Ising and multinomial Ising models are encompassed by our framework.
Beyond the measure, a unifying property of the two examples is that the integrability condition is satisfied for all matrices. In a formula, this reads
\[ \gamma(M) < \infty \quad \text{for all} \quad M \in \mathbb{R}^{p \times p}. \]
Since the set of all matrices in \( \mathbb{R}^{p \times p} \) is open, the integrability property (3) means that
\[ \mathcal{M}^* = \{ M \in \mathbb{R}^{p \times p} \}. \]
This ensures that any convex and (relatively) open set \( \mathcal{M} \subset \mathbb{R}^{p \times p} \) meets our technical assumptions. Importantly, we show later that the same properties are shared by exponential trace models for Poisson data.

**Ising.** The Ising model has a variety of applications, for example, in Statistical Mechanics and Quantum Field Theory [13, 42]. Its densities are proportional to
\[ e^{\sum_{j=1}^{p} a_{jj} x_j + \sum_{j,k} a_{jk} x_j x_k} \]
with \( a_{jk} = a_{kj} \), and the domain is \( x_1, \ldots, x_p \in \{0, 1\} \). As an illustration, consider a material that consists of \( p \) molecules with one “magnetic” electron each. The binary variable \( x_j \) then corresponds to the electron’s spin (up or down in a given direction) in the \( j \)th molecule, and the factor \( a_{jk} \) determines whether the spins of the electrons in the \( j \)th and \( k \)th molecule tend to align (\( a_{jk} > 0 \) as in ferromagnetic materials) or tend to take opposite directions (\( a_{jk} < 0 \) as in anti-ferromagnetic materials). The Ising model is a special case of our framework (2). Indeed, since \( 1^2 = 1 \) and \( 0^2 = 0 \), we can obtain the desired densities by setting \( D = \{0, 1\}^p \), \( q = p \), \( M_{ij} = -a_{ij} \), \( T_{ij}(x) = x_i x_j \), and \( \xi \equiv 0 \). Since the domain is finite, the integrability condition \( \gamma(M) < \infty \) is naturally satisfied for all matrices \( M \).
Multinomial Ising. The spin of spin 1/2 particles (such as the electron) can take two values. Thus, the corresponding measurements can be represented by the domain \{0, 1\} as in the Ising model discussed above. In contrast, the spin of spin \(s\) particles with \(s \in \{1, 3/2, \ldots\}\) (such as W and Z vector bosons and composite particles) can take \(2s + 1\) values, which cannot be represented by binaries directly. Therefore, the multinomial Ising model, which extends the Ising model to multinomial domains, is of considerable interest in quantum physics. For similar reasons, multinomial Ising models are of interest in other fields, see [6] for an example in sociology.

We now show that also the multinomial Ising model is a special case of our framework. For this, we encode the data in an enlarged binary vector. We first denote the original multinomial data by \(Y \in \{0, \ldots, m-1\}\). This could correspond to \(l\) spin \((m-1)/2\) particles. We then represent the data with an enlarged binary vector \(X \in \{0, 1\}^p\), \(p = l \cdot (m-1)\), by setting

\[
X_i = 1 \left\{ Y_j = i - (m-1) \left\lfloor \frac{i-1}{m-1} \right\rfloor, j = \left\lceil \frac{i}{m-1} \right\rceil \in \{0, 1\} \right. \}
\]

Each coordinate of the original data (for example, each spin) is now represented by \(m-1\) binary variables. We can use the same model as above, except for imposing the additional requirement \(M_{ij} = 0\) if \(\left\lceil \frac{i}{m-1} \right\rceil = \left\lceil \frac{j}{m-1} \right\rceil\) to avoid self-interactions. These settings yield the standard extension of the Ising model to multinomial data, cf. [29]. In particular, for \(m = 2\), we recover the Ising model above. Again, since the domain is finite, the integrability condition \(\gamma(M) < \infty\) is naturally satisfied for all matrices \(M\).

2.2.2. Continuous Measure. We now consider two examples in which the base measure \(\nu\) is the Lebesgue measure on \(\mathbb{R}^p\). Specifically, we show that Gaussian and non-paranormal graphical models are encompassed by our framework.

We show that the integrability condition is satisfied for all matrices that are positive definite. In a formula, this reads

\[
(4) \quad \gamma(M) < \infty \quad \text{for all positive definite } M \in \mathbb{R}^{p \times p}.
\]

Since the set of all positive definite matrices in \(\mathbb{R}^{p \times p}\) is open, this means that

\[
\mathcal{M}^* \supset \{M \in \mathbb{R}^{p \times p} : M \text{ positive definite}\}.
\]

One can thus take any convex and (relatively) open set of positive definite matrices as parameter space \(\mathcal{M}\). Later, we will show that very similar models can also be formulated for exponential data, for example.
The most popular examples are Gaussian graphical models [23]. For centered data (which can be assumed without loss of generality), these models correspond to random vectors $X \sim \mathcal{N}_p(0, \Sigma)$, where $\Sigma$ (and thus also $\Sigma^{-1}$) is a symmetric, positive definite matrix. We can generate these models in our framework (2) by setting $D = \mathbb{R}^p$, $q = p$, $M = \Sigma^{-1}$, $T_{ij}(x) = x_ix_j/2$, and $\xi \equiv 0$.

Gaussian graphical models are well-studied and are also the starting point for our contribution. It is thus especially important to disentangle general properties of graphical models in our framework from peculiarities of the Gaussian case. First, the correspondence of the conditional independence structure and the non-zero pattern of the inverse covariance matrix is specific to the Gaussian case. This has already been pointed out in [29], where relationships between the dependence structure and the inverse covariance matrix in Ising models and other exponential families with additional interaction terms are studied. However, instead of concentrating on possible connections, we argue that it is important to distinguish clearly between the two concepts. In our framework, the (conditional) dependencies are completely captured by the matrix $M$. It is thus reasonable to consider $M$ the fundamental quantity. Instead, the equality of $M^{-1}$ and $E_M[XX^\top]$ is specific to the Gaussian case, and the (generalized) covariances $E_M[XX^\top] (2E_M T(X))$ and their inverse should be viewed only as a characteristic of the model. Second, the type of the node conditionals can change once dependencies are introduced. In the Gaussian case, the node conditionals are normal irrespective of $M$. In most other cases, however, the types of node conditionals cannot be the same in the dependent and independent case - unless additional assumptions are introduced. We will discuss this in the later examples below.

**Non-paranormal.** Well-known generalizations of Gaussian graphical models are non-paranormal graphical models [28]. These models correspond to vectors $X$ such that $(g_1(X_1), \ldots, g_p(X_p))^\top \sim \mathcal{N}_p(0, \Sigma)$ for real-valued, monotone, and differentiable functions $g_1, \ldots, g_p$ and symmetric, positive definite matrix $\Sigma$. These models can be generated in our framework by setting $D = \mathbb{R}^p$, $q = p$, $M = \Sigma^{-1}$, $T_{ij}(x) = g_i(x_i)g_j(x_j)/2$, and $\xi(x) = \sum_{j=1}^p \log |g'_j(x_j)|$, cf. [28, Equation (2)]. The normalization constant is (irrespective of symmetry) $\gamma(M) = \log((2\pi)^p|M^{-1}|)/2 < \infty$ both in the Gaussian and the non-paranormal case, so that in both cases, property (4) is satisfied.

2.3. **Non-standard Examples.** The idea that Ising models as well as other standard graphical models can be written as exponential families is not new [35, Chapter 3.3]. However, we argue that the details of the notions and
formulations are essential, especially when it comes to establishing models for data that are not covered by standard graphical models. We will outline this in the following. We first discuss Poisson and exponential distributions. In particular, we establish thorough proofs for the integrability of the square-root models in [18] and introduce extensions that show that the square-root is just one out of many possible operations for the interaction terms and that a variety of distributions besides Poisson can be handled. We finally look beyond the horizon of pairwise interactions, discussing graphs with covariates and mixture models.

**Poisson.** A main objective in systems biology is the inference of microbial interactions. The corresponding data is multivariate count data with infinite range [12]. Other fields where such data is prevalent include particle physics (radioactive decay of particles) and criminalistics (number of crimes and arrests). However, copula-based approaches to infinite count data inflict severe identifiability issues [14], while standard extensions of the independent case lead to integrability issues, see below. Multivariate Poisson data has thus obtained considerable attention in the recent Machine Learning literature [18, 19, 40, 41], but much less in statistics.

We show in the following that within framework (2), one can solve the problems associated with the standard approaches while preserving the Poisson flavor of the individual coordinates, especially in the limit of small interactions. For this, we use our framework with the specifications \( D = \{0, 1, \ldots, p\} \), \( q = p \), \( \xi(x) = -\sum_{j=1}^{p} \log(x_j!) \), and functions \( T \) that satisfy \( T_{ii}(x) = T_{ii}(x_i) = x_i \) and \( T_{ij}(x) = T_{ij}(x_i, x_j) \leq c(x_i + x_j) \) for some \( c \in (0, \infty) \). We note that the case \( T_{ij}(x) = \sqrt{x_i x_j} \) has been introduced in the Machine Learning literature [18], but the technical aspects of this case have not been studied, and the general setting has not been formulated altogether.

Most importantly, we need to show property (3). To this end, we have to verify that for all matrices \( M \in \mathbb{R}^{p \times p} \)

\[
\sum_{x_1, \ldots, x_p = 0}^{\infty} \frac{e^{-\sum_{j} M_{jj} x_j - \sum_{i,j,i \neq j} M_{ij} T_{ij}(x_i, x_j)}}{\prod_j x_j!} < \infty.
\]

Since \(-M_{ij} T_{ij}(x_i, x_j) \leq \tilde{c} x_i + \tilde{c} x_j \), where \( \tilde{c} := c \max_{i,j} |M_{ij}| \), a sufficient condition is

\[
\sum_{x_1, \ldots, x_p = 0}^{\infty} \frac{e^{-\sum_{j} (M_{jj} - 2\tilde{c}) x_j}}{\prod_j x_j!} < \infty.
\]

Hence, defining \( C(M, j) := e^{-(M_{jj} - 2\tilde{c})} \in (0, \infty) \), \( j \in \{1, \ldots, p\} \), the integra-
bility condition is implied for any $M$ by the fact

$$
\sum_{x_1, \ldots, x_p=0}^{\infty} \prod_{j=1}^{p} \frac{C(M,j)^{x_j}}{x_j!} = \prod_{j=1}^{p} e^{C(M,j)} < \infty.
$$

This proves property (3).

In contrast, the corresponding integrability conditions in standard approaches to this data type inflict severe restrictions on the parameter space. To see this, recall that the joint density of $p$ independent Poisson random variables with parameters $a_1, \ldots, a_p > 0$ is proportional to

$$
\exp \left( \sum_{j=1}^{p} \log(a_j) x_j - \sum_{j=1}^{p} \log(x_j!) \right).
$$

The standard approach to include interactions is to add terms of the form $a_{ij} x_i x_j$. This yields densities proportional to

$$
\exp \left( \sum_{j=1}^{p} \log(a_j) x_j + \sum_{i \neq j} a_{ij} x_i x_j - \sum_{j=1}^{p} \log(x_j!) \right).
$$

The dominating terms in this expression are the interaction terms $a_{ij} x_i x_j$. Using Stirling’s approximation, we find that $x^2 / \log(x!) \to \infty$ for $x \to \infty$, showing that the density cannot be normalized unless $a_{ij} \leq 0$ for all $i, j$. This means that the standard approach excludes positive interactions between the nodes.

Let us finally look at the node conditionals for a specific $T$. We choose $T_{ij}(x) = \sqrt{x_i x_j}$ for simplicity. The node conditionals become

$$
f_M(x_j | x_{-j}) \sim e^{-M_{jj} x_j - \log(x_j!)} L_{\text{Int}},
$$

where

$$
L_{\text{Int}} = e^{-\sqrt{\pi} \sum_{k \in \mathcal{N}(j)} (M_{jk} + M_{kj}) / \sqrt{x_k}}.
$$

The off-diagonal terms in $M$ model the interactions of $j$ with the other nodes. If the factors $M_{jk}$ are small, $L_{\text{Int}} \approx 1$, and thus, the node $j$ approximately follows a Poisson distribution with parameter $e^{-M_{jj}}$. In particular, if $M$ is diagonal, the nodes are independent Poisson distributed random variables.

In comparison, the standard approach represented by Display (5) results in exact Poisson node conditionals for any non-positive correlations. Conversely, it has been shown in [5, Proposition 1 and Lemma 1] that one can
find a distribution with Poisson (or exponential) node conditionals *only* if all interactions are non-positive. Thus, an unavoidable price for “pure” node conditionals is a strong, in practice typically unrealistic assumption on the parameter space. Our framework avoids this assumption and is still close to the exact Poisson (exponential) distributions if the interactions are small.

**Exponential.** The exponential case is the counterpart of the Poisson case discussed above. In particular, the standard approach to correlated exponential data is confronted with the same integrability issues as above, while approaches via framework (2) easily satisfy the integrability conditions.

To model exponential data, we consider $D = [0, \infty)^p$, $q = p$, and $\xi \equiv 0$. Again a number of transformations $T$ would have the desired properties; however, to avoid digression, we only consider the square-root transformations $T_{ij}(x) = \sqrt{x_i x_j}$ that correspond to [18]; generalization are possible along the same lines as in the Poisson case. We can now check the integrability condition (4). Denoting the smallest $\ell_2$-eigenvalue of $M$ by $\kappa(M) > 0$, we find

$$e^{\gamma(M)} = \int_0^\infty \cdots \int_0^\infty e^{-\sum_{i,j=1}^p M_{ij} \sqrt{x_i x_j}} \, dx_1 \cdots dx_p$$

$$\leq \int_0^\infty \cdots \int_0^\infty e^{-\kappa(M)x_1^2} \, dx_1 \cdots dx_p$$

$$= \left( \int_0^\infty e^{-\kappa(M)x} \, dx \right)^p$$

$$= \kappa(M)^{-p} < \infty.$$

Hence, $\gamma(M) < \infty$ for all positive definite matrices $M$.

In contrast, adding linear interaction terms to the independent joint density forbids positive correlations. One can check this similarly as in the Poisson case above.

The node conditionals finally become

$$f_M(x_j|x_{-j}) \sim e^{-M_{jj} x_j} L_{\text{Int}},$$

where

$$L_{\text{Int}} = e^{-\sqrt{x_j} \sum_{k \in N(j)} (M_{jk} + M_{kj}) \sqrt{x_k}}.$$

The off-diagonal terms in $M$ model the interactions of $j$ with the other nodes. If the factors $M_{jk}$ are small, $L_{\text{Int}} \approx 1$, and thus, node $j$ approximately follows an exponential distribution with parameter $M_{jj}$. In particular, if $M$ is diagonal, all node conditionals follow independent exponential distributions.
Composite Models. As an example for composite models, let us consider data with Poisson and exponential elements. Note first that the conditions on the set of matrices $\mathcal{M}$ are different in the discrete examples and the continuous examples: In the discrete examples, we have shown $\gamma(M) < \infty$ for any matrix $M$. In the continuous examples, we have shown $\gamma(M) < \infty$ under the additional assumption that $M$ is positive definite. In the case of composite models, one can interpolate the conditions. However, for the sake of simplicity, we instead assume that the matrices $M$ are positive definite. We then consider $p_1, p_2 \in \{1, 2, \ldots\}$, $p_1 + p_2 = p$, $\mathcal{D} = \{0, 1, \ldots\}^{p_1} \times [0, \infty)^{p_2}$, $\mathcal{M} \subset \{M \in \mathbb{R}^{p \times p} : M \text{ symmetric, positive definite}\}$, $\mathcal{M}$ open and convex, and $\xi(x) = -\sum_{j=1}^{p_1} \log(x_j!)$. Hence, the first $p_1$ elements of the random vector $X$ are discrete, while the other $p_2$ elements are continuous. Using again the square-root transformation, the Poisson-type node conditionals for $j \in \{1, \ldots, p_1\}$ are

$$f_M(x_j|x_{-j}) \sim e^{-M_{jj}x_j - \log(x_j!)} L_{\text{Int}},$$

where

$$L_{\text{Int}} = e^{-\sqrt{\mathcal{D}} \sum_{k \in \mathcal{N}(j)(M_{jk} + M_{kj})} \sqrt{\mathcal{N}}}.$$

The exponential-type node conditionals for $j \in \{p_1 + 1, \ldots, p_1 + p_2\}$ have the corresponding form. The expressions highlight that the densities can include interactions between the discrete and continuous elements of $X$, while the Poisson/exponential-flavors of the nodes are still preserved.

To show that $\gamma(M) < \infty$, we proceed similarly as in the examples above. More precisely, denoting the smallest $\ell_2$-eigenvalue of $M$ by $\kappa(M) > 0$, we find

$$e^{\gamma(M)} = \sum_{x_1, \ldots, x_{p_1} = 0}^{\infty} \int_0^\infty \cdots \int_0^\infty e^{-\sum_{i,j=1}^{p_1} M_{ij} \sqrt{x_i} x_j} \prod_{j=1}^{p_1} x_j! \, dx_{p_1+1} \cdots dx_p \leq \sum_{x_1, \ldots, x_{p_1} = 0}^{\infty} \int_0^\infty \cdots \int_0^\infty e^{-\kappa(M)|x|_1} \prod_{j=1}^{p_1} x_j! \, dx_{p_1+1} \cdots dx_p \leq \sum_{x_1, \ldots, x_{p_1} = 0}^{\infty} e^{-\kappa(M)(x_1 + \cdots + x_{p_1})} \prod_{j=1}^{p_1} x_j! \times \int_0^\infty \cdots \int_0^\infty e^{-\kappa(M)(x_{p_1+1} + \cdots + x_{p_2})} \, dx_{p_1+1} \cdots dx_p = (e^{e^{-\kappa(M)}})^{p_1} \left( \int_0^\infty e^{-\kappa(M)x} \, dx \right)^{p_2} < \infty.$$
Laplace and Beyond. There is much room for our creativity in constructing models. For example, we can readily establish models for Laplace (double-exponential) data by inserting absolute values throughout, for example, \( T_{ij}(x) = \sqrt{|x_i x_j|} \). Indeed, again denoting the smallest \( \ell_2 \)-eigenvalue of \( M \) by \( \kappa(M) > 0 \), we find

\[
e^{-\gamma(M)} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\sum_{i,j=1}^{p} M_{ij} \sqrt{|x_i x_j|}} \ dx_1 \cdots dx_p \\
\leq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\kappa(M) |x|^1} \ dx_1 \cdots dx_p \\
= \left( 2 \int_{0}^{\infty} e^{-\kappa(M) x} \ dx \right)^p \\
= \left( \frac{2}{\kappa(M)} \right)^p < \infty.
\]

Hence, \( \gamma(M) < \infty \) for all positive definite matrices \( M \). In general, the essentially only limit is that to ensure integrability, the interaction terms have to be of “smaller order” than the terms that correspond to the independent case. In the following, we show in particular that interactions do not need to be pairwise.

2.3.1. Restricted Pairwise Interaction Models. We are sometimes interested in only a part of the dependence structure, so that applying restrictions to all interactions might be unnecessarily stringent. As a toy example, let us study interactions of one variable \( X_p \) with a number of covariates \( X_1, \ldots, X_{p-1} \). Assuming that the joint distribution of the covariates \( X_1, \ldots, X_{p-1} \) and the distribution of \( X_p \) are known, simple yet informative models could be of the form

\[
f_M(x) = e^{-\langle M, T(x) \rangle + \xi(x_1, \ldots, x_{p-1}) + \dot{\xi}(x_p) - \gamma(M)},
\]

where - just to fix ideas -

\[
M = \begin{bmatrix}
0_{p \times (p-1)} \\
M_{1p} \\
M_{pp}
\end{bmatrix}
\quad \text{and} \quad
T(x) = \begin{bmatrix}
x_1 x_p \\
0_{p \times (p-1)} \\
x_p x_p
\end{bmatrix}.
\]

The quantities \( \xi, \dot{\xi} \) are known but arbitrary functions of \( x_1, \ldots, x_{p-1} \) and \( x_p \), respectively; in particular, we do not assume that \( f_M \) is a pairwise interaction model. Still, the model is in a sense “locally pairwise,” and the matrix \( M \) captures concisely the conditional dependence structures of interest. Indeed, one can check that

\[
X_i \indep X_p | X_{-\{i,p\}} \quad \text{if and only if} \quad M_{ip} \neq 0.
\]
More generally, consider a graph $G = (V, E)$ with edge set $E \subset A \subset \{(i, j) : i, j \in V, i \neq j\}$. We say that $f_M$ is $A$-restricted pairwise if for all $(i, j) \in A$,

$$X_i \perp \! \! \! \perp X_j | X_{-\{i,j\}} \quad \text{if and only if} \quad (i, j) \notin E.$$  

Then, $G = (V, E)$ with $V = \{1, \ldots, p\}$ and $E = \{(i, j) \in A : M_{ij} \neq 0\}$ provides a concise description dependence structure under consideration. An illustration with the above toy example is provided in Figure 2.

**Mixture Models.** We now consider network structures that can depend on observed qualitative covariates. As a toy example, we look at Gaussian graphical models for $Y \in \mathbb{R}^{p-1}$ (brain regions or voxels, ...) with an additional binary covariate (sex, disease group, ...) $Z \in \{0, 1\}$.

In line with the earlier description of Gaussian graphical models, a corresponding model for $X := (Y^\top, Z)^\top \in \mathbb{R}^p$ is

$$h_{M^1, M^2}(x) := (I\{z = 0\}e^{-(M^1, yy^\top/2)^\top} + I\{z = 1\}e^{-(M^2, yy^\top/2)^\top})e^{-\gamma(M^1, M^2)},$$

where $M^1, M^2 \in \mathbb{R}^{(p-1)\times (p-1)}$ are symmetric matrices and $\gamma(M^1, M^2)$ is the normalization. One can readily write the density $h_{M^1, M^2}$ in the form (2) as

$$f_{M}(x) = e^{-\langle M, T(x) \rangle_{tr}} - \gamma(M),$$

where $M := \text{diag}(M^1, M^2)$ and $T(x) := \text{diag}(I\{z = 0\}yy^\top, I\{z = 1\}yy^\top)/2$. The parameter set $\mathcal{M}$ consists then of the symmetric, block-diagonal matrices in $\mathbb{R}^{2(p-1)\times 2(p-1)}$. Note that $\mathcal{M}$ is not open $\mathbb{R}^{2(p-1)\times 2(p-1)}$ but only in an appropriate subspace, which highlights once again that the conditions on the parameter set need to be sufficiently week. Note also that the above model is again not a pairwise interaction model.
A natural question is now why not simply estimating the graphs corresponding to the different outcomes of \( Z \) independently from each other. A main argument against this is that sample sizes are often small in practice. Small sample sizes mean that separate parameter estimations for the models might not be successful \([4]\), and instead, exploiting commonalities between the networks might be essential. One approach to exploit commonalities could be to add fusion-type penalties \([32]\) into the estimation, assuming that the networks are similar overall. Another approach is to use quasi maximum likelihood techniques. For example, we could model the connectivities among nodes \( 1, \ldots, \hat{p} \) in dependence of \( Z \) and assume that the other parts of the network are unaffected by \( Z \). To formulate this, let \( \hat{Y} \in \mathbb{R}^{\hat{p}} \) be the samples corresponding to nodes \( 1, \ldots, \hat{p} \) and \( \ddot{Y} \in \mathbb{R}^{p - 1 - \hat{p}} \) the remaining samples. A Gaussian-type model for \( X = (\hat{Y}^\top, \ddot{Y}^\top, Z)^\top \in \mathbb{R}^p \) is then

\[
h_{M_1, M_2, M_{12}, M_3}(x) := (1 \{z = 0\} e^{-(M_1, \hat{y}\hat{y}^\top / 2)^\text{tr}} + 1 \{z = 1\} e^{-(M_2, \hat{y}\hat{y}^\top / 2)^\text{tr}}) \times e^{-(M_{12}, \hat{y}\ddot{y}^\top / 2)^\text{tr}} - (M_1, M_2, M_{12}, M_3) - (1 \{z = 0\} e^{-(M_1, \hat{y}\hat{y}^\top / 2)^\text{tr}}) - (1 \{z = 1\} e^{-(M_2, \hat{y}\hat{y}^\top / 2)^\text{tr}} - (M_1, M_2, M_{12}, M_3)).
\]

We can again write this density in the form (2) as

\[
f_M(x) = e^{-\langle M, T(x) \rangle_{\text{tr}} - \gamma(M)},
\]

where \( M := \text{diag}(M_1, M_2, M_{12}, M_3) \) and

\[
T(x) := \begin{bmatrix}
1 \{z = 0\} \hat{y} \hat{y}^\top / 2 \\
1 \{z = 1\} \hat{y} \hat{y}^\top / 2 \\
\hat{y} \ddot{y}^\top / 2 \\
\ddot{y} \ddot{y}^\top / 2
\end{bmatrix}.
\]

We can now describe the dependence structures in terms of graphs. Since the above densities are sums, a direct application of the Hammersley-Clifford machinery would not be very enlightening. Instead, we describe the dependence structures in function of \( Z \). To be a bit more general, let \( X = (Y^\top, Z)^\top \in \mathbb{R}^{p - 1} \times \{1, \ldots, l\} \) have density \( h \), and consider graphs \( G^1 = (V, E^1), \ldots, G^l = (V, E^l) \) over the vertex set \( V = \{1, \ldots, p - 1\} \). Assume that for all \( k \in \{1, \ldots, l\} \)

\[
h(y|z = k) = \prod_{m \in V} h_m(y_m, k) \prod_{(i, j) \in E^k} h_{ij}(y_i, y_j, k)
\]

with positive functions \( h_m, h_{ij} \). This assumption relates to the notion of conditional random fields that has been introduced for segmenting and labeling sequence data \([22]\), see also \([25]\) for further considerations. We can
now apply the Hammersley-Clifford theorem for fixed $Z$. This yields that for $Z = k$, the edge set $E^k$ captures the conditional dependence structure of $Y$. In the first toy model at the very top, one can check readily that $f_M$ factorizes as above with respect to the graphs $G^1 = (V, E^1), G^2 = (V, E^2)$, where $V = \{1, \ldots, p - 1\}, E^1 = \{(i, j) : i, j \in V, i \neq j, M^1_{ij} \neq 0\}$, and $E^2 = \{(i, j) : i, j \in V, i \neq j, M^2_{ij} \neq 0\}$. In summary, although the mixture models are not pairwise interaction models, they have interesting and concisely described dependence structures, and they are equipped with all the guarantees that are stated in the following section.

3. Estimation and Inference. We now turn to estimation and inference. In Section 3.1, we show that maximum likelihood estimation has desirable properties in our framework. In Section 3.2, we then show that the maximum likelihood estimator is asymptotically normal, and we discuss the construction of tests and confidence intervals.

3.1. Maximum Likelihood Estimation. We study maximum likelihood estimation in our framework. For this, we assume given $n$ i.i.d. data samples $X^1, \ldots, X^n$ from a distribution of the form (2). Also, we assume given the model specifications $D, \nu, T, \xi$ and the parameter space $\mathcal{M}$, that is, we assume that the model class is known. In contrast, the correct model parameters specified in the matrix $M$ are unknown. Our goal is to estimate $M$ from the data.

As a toy example, consider data about $p$ different populations of freshwater fish in $n$ similar lakes. More specifically, consider vector-valued observations $X^1, \ldots, X^n \in \{0, 1, \ldots\}^p$, where $(X^i)_j$ is the number of fish of type $j$ in lake $i$. We want to use these data to uncover the relationships among the different populations. A model suited for this task is the Poisson model discussed earlier. For example, we might set $D = \{0, 1, \ldots\}^p$, $q = p$, $\{M \in \mathbb{R}^{p \times p} : M \text{ symmetric}\}$, $T_{ij}(x) = \sqrt{x_i x_j}$, and $\xi(x) = -\sum_{j=1}^p \log(x_i!)$. The relationships among the fish populations are then encoded in $M$, which then needs to be estimated from the observations.

Before heading on, we add some convenient notation. We summarize the data in $X := (X^1, \ldots, X^n)$ and denote the corresponding function argument by $x := (x^1, \ldots, x^n)$ for $x^1, \ldots, x^n \in D$. The generalized Gram matrix is denoted by

$$T(x) := \frac{1}{n} \sum_{i=1}^n T(x^i).$$

The negative joint log-likelihood function $-\ell_M$ for $n$ i.i.d random vectors
corresponding to the model (2) is finally given by

$$-\ell_M(\mathbf{x}) = n\langle M, \mathbf{T}(\mathbf{x}) \rangle_{tr} - \sum_{i=1}^{n} \xi(x_i) + n\gamma(M).$$

We can now state the maximum likelihood estimator and its properties. For further reference, we first state the essence of the previous display in the following lemma.

**Lemma 3.1 (Log-likelihood).** Given any \( M \in \mathcal{M}^* \), the negative joint log-likelihood function \(-\ell_M\) of \( n \) i.i.d. random vectors distributed according to \( f_M \) in (2) can be expressed by

$$-\ell_M(\mathbf{x}) = n\langle M, \mathbf{T}(\mathbf{x}) \rangle_{tr} + n\gamma(M) + c,$$

where \( c \in \mathbb{R} \) does not depend on \( M \).

Motivated by Lemma 3.1, we introduce the maximum likelihood estimator of \( M \) by

$$\hat{M} := \arg\min_{\tilde{M} \in \mathcal{M}} \{ -\ell_{\tilde{M}}(X) \} = \arg\min_{\tilde{M} \in \mathcal{M}} \{ \langle \tilde{M}, \mathbf{T}(\mathbf{x}) \rangle_{tr} + \gamma(\tilde{M}) \}.$$

The estimator exists in all generic examples. More generally, under our assumption that \( M \in \mathcal{M} \subset \mathcal{M}^* \) and \( \mathcal{M} \) is open and convex, it exists for \( n \) sufficiently large, cf. [1]. The objective function of the estimator has two important properties. First, it is convex:

**Lemma 3.2 (Convexity).** For any \( \mathbf{x} \in \mathcal{D}^p \), the function

$$\mathcal{M}^* \to \mathbb{R}
M \mapsto \langle M, \mathbf{T}(\mathbf{x}) \rangle_{tr} + \gamma(M)$$

is convex.

Second, its derivatives can be computed explicitly.

**Lemma 3.3 (Derivatives).** For any \( \mathbf{x} \in \mathcal{D}^p \), the function

$$\mathcal{M}^* \to \mathbb{R}
M \mapsto \langle M, \mathbf{T}(\mathbf{x}) \rangle_{tr} + \gamma(M)$$

is differentiable.
is twice differentiable with partial derivatives

\[ \frac{\partial}{\partial M_{ij}} \left( \langle M, T(x) \rangle_{\text{tr}} + \gamma(M) \right) = T_{ij}(x) - \mathbb{E}_M T_{ij}(X) \]

and

\[ \frac{\partial}{\partial M_{ij}} \frac{\partial}{\partial M_{kl}} \left( \langle M, T(x) \rangle_{\text{tr}} + \gamma(M) \right) = n \mathbb{E}_M \left[ (T_{ij}(X) - \mathbb{E}_M T_{ij}(X))(T_{kl}(X) - \mathbb{E}_M T_{kl}(X)) \right] \]

for \( i, j, k, l \in \{1, \ldots, q\} \).

Convexity and the explicit derivatives are desirable for both optimization and theory. From an optimization perspective, the two properties are valuable, because they render the objective function amenable to gradient-type minimization. From a theoretical perspective, the two properties are valuable, because they imply that

\[ \hat{M} = \arg\min_{M \in \mathcal{M}} \left\{ \mathbb{E}_M \left[ \langle \tilde{M}, T(X) \rangle_{\text{tr}} + \gamma(\tilde{M}) \right] \right\}, \]

showing that \( \hat{M} \) is a standard M-estimator, and because they imply that \( \hat{M} \) can be written as a Z-estimator (note that \( \hat{M} \) is necessarily in the interior of \( \mathcal{M} \)) with criterion

\[ \overline{T}(X) = \mathbb{E}_\hat{M} \overline{T}(X). \]

A simple special case is Gaussian data. Recall that in this case, \( M \) is the inverse of the usual population covariance matrix. Moreover, one can check that

\[ T(x) = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \]

and \( \hat{M} = T(X)^{-1} \). Hence, in this case, the estimator \( \hat{M} \) is the inverse of the (usual) sample covariance matrix.

### 3.2. Asymptotic Normality

We now show that the maximum likelihood estimator of \( M \) is asymptotically normal with covariance equal to the inverse Fisher information. This allows for the construction of efficient asymptotic tests and confidence sets.

The asymptotic normality is established in the following result.

**Theorem 3.1.** The estimator \( \hat{M} \) in (6) is asymptotically normal with covariance equal to the inverse Fisher information.
Theorem 3.1 implies in particular that the maximum likelihood estimator of $\mathbf{M}$ is asymptotically efficient [30, Chapter 4.5]. Note that in our notation, the Fisher information is a tensor acting on pairs of matrices in $\mathbb{R}^{q \times q}$. Taking inverses is understood in terms of the vectorized parameter spaces that associate $\mathbb{R}^q \leftrightarrow \mathbb{R}^{q \times q}$. The following remark contains the explicit form of the Fisher information.

**Remark 3.1.** The Fisher information $\mathbf{I}_M$ is given by

$$(\mathbf{I}_M)_{ijkl} = n \mathbb{E}_M \left[ (\mathbf{T}_{ij}(\mathbf{X}) - \mathbb{E}_M \mathbf{T}_{ij}(\mathbf{X})) (\mathbf{T}_{kl}(\mathbf{X}) - \mathbb{E}_M \mathbf{T}_{kl}(\mathbf{X})) \right]$$

for $i, j, k, l \in \{1, \ldots, q\}$. This explicit expression can be used to show three important properties: (i) It can be used to show that $\mathbf{I}_M$ is invertible, cf. the proof of Lemma 3.2. (ii) It can be used to show that the density with respect to the Lebesgue measure of the normal distribution with mean zero and covariance equal to the Fisher information has a concise form. Indeed, we find that the density as a function of $\mathbf{A} \in \mathbb{R}^{q \times q}$ is proportional to

$$e^{-\mathbb{E}_M \langle \mathbf{A}, \mathbf{T}(\mathbf{X}) - \mathbb{E}_M \mathbf{T}(\mathbf{X}) \rangle^2 / 2}.$$ 

This formulation highlights once again the prominent role of the trace in our framework. (iii) Finally, it can be used to show that for pairwise interaction models, the non-zero pattern in $\mathbf{M}$ not only determines the dependencies among the coordinates of $\mathbf{X}$ but also the asymptotic dependencies among the coordinates of $\hat{\mathbf{M}}$. Most illustrative for this is the case where $\mathbf{M}$ is diagonal, that is, the coordinates of $\mathbf{X}$ are independent. Then, the above expression implies that $\mathbf{I}_M$ is diagonal, so that also the coordinates of $\hat{\mathbf{M}}$ are asymptotically independent.

A main feature of Theorem 3.1 is that it readily provides asymptotic tests and confidence sets. For example, one can apply Wald-type tests, Rao-type score tests, and likelihood ratio tests, see [30] for details about these types of tests. If the data are continuous, score matching is yet another alternative, see [7, 26] and references therein. Consider test settings with null and alternative hypotheses of the form

$$H_N : \lambda(\mathbf{M}) = 0 \quad \text{versus} \quad H_A : \lambda(\mathbf{M}) \neq 0$$

for a continuously differentiable $\lambda$ mapping $\mathbb{R}^{q \times q}$ into $\mathbb{R}^m$. A corresponding Wald-type test statistic is given by

$$\hat{\mathbf{W}} := \lambda(\hat{\mathbf{M}}) \mathbf{V}^{-1} \lambda(\hat{\mathbf{M}}),$$

where $\mathbf{V}$ is the estimated covariance matrix of $\lambda(\hat{\mathbf{M}})$. The asymptotic distribution of $\hat{\mathbf{W}}$ is a chi-square distribution with $m$ degrees of freedom under $H_N$.
where $\hat{V} \equiv \hat{V}(\hat{M})$ is defined via

$$
\hat{V}_{ij} := \sum_{k,l=1}^{q} \sum_{k',l'=1}^{q} \frac{\partial \lambda_i(M)}{\partial M_{kl}} \bigg|_{M=\hat{M}} (\hat{M}^{-1})_{k'l'} \frac{\partial \lambda_j(M)}{\partial M_{k'l'}} \bigg|_{M=\hat{M}}
$$

for $i, j \in \{1, \ldots, m\}$ and a consistent estimator $\hat{I}_M$ of $I_M$. A natural choice for $\hat{I}_M$ is the empirical version of $I_M$, that is,

$$
(\hat{I}_M)_{ijkl} = \frac{1}{n-1} \sum_{i=1}^{n} (T_{ij}(X^i) - T_{ij}(\bar{X})) (T_{kl}(X^i) - T_{kl}(\bar{X})).
$$

Since $\hat{W}$ converges in distribution under the null hypothesis $H_N$ to the chi-square distribution with $m$ degrees of freedom [30, Theorem 6.6], the quantiles of $\chi^2_m$ can be used to construct asymptotic tests (and similarly, asymptotic confidence intervals).

Let us come back to our toy example introduced earlier. Assume fish of type $i$ and $j$ are suspected to compete for the same food sources. In our framework, an indication for this would be $M_{ij} \neq 0$, that is, the sizes of the populations of fish $i$ and $j$ depend directly on each other. We are thus interested in testing

$$
H_N : M_{ij} = 0 \quad \text{versus} \quad H_A : M_{ij} \neq 0.
$$

For this, one can use the Wald-type test statistic defined in (7), which simplifies to

$$
\hat{W} = (\hat{I}_M)_{ijij} \hat{M}_{ij}^2.
$$

A p-value can now be obtained via the quantiles of $\chi^2_1$.

Another concept of interest is the class of confidence subgraphs. Consider a model with parameter $M \in \mathbb{R}^{p \times p}$ and corresponding graph $G = (V, E)$ that has vertex set $V = \{1, \ldots, p\}$ and edge set $E = \{(i, j) : i, j \in \{1, \ldots, p\}, i \neq j, M_{ij} \neq 0\}$ as discussed in the examples sections. Consider also a data-driven graph $\hat{G} = (V, \hat{E})$ that has the same vertex set $V = \{1, \ldots, p\}$ but a data-driven edge set $\hat{E} \equiv \hat{E}(\bar{X})$. The graph $\hat{G}$ is called a confidence subgraph at level $\alpha \geq 0$ if $P_M(\hat{E} \subset E) \geq 1 - \alpha$. Of course, $\hat{G}$ with $\hat{E} = \emptyset$ is always a confidence subgraph, but much more interesting are larger confidence subgraphs. The construction of confidence subgraphs for Gaussian graphical models is described in [8, 9], see also [36, 37]. In our more general framework, one can proceed along the same lines. For example, one can apply standard multiple testing adjustments, such as Bonferroni-Holm procedures [17], to the pairwise tests above.

We finally mention that model misspecifications can be treated along the classical lines in [38].
4. Discussion. The exponential family framework for graphical models allows for a wide range of examples as highlighted in Sections 2.2 and 2.3 and yet ensures a rigid theoretical structure as demonstrated in Section 3.

A direction for future research is total positivity. This concept has recently received considerable attention in the context of graphical models, see [11, 21, 31] and references therein. On a high level, total positivity requires that the coordinates of the random vector are positively correlated. It would be of interest to study the effects of this additional requirement on our framework.

Acknowledgements. We particularly thank Jon Wellner for the many inspiring discussions and insightful suggestions. We also thank Rui Zhuang, Marco Rossini, Noah Simon, Mathias Drton, Sam Koelle, Zack Almquist, and Lina Lin for their valuable input.

REFERENCES

[1] Robert H. Berk. Consistency and asymptotic normality of MLE’s for exponential models. *Ann. Mat. Statist.*, 43:193–204, 1972.

[2] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, 36:192–236, 1974. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

[3] Stephen G. Brush. History of the Lenz-Ising model. *Reviews of modern physics*, 39(4):883, 1967.

[4] Yunqi Bu and Johannes Lederer. Integrating additional knowledge into estimation of graphical models. *arXiv:1704.02739*, 2017.

[5] Shizhe Chen, Daniela M. Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015.

[6] Jana Diesner and Kathleen M. Carley. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security*, Newport Beach, CA. Citeseer, 2005.

[7] Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *arXiv:1606.02359*, 2016.

[8] Mathias Drton and Michael D. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.*, 22(3):430–449, 2007.

[9] Mathias Drton and Michael D. Perlman. A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference*, 138(4):1179–1200, 2008.

[10] Morris L. Eaton. Multivariate Statistics: A Vector Space Approach. *IMS Lecture Notes Monogr. Ser.*, 53, 2007.

[11] Shaun Fallat, Steffen Lauritzen, Kayvan Sadeghi, Caroline Uhler, Nanny Wermuth, and Piotr Zwiernik. Total positivity in markov structures. *arXiv:1510.01290*, 2015.

[12] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*, 8(7):e1002606, 2012.
[31] Martin Slawski and Matthias Hein. Estimation of positive definite $M$-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra Appl.*, 473:145–179, 2015.

[32] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[33] Aad W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[34] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[35] Martin J. Wainwright and Michael I. Jordan. *Graphical models, exponential families, and variational inference*, volume 1. Now Publishers Inc., 2008.

[36] Larry Wasserman, Mladen Kolar, and Alessandro Rinaldo. Estimating undirected graphs under weak assumptions. *arXiv:1309.6933*, 2013.

[37] Larry Wasserman, Mladen Kolar, and Alessandro Rinaldo. Berry-Esseen bounds for estimating undirected graphs. *Electron. J. Stat.*, 8(1):1188–1224, 2014.

[38] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[39] Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571, 2012.

[40] Eunho Yang, Pradeep Ravikumar, Genevera I. Allen, and Zhandong Liu. On graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, 16:3813–3847, 2015.

[41] Eunho Yang, Pradeep K. Ravikumar, Genevera I. Allen, and Zhandong Liu. On Poisson Graphical Models. In *NIPS*, pages 1718–1726, 2013.

[42] Jean-Bernard Zuber and Claude Itzykson. Quantum field theory and the two-dimensional Ising model. *Physical Review D*, 15(10):2875, 1977.
A.1. Proof of Lemma 2.1.

**Proof of Lemma 2.1.** We prove the two properties in order. The main proof ideas can also be found in [1, Pages 193-195].

**Property 1.** We first show that $\mathcal{M}^*$ is convex.
For this, consider $\alpha \in [0, 1]$ and $M, M' \in \mathcal{M}^*$. Then, by definition of the normalization and by convexity of the exponential function,

$$e^{\gamma(M + (1 - \alpha)M') = \int_D e^{-\gamma(M + (1 - \alpha)M')} d\nu} \leq \int_D (\alpha e^{-\gamma(M)} + (1 - \alpha) e^{-\gamma(M')}) d\nu = \alpha e^{\gamma(M)} + (1 - \alpha) e^{\gamma(M')} < \infty.$$ 

Hence, $\gamma(M + (1 - \alpha)M') < \infty$, and thus, $\alpha M + (1 - \alpha)M' \in \mathcal{M}^*$. This concludes the proof of the first property.

**Property 2.** We now show that for any $M \in \mathcal{M}^*$, the coordinates of $T(X)$ have moments of all orders with respect to $f_M$.
To this end, fix an $M \in \mathcal{M}^*$. Since $\mathcal{M}^*$ is open, there is a neighborhood $\mathcal{M}_M$ of $0_{q \times q}$ such that $\{M - A : A \in \mathcal{M}_M\} \subset \mathcal{M}^*$. For any $A \in \mathcal{M}_M$, the moment generating function of $T(X)$ is finite:

$$E_M e^{\langle A, T(X) \rangle} = \int_D e^{-\gamma(M - A)} d\nu = e^{\gamma(M - A)} < \infty.$$ 

This is a sufficient condition for the existence of all moments of $T(X)$ [30, Page 33] and thus concludes the proof of the second property. 

A.2. Proof of Lemma 3.2.

**Proof of Lemma 3.2.** The claim follows readily from Lemma 3.3, which is proved in the next section. Indeed, using the second derivates stated in Lemma 3.3, we find for any $M \in \mathcal{M}^*$ and $M' \in \mathbb{R}^{q \times q}$,

$$\sum_{i,j,k,l=1}^q M'_{ij} \frac{\partial}{\partial M_{ij}} \frac{\partial}{\partial M_{kl}} \left(\langle M, T(x) \rangle + \gamma(M)\right) M'_{kl} = n \sum_{i,j,k,l=1}^q M'_{ij} E_M \left[\langle T_{ij}(X) - E_M T_{ij}(X), T_{kl}(X) - E_M T_{kl}(X) \rangle\right] M'_{kl} = n E_M (M', T(X) - E_M T(X))^2.$$
The display implies that for any $M \in \mathcal{M}$ and $M' \in \mathbb{R}^{q \times q}$,

$$\sum_{i,j,k,l=1}^{q} M'_{ij} \frac{\partial}{\partial M_{ij}} \frac{\partial}{\partial M_{kl}} \left( (M, \overline{T}(x))_{tr} + \gamma(M) \right) M'_{kl} \geq 0.$$ 

This ensures convexity, and thus concludes the proof of Lemma 3.2. \qed

A.3. Proof of Lemma 3.3.

**Proof of Lemma 3.3.** We prove the two claims in order.

**Part 1.** We start by taking the first derivative, showing that

$$\nabla_{ij} \left( (M, \overline{T}(x))_{tr} + \gamma(M) \right) = \overline{T}_{ij}(x) - \mathbb{E}_M \overline{T}_{ij}(X),$$

where we use the shorthand notation $\nabla_{ij} := \frac{\partial}{\partial M_{ij}}$.

Since the trace is linear, the derivative of the first term is

$$\nabla_{ij} (M, \overline{T}(x))_{tr} = \overline{T}_{ij}(x).$$

For the second term, recall that the normalization $\gamma$ is given by

$$\gamma(M) = \log \int_D e^{-\langle M, T(x) \rangle_{tr} + \xi(x)} d\nu.$$ 

Taking exponentials on both sides, we find

$$e^{\gamma(M)} = \int_D e^{-\langle M, T(x) \rangle_{tr} + \xi(x)} d\nu.$$ 

We can now take derivatives and get

$$e^{\gamma(M)} \nabla_{ij} \gamma(M) = \int_D \nabla_{ij} e^{-\langle M, T(x) \rangle_{tr} + \xi(x)} d\nu$$

$$= -\int_D T_{ij}(x) e^{-\langle M, T(x) \rangle_{tr} + \xi(x)} d\nu,$$

where we again use the linearity of the trace. Bringing the exponential factor back into the integral and using the assumed independence of the observations then yields

$$\nabla_{ij} \gamma(M) = -\int_D T_{ij}(x) e^{-\langle M, T(x) \rangle_{tr} + \xi(x)} d\nu = -\mathbb{E}_M T_{ij}(X) = -\mathbb{E}_M \overline{T}_{ij}(X).$$

This provides the derivative for the second term. Collecting the pieces concludes the proof of the first part.
Part 2. We now compute the second derivative, showing that
\[
\nabla_{ij} \nabla_{kl} \left( \langle M, T(x) \rangle_{tr} + \gamma(M) \right) = \mathbb{E}_M \left[ \left( T_{ij}(X) - \mathbb{E}_M T_{ij}(X) \right) \left( T_{kl}(X) - \mathbb{E}_M T_{kl}(X) \right) \right],
\]
where we again use the shorthand notation \( \nabla_{ij} = \frac{\partial}{\partial M_{ij}} \).

To prove this claim, recall that by Part 1,
\[
\nabla_{kl} \left( \langle M, T(x) \rangle_{tr} + \gamma(M) \right) = T_{kl}(x) - \mathbb{E}_M T_{kl}(X).
\]

Since the first term is independent of \( M \), we can focus on the second term. Independence of the observations and the model (2) provide
\[
\mathbb{E}_M T_{kl}(X) = \mathbb{E}_M T_{kl}(X) = \int_D T_{kl}(x) e^{-\langle M, T(x) \rangle_{tr} + \xi(x) - \gamma(M)} d\nu.
\]

Taking derivatives, we find similarly as in Part 1
\[
- \nabla_{ij} \mathbb{E}_M T_{kl}(X)
= - \int_D T_{kl}(x) \nabla_{ij} e^{-\langle M, T(x) \rangle_{tr} + \xi(x) - \gamma(M)} d\nu
= - \int_D T_{kl}(x) (-T_{ij}(x) - \nabla_{ij} \gamma(M)) e^{-\langle M, T(x) \rangle_{tr} + \xi(x) - \gamma(M)} d\nu
= - \int_D T_{kl}(x) (-T_{ij}(x) + \mathbb{E}_M T_{ij}(X)) e^{-\langle M, T(x) \rangle_{tr} + \xi(x) - \gamma(M)} d\nu
= \int_D (T_{ij}(x) - \mathbb{E}_M T_{ij}(X)) \left( T_{kl}(x) - \mathbb{E}_M T_{kl}(X) \right) e^{-\langle M, T(x) \rangle_{tr} + \xi(x) - \gamma(M)} d\nu
= \mathbb{E}_M \left[ (T_{ij}(X) - \mathbb{E}_M T_{ij}(X)) (T_{kl}(X) - \mathbb{E}_M T_{kl}(X)) \right]
= n \mathbb{E}_M \left[ (T_{ij}(X) - \mathbb{E}_M T_{ij}(X)) (T_{kl}(X) - \mathbb{E}_M T_{kl}(X)) \right].
\]

Plugging this in above concludes the proof of Part 2. \( \square \)

A.4. Proof of Theorem 3.1.

Proof of Theorem 3.1. A benefit of the introduced framework is that the models form exponential families. This connects our work with a rich literature. In particular, consistency and asymptotic normality of the maximum likelihood estimator can be proved following the arguments in the classical paper [1], see especially [1, Theorems 4.1 and 6.1]. We refer to that paper for details.

There are also more explicit ways for proving Theorem 3.1. For example, a proof can be established via general maximum likelihood theory, see [33,
Theorem 5.7 on Page 45 and Example 19.7 on Page 271] and [34, Example 3.2.24 on Page 306]. For this, three properties of the models need to be verified: first, the function \(M \mapsto \langle M, \mathcal{T}(x) \rangle_{\text{tr}} + \gamma(M)\) fulfills a Lipschitz condition; second, the population version \(M \mapsto \mathbb{E}_M \left[ \langle M, \mathcal{T}(X) \rangle_{\text{tr}} + \gamma(M) \right]\) has a well-separated minimum; and finally, the function \(M \mapsto f_M(x)^{1/2}\) is - in a sense - differentiable. We leave out the details.