Two-sample inference for high-dimensional Markov networks

Byol Kim | Song Liu | Mladen Kolar

Abstract
Markov networks are frequently used in sciences to represent conditional independence relationships underlying observed variables arising from a complex system. It is often of interest to understand how an underlying network differs between two conditions. In this paper, we develop methods for comparing a pair of high-dimensional Markov networks where we allow the number of observed variables to increase with the sample sizes. By taking the density ratio approach, we are able to learn the network difference directly and avoid estimating the individual graphs. Our methods are thus applicable even when the individual networks are dense as long as their difference is sparse. We prove finite-sample Gaussian approximation error bounds for the estimator we construct under significantly weaker assumptions than are typically required for model selection consistency. Furthermore, we propose bootstrap procedures for estimating quantiles of a max-type statistics based on our estimator, and show how they can be used to test the equality of two Markov networks or construct simultaneous confidence intervals. The performance of our methods is demonstrated through extensive simulations. The scientific usefulness is illustrated with an analysis of a new fMRI data set.

Keywords
differential networks, high-dimensional inference, Kullback–Leibler importance estimation procedure, Markov networks, post-regularization inference
INTRODUCTION

Markov networks, also known as Markov random fields or undirected probabilistic graphical models, are successfully used in many application domains to represent interactions between measured components of a complex system and help scientists in uncovering structured information from large amounts of unstructured data (Koller & Friedman, 2009; Lauritzen, 1996; MacKay, 2003). In genetics, the graph structure can be used, for example, to model regulatory activities in gene expressions (Dobra et al., 2004; Hartemink et al., 2001), while in neuroscience, it can be used to model brain network in order to identify features associated with different mental diseases (Supekar et al., 2008). Other successful application areas include social and political sciences (Banerjee et al., 2008), analysis of financial data (Barber & Kolar, 2018), and many others. One of the fundamental problems in statistics is that of learning the graph structure of a probabilistic graphical model based on independent and identically distributed (i.i.d.) samples. See Drton and Maathuis (2017) for a recent overview.

The focus of this paper is on developing a method for statistical inference of parameters in a differential network. For a recent survey, see Shojaie (2021), and references therein. In many applications, interest centres not on a particular network, but rather on whether and how the network changes between different states. For example, genes may regulate each other differently when the external environment is altered. The way different regions of a brain interact together may be altered depending on the activity that a patient is performing. A single graphical model lacks the ability to capture such changes and cannot reflect the dynamic nature of such data, therefore limiting our ability to gain key insights into the underlying system under consideration.

We develop a collection of methods for performing statistical inference on the difference of parameters in high-dimensional Markov networks. Subtleties arise when the target of inference is the difference of parameters rather than the parameters themselves. In high-dimensional regimes, consistent estimation requires an assumption of inherent low-dimensionality such as sparsity (Cai et al., 2011; Friedman et al., 2008; Ravikumar et al., 2011; Yuan, 2010; Yuan & Lin, 2007). Therefore, a crude procedure that estimates the network parameters separately, and then takes the difference, can only work when all the individual networks are sparse. This is quite restrictive for applications where the individual networks may be dense, but the differences are expected to be sparse, say, due to the experimental set-up. Moreover, even when the assumption is satisfied, many such methods have tuning parameters that have an influence on the estimated structure, and it is unclear how they should be combined in practice to yield a consistent estimate of the difference.

This has led many researchers either to jointly estimate structurally similar networks (Chiquet et al., 2011; Danaher et al., 2014; Guo et al., 2011; Ma & Michaillidis, 2016; Majumdar & Michaillidis, 2018; Mohan et al., 2014) or to directly estimate the difference (Fazayeli & Banerjee, 2016; Liu et al., 2017; Xu & Gu, 2016; Zhao et al., 2014). The latter approaches tend to have better sample complexity as well as greater applicability. The methods we propose also belong to the latter category.

Our proposal tries to fill two gaps in the existing literature on differential network estimation. First, the majority of the literature on graphical models are developed assuming a particular observation model, and the growing literature on difference estimation is no exception. For example, Xia et al. (2015) assume that the data are Gaussian, whereas Cai et al. (2019) use an Ising model. By contrast, we work with general Markov random fields; we present a unified framework for statistical inference in differential networks, without the need for developing separate methodology for different distributional assumptions. Following the
development of Sugiyama et al. (2008), Sugiyama et al. (2012), Liu et al. (2014, 2017), Fazayeli & Banerjee (2016), we take the density ratio approach and estimate the difference directly. The last three assume a high-dimensional regime, and study consistency of point estimators defined as solutions to penalized procedures, but the question of statistical inference is unaddressed.

This brings us to the second gap. Most of the existing literature on network difference estimation focuses on producing consistent point estimates, leaving the question of quantifying uncertainty in those estimates largely untouched. The methods we develop in this paper can be used to construct confidence intervals and carry out hypothesis tests about the difference of networks parameters. The theoretical guarantees we provide hold under a fairly weak set of assumptions. In particular, they do not rely on perfect model selection at any stage, which would have necessitated strong assumptions, for example, incoherence and strong signal strength. Certain features of our problem, for example, non-linearity, introduce technical challenges in establishing our theoretical results.

Our paper contributes to the growing literature on statistical inference on high-dimensional parameter estimates. Hypothesis testing and confidence intervals for high-dimensional M-estimators are studied in Zhang and Zhang (2013), Belloni and Chernozhukov (2013), Belloni et al. (2016), Javanmard and Montanari (2014), Meinshausen (2015), van de Geer et al. (2014). Related ideas have been developed in the context of Gaussian graphical models (Janková and van de Geer, 2015, 2017; Ren et al., 2015), elliptical copula models (Barber & Kolar, 2018; Lu et al., 2018), and Markov networks (Wang & Kolar, 2016; Yu et al., 2016). Existing inferential techniques for high-dimensional differential networks rely on Gaussian observation model and separate estimation (Belilovsky et al., 2016; Liu, 2017; Xia et al., 2015). By contrast, our methods also apply to non-Gaussian data and are based on direct difference estimation.

Our Gaussian bootstrap approximation results can be viewed as another contribution along the lines of Chen (2018) and Xue and Yao (2020), which build on the ideas of Chernozhukov et al. (2013, 2015, 2017). In particular, the testing procedure developed in Xue and Yao (2020) relies on a Gaussian approximation result of the difference of two independent sums in high dimensions. Our equal graph test is similar in flavour, but as our estimator cannot be represented as an independent sum, the proof of validity requires a careful control of the remainder.

The rest of this paper is organized as follows. Section 2 discusses some background. Our methods are presented in Section 3, and their theoretical guarantees are given in Section 4. We report the results of our extensive simulation study in Section 5, and analyse a real fMRI data set in Section 6. We conclude with a discussion of alternatives and future directions in Section 7. The proofs of the main results are found in Supplement. A Julia package implementing the proposed methods may be obtained from https://github.com/mlakolar/KLIEPInference.jl, together with the code to reproduce the results.

2 | PRELIMINARIES

We list notations that are used frequently throughout this paper. Vectors are distinguished by bold font, for example, \( \mathbf{v} \). Bold uppercase letters are reserved for matrices, for example, \( \mathbf{M} \). For \( d \in \mathbb{N}, [d] = \{1, \ldots, d\} \). For \( k \in [d], \mathbf{e}_k \in \mathbb{R}^d \) is the kth standard basis vector. For \( \mathbf{v} \in \mathbb{R}^d \) and \( k \in [d] \), we write \( v_k \) for the kth component of \( \mathbf{v} \). For \( S \subseteq [d], \mathbf{v}_S \in \mathbb{R}^d \) with \( v_{S,k} = v_k \) for \( k \in S \), \( v_{S,k} = 0 \) else. Let \( I \subseteq [d] \) be an index set. For a set of scalars \( \{v_k\}_{k \in I}, \{(v_k)_{k \in I}\} \) denotes the \( |I| \)-vector...
with the components given by the set. Similarly, for a set of vectors \( \{\mathbf{v}_k\}_{k \in I} \) with all \( \mathbf{v}_k \in \mathbb{R}^d \), \( [\mathbf{v}_k]_{k \in I} \) denotes the \( d \times |I| \)-matrix with the columns given by the set. Given \( \mathbf{v}_1 \in \mathbb{R}^{d_1} \) and \( \mathbf{v}_2 \in \mathbb{R}^{d_2} \), \( \mathbf{v}_1 \cup \mathbf{v}_2 \in \mathbb{R}^{d_1+d_2} \) denotes their concatenation. For \( \mathbf{v} \in \mathbb{R}^d \), the partition of \( \mathbf{v} \) induced by a partition \( d_1 + d_2 = d \) is denoted \( \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \). Similarly, for \( \mathbf{M} \in \mathbb{R}^{d \times d} \), \( \mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \).

The inner product is denoted as \( \langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v} = \sum_{k=1}^d u_k v_k \). We use \( \| \cdot \| \) to denote a norm on \( \mathbb{R}^d \), and \( \| \cdot \|_s \) to denote its dual, \( \| \mathbf{v} \|_s = \sup_{\|\mathbf{u}\|_1 \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle \). For \( p \in [1, \infty) \), \( \| \mathbf{v} \|_p = \left( \sum_{k=1}^d |v_k|^p \right)^{1/p} \) is the usual \( \ell^p \)-norm of \( \mathbf{v} \). This is extended first to \( q \in (0, 1] \) as \( \| \mathbf{v} \|_q = \sum_{k=1}^d |v_k|^q \), and then to \( q = 0 \) by adopting the convention \( 0^0 \equiv 0 \) so that \( \| \mathbf{v} \|_0 = |\text{supp}(\mathbf{v})| = |\{ k : v_k \neq 0 \}| \). For \( q \in [0, 1) \), \( \ell^q \)-norms can be thought of as generalized sparsity measures.

A Markov network describes conditional dependencies among a collection of random variables (Drton & Maathuis, 2017; Lauritzen, 1996). Let \( \mathbf{x} = (x_v)_{v \in V} \) be a random vector taking values in \( \mathcal{X} \subseteq \mathbb{R}^m \). Consider an undirected graph \( G \) on the node set \( V = |m| \), that is, a pair \( G = (V, E) \), where \( E \), called the edge set, contains unordered pairs of nodes. We say that nodes \( u, v \in V \) are connected by an edge if \( \{u, v\} \in E \). Formally, a Markov network associated with \( G = (|m|, E) \) is a collection of \( m \)-variate distributions such that \( x_u \perp \perp x_v | (x_w)_{w \neq u, v} \) if and only if \( \{u, v\} \notin E \). Thus, the edge set \( E \) describes which pairs of random variables are conditionally independent given all the other variables.

Let \( \mathcal{C}(G) \) denote the set of all cliques of \( G \), that is, subsets of \( V \) for which every pair of nodes is connected by an edge. It is well known that any such \( \mathbf{x} \) with a strictly positive density is an exponential family \( f(\mathbf{x}; \gamma) = \exp(\gamma^\top \psi(\mathbf{x}))/Z(\gamma) \) for some \( \gamma = (\gamma_C)_{C \in \mathcal{C}(G)} \) \( \psi = (\psi_C)_{C \in \mathcal{C}(G)} \), and the normalizing constant \( Z(\gamma) = \int \exp(\gamma^\top \psi(\mathbf{x}))d\mathbf{x} \), where \( \gamma_C \in \mathbb{R} \) and \( \psi_C \) is a function of the clique variables \( (x_v)_{v \in C} \) only (Hammersley & Clifford, 1971). A parametric class \( \mathcal{F}_\gamma \) of Markov networks is obtained by assuming a fixed \( \psi \).

A special case of significance is the class of pairwise Markov networks (Wainwright & Jordan, 2008; Yang et al., 2015) that have densities of the form

\[
f(\mathbf{x}; \gamma) = \frac{1}{Z(\gamma)} \exp \left( \sum_{v=1}^m \gamma_v \psi_v(x_v) + \sum_{u=1}^m \sum_{v=u+1}^m \gamma_{uv} \psi_{uv}(x_u, x_v) \right),
\]

2.1 Statement of the problem
where \( y = (y_v)_{v=1}^m \cup (y_{uv})_{1 \leq u < v \leq m} \). \( \psi = (\psi_v)_{v=1}^m \cup (\psi_{uv})_{1 \leq u < v \leq m} \). For this class, each component function of \( \psi \) is at most a function of two variables, and hence \( x_u \perp \perp x_v \mid (x_{w})_{w \neq u,v} \) if and only if \( y_{uv} = 0, u \neq v \). Thus, for a pairwise Markov network, the edge set \( E \) dictates which of the pairwise parameters \( y_{uv} \) are non-zero. A number of well-studied models belong to the pairwise class.

**Example 1** (Ising models) An Ising model is a family of discrete probability distribution on the vertices of the \( m \)-dimensional hypercube \( \mathbb{X} = \{ \pm 1 \}^m \) given by the probability mass function of the form Equation (1) with \( \psi_v(x_v) = x_v, \psi_{uv}(x_u, x_v) = x_u x_v \), and \( y_v, y_{uv} \in \mathbb{R} \). Thus, the Markov network associated with \( G = ([m], E) \) are all Ising models with \( y_{uv} \neq 0 \) if and only if \( \{ u, v \} \in E \).

**Example 2** (Gaussian graphical models) The most-studied example of a probabilistic graphical model is the case of the undirected Gaussian graphical model. Suppose \( x \sim \mathcal{N}(\mu, \Sigma) \). Then, \( x \) has a density of the form Equation (1) with \( \psi_v(x_v) = x_v, \psi_{uv}(x_u, x_v) = x_u x_v \), \( y_v = (\Sigma^{-1})_{uv} \), and \( y_{uv} = -[\Sigma^{-1}]_{uv}/2 \). Thus, if \( x \) is in a Gaussian graphical model with the graph \( G = ([m], E) \), then the inverse covariance matrix satisfies \( [\Sigma^{-1}]_{uv} \neq 0 \) if and only if \( \{ u, v \} \in E \).

Suppose \( f_x = f(\cdot; y_x) \) and \( f_y = f(\cdot; y_y) \) are two distributions from the same pairwise family (1) corresponding to parameters \( y_x \) and \( y_y \), respectively. Then, the change from \( f_x = f(\cdot; y_x) \) to \( f_y = f(\cdot; y_y) \) can be described by the difference \( \theta^* = y_x - y_y \). In particular, whenever \( x_u \perp \perp x_v \mid (x_{w})_{w \neq u,v} \) is true for only one of \( f_x \) or \( f_y \), we have \( \theta_{uv}^* = y_{uv} - y_{uv} \neq 0 \). More generally, the support of \( \theta^* \) gives the pairs of random variables for which the conditional dependence relationship has changed.

The **differential network** is defined as the difference \( \theta^* \) of \( y_x \) and \( y_y \). We represent the differential network with a graph \( G = (V, E) \) where an edge \( \{ u, v \} \in E \) is drawn between vertices \( u \) and \( v \) if and only if \( \theta_{uv}^* \neq 0 \). Our goal here is to learn the differential network given independent and identically distributed (i.i.d.) observations from each of \( f_x = f(\cdot; y_x) \) and \( f_y = f(\cdot; y_y) \). More precisely, using \( x^{(1)}, \ldots, x^{(n_x)} \sim f_x \) and \( y^{(1)}, \ldots, y^{(n_y)} \sim f_y \), we would like to construct confidence intervals or conduct hypothesis tests over possibly high-dimensional sub-vectors of \( \theta^* \) with provably valid simultaneous guarantee at arbitrary user-specified confidence level of \( 1 - \alpha \) for small \( \alpha \in [0, 1] \). This requires an estimate of \( \theta^* \), which we construct in the next section based on the density ratio \( f_x/f_y \) without separately estimating the individual parameters \( y_x \) and \( y_y \).

### 2.2 Direct difference estimation via density ratio

We describe the Kullback–Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2008) and how it can be used to directly estimate the differential network \( \theta^* \).

KLIEP is a framework for estimating the density ratio of two probability distributions based on i.i.d. observations from each. When the distributions are from the same parametric exponential family, the density ratio depends on the underlying pair of parameters only through their difference while maintaining the exponential form. Indeed, let \( r_{\theta^*} = f_x/f_y \). Then,

\[
r_{\theta^*}(x) = \frac{f_x(x)}{f_y(x)} = \frac{Z(y_x) \exp \left( y_x^\top \psi(x) \right)}{Z(y_y) \exp \left( y_y^\top \psi(x) \right)} = \frac{\exp \left( \theta^\top \psi(x) \right)}{Z_y(\theta^*)},
\]
where we have \( Z_y(\theta^*) = \mathbb{E}_y[\exp(\theta^{*\top}\psi(y))] \), because

\[
Z_y(\theta^*) = \frac{Z(y_x)}{Z(y_y)} = \frac{\int \exp(\gamma_x^\top\psi(x)) \, dx}{\int \exp(\gamma_y^\top\psi(x)) \, dx} = \mathbb{E}_x[\exp(\theta^{*\top}\psi(x))] = \mathbb{E}_y[\exp(\theta^{*\top}\psi(y))].
\]

This can be used to derive a procedure that directly learns \( \theta^* \) without having to learn either \( y_x \) or \( y_y \). Let \( D_{KL}(f \parallel g) \) be the Kullback-Leibler (KL) divergence for probability densities \( f \) and \( g \). Recall that \( D_{KL}(f \parallel g) \geq 0 \) with equality if and only if \( f = g \) almost everywhere. Since \( f_x = r_{\theta^*}f_y \), \( \theta^* = \arg\min_\theta D_{KL}(f_x \parallel r_{\theta}f_y) \). Moreover, it is proved in Supplement A.1 that

\[
\theta^* = \arg\min_\theta D_{KL}(f_x \parallel r_{\theta}f_y) = \arg\min_\theta \left\{ -\mathbb{E}_x[\theta^\top\psi(x)] + \log \mathbb{E}_y[\exp(\theta^{*\top}\psi(y))] \right\}, \tag{2}
\]

where \( \mathbb{E}_x \) denotes the expectation with respect to \( f_x \) and \( \mathbb{E}_y \) the expectation with respect to \( f_y \). The empirical KLIPE loss \( \ell_{\text{KLIPE}} \) is obtained by replacing each expectation with the corresponding sample average:

\[
\ell_{\text{KLIPE}}(\theta) = \ell_{\text{KLIPE}}(\theta; x^{(1)}, \ldots, x^{(n_x)}, y^{(1)}, \ldots, y^{(n_y)}) = -\frac{1}{n_x} \sum_{i=1}^{n_x} \theta^\top\psi(x^{(i)}) + \log \left\{ \frac{1}{n_y} \sum_{j=1}^{n_y} \exp(\theta^{*\top}\psi(y^{(j)})) \right\}. \tag{3}
\]

Minimizing \( \ell_{\text{KLIPE}} \) yields the KLIPE estimate \( \hat{\theta}_{\text{KLIPE}} = \arg\min_\theta \ell_{\text{KLIPE}}(\theta) \) as a direct estimate of the differential network \( \theta^* \). The function \( \ell_{\text{KLIPE}} \) is convex in \( \theta \), and when it is strictly convex—which requires \( n_y > p \)—the KLIPE estimate \( \hat{\theta}_{\text{KLIPE}} \) is known to be approximately normal and unbiased (Sugiyama et al., 2012, Chapter 13).

In the high-dimensional setting with \( n_y \leq p \), the minimizer of \( \ell_{\text{KLIPE}} \) is no longer unique, and regularization becomes necessary for consistent estimation. In this setting, Liu et al. (2017) and Fazayeli and Banerjee (2016) proposed regularized versions of KLIPE using norm penalties. In particular, Liu et al. (2017) proposed the sparse KLIPE

\[
\hat{\theta} = \hat{\theta}(\lambda) = \arg\min_\theta \ell_{\text{KLIPE}}(\theta; x^{(1)}, \ldots, x^{(n_x)}, y^{(1)}, \ldots, y^{(n_y)}) + \lambda\|\theta\|_1, \tag{4}
\]

where \( \lambda > 0 \) is a regularization parameter to be chosen by the user. They show that when \( \theta^* \) is sparse, the support of \( \hat{\theta} \) consistently recovers the support of \( \theta^* \) for suitable choices of \( \lambda \). However, such results typically require additional conditions, for example a lower bound on the minimal signal strength and incoherence of the Hessian, which may be restrictive for many real data applications. Furthermore, these are essentially results about the accuracy of the point estimates, whereas to construct confidence intervals or conduct hypothesis tests, one needs information about the distribution of the estimators. This is difficult for regularized estimators, as we shall see next.
### 2.3 De-biasing

Challenges arise when a regularized estimator \( \hat{\theta} \), for example the sparse KLIEP estimator (4), is used for statistical inference. Regularization produces a non-negligible bias, and the distribution of the resulting estimator is typically intractable (see Ning & Liu, 2017, and references therein).

We propose to deal with this issue by de-biasing each component of \( \hat{\theta} \). For convenience, adopt a linear indexing so that \( y = (y_k)_{k=1}^p \) and \( \psi = (\psi_k)_{k=1}^p \), where \( p \) is the total number of parameters. Suppose we wish to obtain a de-biased estimate of \( \theta^*_k \) for some \( k \in [p] \). Let \( \theta^*_k = \theta^*_{[p]\backslash\{k\}} \in \mathbb{R}^{p-1} \) denote the vector of remaining \( p-1 \) parameters. This is the nuisance parameter for carrying out statistical inference for \( \theta^*_k \). Abusing the notation somewhat, we write the resulting partition as \( \theta = (\theta_k, \theta_{k^c}) \). Define \( \omega_k^* \) as the vector satisfying \( \mathbb{E}[\nabla^2 \ell_{KLIEP}(\theta^*)] \omega_k^* = e_k \), and let \( \tilde{\omega}_k \) be a consistent estimator of \( \omega_k^* \).

Our method offers two options for constructing an approximately normal and unbiased estimator \( \hat{\theta}_k \) of \( \theta^*_k \) that are asymptotically equivalent (Chernozhukov et al., 2015). The first option is to use the one-step estimator (van de Geer et al., 2014; van der Vaart, 1998; Zhang & Zhang, 2013):

\[
\hat{\theta}_k^{1+} = \hat{\theta}_k - \tilde{\omega}_k^\top \nabla \ell_{KLIEP}(\hat{\theta}).
\] (5)

This approximately solves a modified score equation \( \tilde{\omega}_k^\top \nabla \ell_{KLIEP}(\theta_k, \hat{\theta}_{k^c}) = 0 \), where \( \hat{\theta}_{k^c} \) is defined via \( \hat{\theta} = (\hat{\theta}_k, \hat{\theta}_{k^c}) \), by taking one Newton iteration starting from \( \hat{\theta}_k \). In Section 4.2, we prove that the one-step estimator \( \hat{\theta}_k^{1+} \) is an approximately normal and unbiased estimator of \( \theta^*_k \).

When \( \hat{\theta} \) and \( \tilde{\omega}_k \) are both sparse vectors, de-biasing may be carried out via the so-called double selection (Chernozhukov et al., 2015). Let \( \hat{\theta} \) be the estimate obtained by re-fitting to the union of the supports of \( \hat{\theta} \) and \( \tilde{\omega}_k \), that is,

\[
\hat{\theta} = \arg \min_{\theta} \ell_{KLIEP}(\theta) \quad \text{subject to} \quad \text{supp}(\theta) \subseteq \{k\} \cup \text{supp}(\hat{\theta}) \cup \text{supp}(\tilde{\omega}_k).
\] (6)

Then, the double-selection estimator \( \hat{\theta}_k^{2+} \) is defined as the \( k \)th component of \( \hat{\theta} \). Intuitively, by including the estimated supports of both \( \theta^* \) and \( \omega_k^* \), the double-selection procedure achieves robustness to errors from either model selection procedure. Provided that \( \hat{\theta} \) is as accurate as \( \hat{\theta} \)—which would be the case for sparse or approximately sparse \( \theta^* \) and \( \omega_k^* \)—\( \hat{\theta}_k^{2+} \) is asymptotically equivalent to \( \hat{\theta}_k^{1+} \).

For a derivation of Equation (5) in the context of KLIEP, see Supplement B.1. A general discussion of the relationship of one-step estimation and double selection may be found in Chernozhukov et al. (2015).

### 3 METHODOLOGY

We propose a procedure for constructing an approximately normal and unbiased estimator of the differential network (Section 3.1). We then give two bootstrap sketching procedures for estimating the quantiles of a max-type statistic based on the estimator from Section 3.1, and show how they can be used for simultaneous inference (Section 3.2).
3.1 Sparse Kullback–Leibler Importance Estimation with de-biasing (SparKLIE+)

We present Procedure 1, which is a general recipe for de-biasing regularized KLIEP estimates for each $\theta^*_k$ in $k \in I$, where $I \subseteq [p]$ is the set of indices for the parameters of inferential interest. The procedure uses a general norm penalty for regularization.

**Procedure 1. Kullback-Leibler Importance Estimation With de-biasing (KLIE+)**

**Input:** Data $X_{n_x} = \{x^{(i)}\}_{i=1}^{n_x}$, $Y_{n_y} = \{y^{(j)}\}_{j=1}^{n_y}$, positive regularization parameters $\lambda_\theta, \lambda_k$, $k \in I$

**Output:** De-biased estimates $\tilde{\theta}_k$, $k \in I$

1. Find an initial estimate of $\theta^*$

\[
\tilde{\theta} = \arg \min_{\hat{\theta}} \ell_{\text{KLIEP}}(\theta; X_{n_x}, Y_{n_y}) + \lambda_\theta \| \theta \|.
\]  

(7)

for $k \in I$ do

2. Find an initial estimate of $\omega_k^*$

\[
\hat{\omega}_k = \arg \min_{\omega} \frac{1}{2} \omega^T \nabla^2 \ell_{\text{KLIEP}}(\tilde{\theta}) \omega - \omega^T \epsilon_k + \lambda_k \| \omega \|.
\]  

(8)

3. De-bias, either by (5) or by (6), to obtain $\hat{\theta}_k$.

end for

return $\hat{\theta}_k$, $k \in I$

A general Gaussian approximation bound for Procedure 1 will be given below in Theorem 1 in Section 4.2. The result is valid as long as the initial estimators from Equations (7) and (8) are sufficiently accurate. For example, this is the case for sparse or approximately sparse $\theta^*$ and $\omega_k^*$ when the $\ell_1$-penalty is used (Lemmas 1 and 2 in Supplement C.3). We call this procedure Sparse Kullback–Leibler Importance Estimation with de-biasing (SparKLIE+), with SparKLIE+1 referring to SparKLIE+ that uses one-step (5) for de-biasing and SparKLIE+2 referring to the double selection (6) option.

**Remark 1** (Alternative procedures for initial estimation) It is possible to use other procedures for either of the initial estimation steps as long as the errors satisfy $\| \hat{\theta} - \theta^* \| \cdot \| \hat{\omega}_k - \omega_k^* \| = o_p(n^{-1/2})$. We give examples in the case of the $\ell_1$-penalty. In Supplement G.1, we give Procedures 4 and 5 which may be performed in Steps 1 or 2, respectively. The main advantage of these procedures is that the user only has to specify a universal penalty level which can be done in a data-independent manner. For example, in Procedure 4, $\lambda_{\theta_0} = 1.01\Phi^{-1}(1 - 0.05/p)$ following Belloni et al. (2014), and in Procedure 5, $\lambda_0 = \sqrt{2\log p/n_y}$ following Sun and Zhang (2013). We may also re-fit the model on the estimated support (Belloni & Chernozhukov, 2013). Finally, it is also possible to use a constrained procedure, similar to the method of Ning and Liu (2017), where instead of Equation (8), one solves

\[
\min \| \omega \|_1 \text{ subject to } \| \nabla^2 \ell_{\text{KLIEP}}(\tilde{\theta}) \omega - \epsilon_k \|_\infty \leq \lambda_k.
\]

**Remark 2** (Regularization parameters) Procedure 1 assumes that the user has already picked out the regularization parameters $\lambda_\theta, \lambda_k$, $k \in I$. However, the optimal choice, as dictated by Lemmas 7 and 8 in Supplement E.1, depends on constants related to the regularity of the
density ratio, which are typically unknown. In Supplement I.3, we empirically study the sensitivity of Procedure 1 to the choice of regularization parameters and find that the performance is robust across a wide range of regularization levels. Furthermore, as stated above in Remark 1, we provide alternative initial estimation procedures in Supplement G.1 that do not require regularization parameter tuning. This is the version of Procedure 1 we use in Sections 5 and 6.

### 3.1.1 Variance of the SparKLIE+ estimator

For statistical inference, we also need a consistent estimator of the variance of $\sqrt{n} \hat{\theta}_k$, $n = n_x + n_y$. Define the empirical density ratio estimate

$$
\hat{r}_\theta(y) = \exp \left( \theta^T \psi(y) \right) / \hat{Z}_y(\theta), \quad \text{where} \quad \hat{Z}_y(\theta) = \frac{1}{n_y} \sum_{j=1}^{n_y} \exp \left( \theta^T \psi(y^{(j)}) \right).
$$

Let $\hat{S}_\psi$ and $\hat{S}_{\theta \hat{\theta}'}$ be the sample covariance matrices of $\{\psi(x^{(i)})\}_{i=1}^{n_x}$ and $\{\psi(y^{(j)})\hat{r}_\theta(y^{(j)})\}_{j=1}^{n_y}$, that is,

$$
\hat{S}_\psi = \frac{1}{n_x} \sum_{i=1}^{n_x} \psi(x^{(i)})\psi(x^{(i)})^T - \overline{\psi}\overline{\psi}^T,
$$

$$
\hat{S}_{\theta \hat{\theta}'}(\theta) = \frac{1}{n_y} \sum_{j=1}^{n_y} \hat{r}_\theta(y^{(j)})\psi(y^{(j)})\psi(y^{(j)})^T - \hat{\mu}(\theta)\hat{\mu}(\theta)^T,
$$

where

$$
\overline{\psi} = \frac{1}{n_x} \sum_{i=1}^{n_x} \psi(x^{(i)}), \quad \hat{\mu}(\theta) = \frac{1}{n_y} \sum_{j=1}^{n_y} \psi(y^{(j)})\hat{r}_\theta(y^{(j)}).
$$

Let $\hat{S}_{\text{pooled}}(\hat{\theta})$ be the pooled covariance

$$
\hat{S}_{\text{pooled}}(\hat{\theta}) = \frac{n}{n_x} \hat{S}_\psi + \frac{n}{n_y} \hat{S}_{\theta \hat{\theta}'}(\hat{\theta}).
$$

Finally, a consistent estimator of the variance of $\sqrt{n} \hat{\theta}_k$ is

$$
\hat{\sigma}_k^2 = \hat{\omega}_k^T \hat{S}_{\text{pooled}}(\hat{\theta}) \hat{\omega}_k.
$$

This estimates the variance of $\sqrt{n} \hat{\omega}_k^* \nabla \ell_{\text{KLIEP}}(\theta^*)$, which we show is asymptotically equivalent to $\sqrt{n} (\hat{\theta}_k - \theta_k^*)$ in the proof of Theorem 1 in Supplement B.2. By Lemma 18 in Supplement F.2, $\hat{\sigma}_k^2$ is consistent if both $\hat{\theta}$ and $\hat{\omega}_k$ are.

Theorem 2 in Section 4.2 implies that if $z_q = \Phi^{-1}(q)$ is the $q$-quantile of a standard Gaussian, then $\mathbb{P} \{ \sqrt{n} (\hat{\theta}_k - \theta_k^*) / \hat{\sigma}_k \leq z_q \} \approx \Phi^{-1}(z_q) = q$. Thus, $\hat{\theta}_k \pm z_{1 - \alpha/2} \times \hat{\sigma}_k / \sqrt{n}$ is an asymptotically valid $100 \times (1 - \alpha)\%$ confidence interval (CI) for $\theta_k^*$. Similarly, the test that rejects for $\sqrt{n} | \hat{\theta}_k - \theta_{0k} | / \hat{\sigma}_k > z_{1 - \alpha/2}$ is asymptotically level-$\alpha$ for the one-dimensional null hypothesis $H_{0k}: \theta_k^* = \theta_{0k}$. In Section 5, we verify with simulations that the approximations are fairly accurate and robust even at small sample sizes.
3.2 High-dimensional inference via bootstrap sketched quantiles

In Section 3.1, we proposed SparKLIE+, a procedure for obtaining an asymptotically unbiased estimator of a component of the differential network. Iterating Step 3 of SparKLIE+ over all edges yields an unbiased estimator \( \hat{\Theta} \) of the differential network \( \Theta^* \). To make inferences about the structure of \( \Theta^* \) using \( \hat{\Theta} \), one may construct a simultaneous confidence region or conduct a simultaneous hypothesis test. This raises issues of multiple comparisons.

We deal with this problem by a bootstrap approximation of the quantiles of the following statistic

\[
T = T_{n_x, n_y} = \max_{k=1}^{p} \sqrt{n} |\hat{\theta}_k - \theta^*_k|, \quad \text{where } n = n_x + n_y. \tag{13}
\]

Let \( c_{T,q} \) be the \( q \)-quantile of \( T \). Then, it is easy to verify that \( \hat{\Theta} \pm c_{T,1-\alpha}/\sqrt{n} \) is a \( 100 \times (1 - \alpha) \)\% confidence region for \( \Theta^* \). Similarly, the test that rejects if \( \max_k |\hat{\theta}_k| > c_{T,1-\alpha}/\sqrt{n} \) controls the family-wise error rate at level \( \alpha \) for the null hypothesis \( H_0: \theta^*_k = 0 \) for all \( k \in [p] \). This approach has the advantage of adapting to the correlations among \( \Theta^* \). Thus, given \( c_{T,q} \) —or an accurate estimator thereof—we can learn the differential network structure while controlling the type I error rate.

However, in high dimensions, it is itself a highly non-trivial problem to estimate \( c_{T,q} \) with sufficient accuracy (see Chernozhukov et al., 2013, 2017; Deng & Zhang, 2020, and references therein). In this section, we present two bootstrap-based methods for estimating \( c_{T,q} \).

Our first proposal employs the Gaussian multiplier bootstrap. Recall the definitions of \( \bar{\psi} \) from Equation (9), and of \( \Psi \) from Equation (10).

**Procedure 2.** Gaussian multiplier bootstrap sketching for estimating quantiles of \( T \)

**Input:** Data \( X_{n_x} = \{x^{(i)}\}_{i=1}^{n_x} \), \( Y_{n_y} = \{y^{(j)}\}_{j=1}^{n_y} \); the outputs \( \hat{\Theta} \) and \( \hat{\omega}_k, k \in I \), of (7) and (8) from Procedure 1

**Output:** A Gaussian bootstrap estimate \( \hat{c}_{T,q} \) of \( c_{T,q} \)

for \( b = 1, \ldots, n_b \) do

Draw Gaussian weights \( \xi^{(b,1)}_x, \ldots, \xi^{(b,n_x)}_x, \xi^{(b,1)}_y, \ldots, \xi^{(b,n_y)}_y \) iid \( \sim \mathcal{N}(0, 1) \).

Compute

\[
\hat{T}^{(b)}(\hat{\omega}_k) = \max_k \sqrt{n} \left\| \hat{\omega}_k - \frac{1}{n_x} \sum_{i=1}^{n_x} \left( \psi(x^{(i)}) - \bar{\psi} \right) \xi^{(b,i)}_x - \frac{1}{n_y} \sum_{j=1}^{n_y} \left( \psi(y^{(j)}) \hat{\theta}(y^{(j)}) - \bar{\mu}(\hat{\Theta}) \right) \xi^{(b,j)}_y \right\|. \tag{14}
\]

end for

**Return** \( \hat{c}_{T,q} \), the \( q \) sample quantile of \( \{\hat{T}^{(b)} : b = 1, \ldots, n_b\} \).

Procedure 2 may be a procedure for estimating the \( (1 - \alpha) \)-quantile of the maximum of \( |\mathcal{N}(0, \hat{\Sigma})| \), where \( \hat{\Sigma} = \hat{\Omega}^T \hat{S}_{\text{pooled}} \hat{\Omega} \), \( \hat{\Omega} = [\hat{\omega}_k]_{k=1}^p \), and \( \hat{S}_{\text{pooled}} \) is defined in Equation (12). Since we can show that
\(\hat{\theta} - \theta^* \approx \mathcal{N}(0, \Sigma^*)\) for some fixed \(\Sigma^*\) and, moreover, \(\hat{\Sigma} \approx \Sigma^*\), we claim that \(\hat{c}_{T,q}\) is a good estimate of the \(q\)-quantile of \(T\). This intuition is formally stated in Theorem 3 in Section 4.3.

Although Procedure 2 is accurate for sufficiently large sample sizes, at smaller values of \(n_x\) and \(n_y\), empirical bootstrap tends to yield more robust estimates of the quantiles. The procedure below, based on the empirical bootstrap, is what we recommend in practice.

**Procedure 3. Empirical bootstrap sketching for estimating quantiles of \(T\)**

**Input:** Data \(X_{n_x} = \{x(i)\}_{i=1}^{n_x}, Y_{n_y} = \{y(j)\}_{j=1}^{n_y}\); the outputs \(\hat{\theta}\) and \(\hat{\omega}_k\), \(k \in \mathcal{I}\), of (7) and (8) from Procedure 1

**Output:** An empirical bootstrap estimate \(\hat{c}_{T,q}\) of \(c_{T,q}\)

for \(b = 1, \ldots, n_b\) do

- Re-sample \(X_{n_x}^{(b)} = \{x^{(b,1)}, \ldots, x^{(b,n_x)}\}\) and \(Y_{n_y}^{(b)} = \{y^{(b,1)}, \ldots, y^{(b,n_y)}\}\) uniformly at random with replacement.

for \(k \in \mathcal{I}\) do

- For replicating SparKLIE+1 estimate (5), \(\hat{\theta}_k^{(b)} = \hat{\theta}_k - \hat{\omega}_k^\top \nabla \ell_{\text{KLIEP}}(\theta, \hat{X}_{n_x}^{(b)}, \hat{Y}_{n_y}^{(b)})\).

- For replicating SparKLIE+2 estimate (6), \(\hat{\theta}_k^{(b)}\), the \(k\)th component of \(\arg\min_{\hat{\theta}} \ell_{\text{KLIEP}}(\theta, \hat{X}_{n_x}^{(b)}, \hat{Y}_{n_y}^{(b)})\) subject to \(\text{supp}(\theta) \subseteq \{k\} \cup \text{supp}(\hat{\theta}) \cup \text{supp}(\hat{\omega}_k)\).

end for

Compute

\[
\hat{T}^{(b)} = \max_k \sqrt{n} |\hat{\theta}_k^{(b)} - \hat{\theta}_k|.
\] (15)

end for

return \(\hat{c}_{T,q}\), the \(q\) sample quantile of \(\{\hat{T}^{(b)} : b = 1, \ldots, n_b\}\).

Note that only Step 3 of Procedure 1 is repeated in Procedure 3. This is akin to the use of \(\hat{\theta}\) and \(\hat{\omega}_k\), \(k \in \mathcal{I}\), in Procedure 2.

We give a heuristic argument in support of Procedure 3, leaving the formal proof to future work. For the sake of argument, consider the infeasible estimator \(\hat{\theta}_k^{(b)} = \theta_k^* - \omega_k^* \nabla \ell_{\text{KLIEP}}(\theta^*)\) or \(\hat{\theta}_k^{(b)}\), the \(k\)th component of \(\arg\min_{\theta} \nabla \ell_{\text{KLIEP}}(\theta)\) subject to \(\text{supp}(\theta) \subseteq \{k\} \cup \text{supp}(\theta^*) \cup \text{supp}(\omega_k^*)\).

In other words, \(\hat{\theta}_k^{(b)}\) or \(\hat{\omega}_k^{(b)}\) is the result of applying Equation (5) or (6), but with the true parameters \(\theta^*\) and \(\omega_k^*\) replacing the initial estimates \(\hat{\theta}\) and \(\hat{\omega}_k\). It is easy to see that both \(\hat{\theta}_k^{(b)}\) and \(\hat{\omega}_k^{(b)}\) are approximately normal and unbiased estimators, and that making the same replacement in Procedure 3 would yield bootstrap replicates of \(\hat{\theta}_k^{(b)}\) and \(\hat{\omega}_k^{(b)}\). Because \(\hat{\theta}\) and \(\hat{\omega}_k\) are consistent estimators, we expect Procedure 3 to be approximately valid for bootstrapping the SparKLIE+ estimator \(\hat{\theta}_k^{(b)}\) or \(\hat{\omega}_k^{(b)}\). This intuition is verified in simulations in Section 5.2.

4 | THEORY

In this section, we establish statistical validity of the inference procedures discussed in Sections 3.1 and 3.2 under two model assumptions introduced in Section 4.1.
4.1 Assumptions

We discuss two sufficient conditions that imply the accuracy of Gaussian approximation. The first is about the regularity of the density ratio \( r_\theta(y) \).

Condition 1 (bounded density ratio model) There exist \( \rho > 0 \) such that

\[
M_r^{-1} \leq r_\theta(y) \leq M_r \text{ a.s. for all } \theta \text{ with } ||\theta - \theta^*|| \leq \rho
\]

for some \( M_r = M_r(\rho) \geq 1 \).

For convenience, we fix \( \rho = ||\theta^*|| \). Proposition 1 says that Condition 1 is equivalent to a boundedness condition on the sufficient statistics, a claim that was stated without proof for \( \ell_2 \)-norm in Liu et al. (2017). We generalize the statement, and prove it in Supplement D.1.

Proposition 1 (bounded sufficient statistics) Condition 1 is satisfied if and only if \( \|\psi(x)\|_s \leq M_\psi \) a.s. for some \( M_\psi < \infty \).

In general, regularity conditions on the density ratio tend to induce even stronger regularity conditions on the sufficient statistics. The identity \( \hat{Z}_y(\theta)/Z_y(\theta) \equiv n_y^{-1} \sum_{j=1}^{n_y} r_\theta(y(j)) \) implies \( \hat{Z}_y(\theta)/Z_y(\theta) \in [M_r^{-1}, M_r] \). Moreover, \( \hat{r}_\theta(y) \equiv (\hat{Z}_y(\theta)/Z_y(\theta))r_\theta(y) \), so that

\[
M_r^{-2} \leq M_r^{-1} (1 - o_\phi(1)) \leq \hat{r}_\theta(y) \leq M_r (1 + o_\phi(1)) \leq M_r^2.
\]

The outer bounds are obvious. The inner bounds require a concentration result (Lemma 5 in Supplement D.1).

When Procedure 1 is implemented with the \( \ell_1 \)-penalty, it is natural to impose Condition 1 with the \( \ell_1 \)-norm, which by Proposition 1 is equivalent to imposing an \( \ell_\infty \)-bound on the sufficient statistics. Thus, this choice of penalty works nicely with models that take values on a bounded domain, such as Ising models or Potts models. Indeed, for the Ising model defined in Example 1, \( \|\psi(x)\|_\infty = 1 \) but \( \|\psi(x)\|_2^2 = p \).

The second are regularity conditions on the population covariances of \( \psi(x) \) under \( f_x \) and \( f_y \), as well as that of \( \langle \psi(y) - \mu_\psi \rangle \psi_{\theta^*}(y) \) under \( f_y \). Recall \( \Sigma_\psi = \text{Cov}_x[\psi(x)] \), and let \( \Sigma_{\psi_r} = \text{Cov}_y[\langle \psi(y) - \mu_\psi \rangle \psi_{\theta^*}(y)] \), where \( \mu_\psi = E_x[\psi(x)] = E_y[\psi(y)\psi_{\theta^*}(y)] \).

Condition 2 (bounded population eigenvalues) There exist \( 0 < \kappa \leq \bar{\kappa} < \infty \) such that

\[
\kappa \leq \min_{||v||=1, v \neq 0} v^T \Sigma_\psi v \leq \max_{||v||=1, v \neq 0} v^T \Sigma_\psi v \leq \bar{\kappa},
\]

\[
\kappa \leq \min_{||v||=1, v \neq 0} v^T \Sigma_{\psi r} v \leq \max_{||v||=1, v \neq 0} v^T \Sigma_{\psi r} v \leq \bar{\kappa}.
\]

Condition 2 is a natural one, and ensures that the problem is well-behaved (Liu et al., 2017). A lower bound on the minimum eigenvalues ensures that the model is non-degenerate. The upper bound ensures that \( \ell_{\text{KLIEP}}(\theta) \) is smooth, and can be regarded as analogous to the assumption on the log-normalizing function in Yang et al. (2015). These bounds will naturally appear in bounding the convergence of \( \nabla^2 \ell_{\text{KLIEP}}(\theta) \) to \( \Sigma_\psi \), as well as in bounding the variance of the estimator \( \sigma_k^2 \).
The conditions imposed here are weaker than those in Liu et al. (2017), as we do not hope to correctly identify the support of the parameter $\theta^*$. In particular, we do not need to assume the incoherence condition, nor do we need to require that the non-zero components of $\theta^*$ be large enough.

Recall $\theta^* = \gamma_x - \gamma_y$ and $\omega_k^* = \Sigma_{\psi}^{-1} e_k$ where $\Sigma_{\psi} = \text{Cov}_k[\psi(x)]$ and $k \in [p]$. To facilitate the discussion of rates in the next two sections, we introduce additional notations. Let $n = n_x + n_y$. We view $n_x$, $n_y$, $p$, $s_\theta = s_\theta,q_\psi = \|\theta^*\|_{q_\psi}$, $s_k$ as sequences indexed by $n$ and possibly diverging to $\infty$. $n_x$ and $n_y$ are characterized by sequences $n_{x,n}$ and $n_{y,n}$ in $(0,1)$ such that $n_{x,n} + n_{y,n} \equiv 1$, $n_{x,n} = n_x n_n$ and $n_{y,n} = n_y n_n$. In particular, this implies that $n \asymp n_x \asymp n_y$.

The bounds we give below are finite-sample in the sense that they are given as functions of $n$, $p$, $s_\theta$, $s_k$. They can be used to study the asymptotic behaviour as $n \to \infty$ by considering a sequence of models $(\theta^*, \Sigma_{\psi}) = (\theta^*_n, \Sigma_{\psi,n})$ such that the induced sequence of $p$, $s_\theta$, $s_k$, etc. satisfy the side conditions of each theorem.

### 4.2 Finite-sample Gaussian approximation result for the SparKLIE+1

Theorem 1 gives a family of Gaussian approximation bounds for Procedure 1.

Let $k \in [p]$. Let $\hat{\theta}$ and $\hat{\omega}_k$ denote the outputs of Steps 1 and 2 of Procedure 1. For $\lambda_\theta, \lambda_k, \delta_\theta, \delta_k, \delta_\sigma \in [0,1)$, define an event

$\mathcal{E}_{\text{one}} = \mathcal{E}_{\text{one}}(\lambda_\theta, \lambda_k, \delta_\theta, \delta_k, \delta_\sigma) =
\begin{align*}
\{ & (G.1) 2\|\nabla^2 \ell_{\text{KLIE}}(\theta^*)\|_* \leq \lambda_\theta, \quad (G.2) 2\|\nabla^2 \ell_{\text{KLIE}}(\theta^*)\omega^*_k - e_k\|_* \leq \lambda_k, \\
& (E.1) \|\hat{\theta} - \theta^*\|_* \leq \delta_\theta, \quad (E.2) \|\hat{\omega}_k - \omega^*_k\| \leq \delta_k, \\
& (B.1) \left|1 - \frac{\tilde{Z}_y(\theta^*)}{\tilde{Z}_x(\theta^*)}\right| \leq \lambda_\theta, \quad (B.2) \left|\frac{1}{n_y} \sum_{j=1}^{n_y} \left\langle \omega^*_k, \mu_y - \psi(y^{(j)}) \right\rangle r_{\theta^*}(y^{(j)})\right| \leq \lambda_k, \\
& (V.1) 4\|\hat{S}_{\psi} - \Sigma_{\psi}\|_* \leq \delta_\sigma, \quad (V.2) 4\|\hat{S}_{\psi}(\theta^*) - \Sigma_{\psi}\|_* \leq \delta_\sigma \}.
\end{align*}$

**Theorem 1** Assume Conditions 1 and 2. Let $\hat{\theta}_k$ be the estimator constructed by Procedure 1 with one-step approximation as

$$\hat{\theta}_k = \hat{\theta} - \hat{\omega}_k^T \nabla \ell_{\text{KLIE}}(\hat{\theta}).$$

Suppose $\mathbb{P}(\mathcal{E}_{\text{one}}) \geq 1 - \epsilon_{\text{one,n}}$ for some $\lambda_\theta, \lambda_k, \delta_\theta, \delta_k, \delta_\sigma \in [0,1)$. Then,

$$\sup_{t \in \mathbb{R}} \mathbb{P}\left\{ \sqrt{n} (\hat{\theta}_k - \theta^*_k) / \hat{\sigma}_k \leq t \right\} - \Phi(t) \leq \Delta_1 + \Delta_2 + \Delta_3 + \epsilon_{\text{one,n}},$$

where

$$\Delta_1 \lesssim \sqrt{\frac{\tilde{r}^2 / \kappa}{n}} \frac{\|\omega^*_k\|}{\sqrt{n}}, \quad \Delta_2 \lesssim \sqrt{n_{x,n} n_{y,n} / \kappa} \|\omega^*_k\| \sqrt{n},$$

$$\Delta_3 \lesssim \sqrt{\frac{\tilde{r}^2 / \kappa}{n}} \|\omega^*_k\|^2 (\delta_\sigma + \delta_\theta) + \delta_k^2.$$
The proof is in Supplement B.2. We highlight some of the main technical difficulties. To prove Theorem 1, we need to find a linear approximation of $\sqrt{n} \left( \hat{\theta}_k - \theta^*_k \right)$ that is easy to analyse. This is not so obvious due to non-linearity of $\ell_1$-KLIEP. Our results require a delicate control of the bias that arises from using the empirical density ratio estimates, as we need to make sure that the error terms are vanishing even after $\sqrt{n}$ scaling. This is in contrast to Liu et al. (2017) or Fazayeli and Banerjee (2016) where it sufficed to control the gradient in the dual norm.

Theorem 1 gives a result for a general norm penalty in Procedure 1. When specialized to SparKLIE+1, we have the following result.

**Theorem 2** Assume Condition 1 with the $\ell_1$-norm and Condition 2. Let $\hat{\theta}_k$ be the SparKLIE+1 estimator with tuning parameters

$$
\lambda_\theta \asymp \left( \frac{\log p}{n} \right)^{1/2} \quad \text{and} \quad \lambda_k \asymp s^{-1/(2-q_k)} \left( \frac{\log p}{n} \right)^{1/2}.
$$

Let $s$ be a sequence of integers satisfying $s \geq s_{\theta,0}, s_{k,q_k}^2 - q_k$. Let $\epsilon_{\text{RSC},n}$ be a sequence in $(0, 1)$ decreasing to 0. Then, subject to additional conditions on $n_y$ and the growth regime detailed in Supplement C.1,

$$
\sup_{t \in \mathbb{R}} \left| \mathbb{P} \left\{ \sqrt{n} \left( \hat{\theta}_k - \theta^*_k \right) / \delta_k \leq t \right\} - \Phi(t) \right| 
\leq O \left( s_{\theta,0}^2 s_{k,q_k}^2 - q_k \left( \frac{\log p}{n} \right)^{1-q_k} \sqrt{n} \right) + \epsilon_{\text{RSC},n} + c \exp \left( - c' \log p \right),
$$

where $c, c' > 0$ are constants that do not depend on $n, p, s_{\theta,0}$ or $s_{k,q_k}$.

The proof in Supplement C.1 relies on numerous technical lemmas to derive the rates of $\hat{\theta}$ and $\hat{\theta}_k$. In particular, we prove a restricted strong convexity (RSC) of the Hessian starting from a population-level assumption (Condition 2). The proof is quite involved as the Hessian is a weighted sample covariance where the weights are given by the empirical density ratio estimates, which makes application of existing results impossible. The details are in Supplement E.3.

**Remark 3** Theorem 2 gives a non-trivial bound only for sufficiently (weakly) sparse $\theta^*$ and $\omega^*_k$.

The additional condition on $n_y$ is a consequence of proving RSC from the population-level assumptions. In particular, it is linked to the probability that the Hessian fails to satisfy RSC. Analogous results for other sparsity regimes can be obtained from Theorem 1 as well (see earlier version of this paper on arXiv). Due to space limitations, we have singled out this regime as being arguably the most interesting.

**Remark 4** We note that the inverse of the Hessian $\Sigma^{-1}_\psi$ is determined by $\gamma_x$, since $\Sigma_\psi = \text{Cov}_X \{ \psi(\mathbf{x}) \}$, and, therefore, the sparsity of $\Sigma^{-1}_\psi$ is related to that of $\gamma_x$. In the case of Gaussian graphical models, we can explicitly characterize $\Sigma^{-1}_\psi$ and we observe that the rows of the inverse of the Hessian are sparse if the maximum degree of the underlying graph is small. The proof strategy critically relies on the properties of a Gaussian distribution and its log-partition function, however, and is intractable for general Markov random fields. Thus, we instead provide numerical evidence on the relationship between the support of $\Sigma^{-1}_\psi$ and that of $\gamma_x$ for Ising models. For our method to perform well, it suffices that the $\ell_q$-“norm” is controlled.
for a small $q \in [0, 1)$, which we numerically verify. See Supplement D.3. Finally, we note that in some cases the rows of $\Sigma_{\psi}^{-1}$ are neither sparse nor approximately sparse, but have bounded $\ell_1^r$ norm. In this case, a possible direction for developing a valid inference procedure would be to modify the three-step procedure in Ma et al. (2017) or Yu et al. (2020).

**Remark 5** As pointed out by a reviewer, there is an inherent asymmetry in KLIEP, and Theorem 2 is one place where this can be observed. Specifically, the quality of Gaussian approximation depends on which set of observations is used as $\mathbf{X}$ and which as $\mathbf{Y}$. First, $r_\theta$ may be more regular than $1/r_\theta$ as measured by the bounds. This affects the magnitude of $\lambda_\theta$ or $\lambda_k$. Second, the larger sample will satisfy the sample complexity condition with a smaller $\varepsilon_{RSC,n}$, which is the probability that the Hessian fails to satisfy RSC. For the bounded sufficient statistics model we consider, we have found the latter to have a larger impact on the results. Therefore, we recommend choosing $f_x$ and $f_y$ so that $n_x \leq n_y$. In Section 7, we discuss alternative approaches to differential network estimation that are not asymmetric in nature. These, however, require imposing stronger conditions.

### 4.3 Finite-sample consistency for Gaussian multiplier bootstrap sketched quantiles

Theorem 3 is a finite-sample consistency result for the Gaussian multiplier bootstrap. Recall $T = \max_k \sqrt{n} | \hat{\theta}_k - \theta_k^\ast |$, and let $\hat{c}_{T,n}$ denote the estimator of $(1 - \alpha)$-quantile of $T$ from Procedure 2. Define $\Sigma_{\text{pooled}}$ analogously as in Equation (12), and let $\Omega^\ast = \Sigma_{\psi}^{-1}$. Recall that the $k$th column of $\Omega^\ast$ is $\omega_k^\ast$. For $\lambda_\theta, (\lambda_k)_{k=1}^p, (\delta_\theta, (\delta_k)_{k=1}^p \in [0, 1)$, define an event

$$
E_{\text{all}} = E_{\text{all}}(\lambda_\theta, (\lambda_k)_{k=1}^p, (\delta_\theta, (\delta_k)_{k=1}^p) = \begin{cases} 
\text{(G.1)} & 2 \| \nabla \varepsilon_{\text{KLIEP}}^{(\theta^\ast)} \| \leq \lambda_\theta, \\
\text{(E.1)} & 2 \| \hat{\theta} - \theta^\ast \| \leq \delta_\theta, \\
\text{(B.1)} & 1 - \frac{2}{\lambda_\theta}, \\
\text{(E.2)} & \| \hat{\omega}_k - \omega_k^\ast \| \leq \delta_k \forall k, \\
\text{(G.2)} & 2 \| \nabla^2 \varepsilon_{\text{KLIEP}}^{(\theta^\ast)} \| \leq \lambda_k \forall k, \\
\text{(B.2)} & \frac{1}{n_y} \sum_{j=1}^{n_y} (\omega_k^\ast, \mu_{\psi}(y^{(j)})) r_{\theta^\ast}(y^{(j)}) \leq \lambda_k \forall k 
\end{cases}.
$$

Put $\nu_n = 1 \lor \max\{ \| \omega_k^\ast \| : k = 1, \ldots, p \}$, and set

$$B_n = \frac{(1 \lor \nu_n)^3(1 \lor M_{\psi})^3 M_{\psi}^3 \nu_n^{\frac{21}{2}}}{\sqrt{k^3 n X Y}} \quad \text{and} \quad \delta_n = \left( \frac{B_n^2 \log^7 (pn)}{n} \right)^{1/6}.
$$

**Theorem 3** Assume Conditions 1 and 2. Let $\hat{\theta}$ be the estimator constructed by Procedure 1 with one-step approximation as

$$\hat{\theta} = \hat{\theta} - \hat{\Omega}^T \nabla \varepsilon_{\text{KLIEP}}(\hat{\theta}),$$

where $\hat{\Omega} = [\hat{\omega}_k]_{k=1}^p \in \mathbb{R}^{p \times p}$ is the matrix with the $k$th column given by $\hat{\omega}_k$. Suppose
\[
D_1 := \max_k \sqrt{\frac{\eta_{x,n} \eta_{y,n}}{\kappa/\kappa^2}} \left( (\delta_\theta + \lambda_\theta)(\delta_k + \lambda_k) + \|\omega_k^*\| \delta_\theta^2 \right) \sqrt{n} \leq \left( \frac{B^2 \log^4(pn)}{n^3} \right)^{1/6},
\]
\[
D_2 := \max_k \frac{\kappa/\kappa^2}{\eta_{x,n}^2 \eta_{y,n}^2} \left( \delta_k^2 \eta_{y,n}^2 \|\omega_k^*\| \eta_{x,n}^2 \|\omega_k^*\|^2 \right) \leq \left( \frac{B^2 \log^4(pn)}{n^3} \right)^{1/6}.
\]

If \( \mathbb{P}(\mathcal{E}_{\text{all}}) \geq 1 - \epsilon_{\text{all},n} \), then
\[
\sup_{a \in (0,1)} \left| \mathbb{P} \left\{ T \leq \hat{c}_{T,1-a} \right\} - (1-a) \right| = O(\delta_n + \epsilon_{\text{all},n})
\]
with probability at least \( 1 - \epsilon_{\text{all},n} - n^{-1} \).

The proof is in Supplement B.3. The bulk of hard work was done in establishing a linear approximation to \( \sqrt{n}(\hat{\theta}_k - \theta^*_k) \) in the proof of Theorem 1. Theorem 3 follows by showing that the error in the linear approximation can be controlled, allowing for application of results in Belloni et al. (2018). Due to the non-linearity of \( \ell'_\text{KLIEP} \) (3) and the fact that we are using a two sample estimator, the detailed calculations are rather complicated.

As an application of Theorem 3, we evaluate the bound in Equation (16) in the case of SparKLIE+1 with \( s_\theta = s_{\theta,0} = \|\theta^*\|_0 \) and \( s_k = s_{k,0} = \|\omega_k^*\|_0 \).

**Theorem 4** Assume Condition 1 with \( \ell'_1 \)-norm and Condition 2. Suppose \( T = \max_k \sqrt{n} |\hat{\theta}_k - \theta^*_k| \), where \( \hat{\theta} \) is the SparKLIE+1 estimator with tuning parameters
\[
\lambda_\theta \asymp \left( \frac{\log p}{n} \right)^{1/2} \quad \text{and} \quad \lambda_k \asymp \left( \frac{s_{k,0} \log p}{n} \right)^{1/2}, \quad k = 1, \ldots, p.
\]

Let \( s \) be a sequence of integers satisfying \( s \geq s_{\theta,0}, s_{k,0}, k = 1, \ldots, p \). Let \( \epsilon_{\text{RSC},n} \) be a sequence in \((0,1)\) decreasing to 0. Then, subject to an additional condition on \( n_X \), detailed in Supplement C.2,
\[
\sup_{a \in (0,1)} \left| \mathbb{P} \left\{ T \leq \hat{c}_{T,1-a} \right\} - (1-a) \right| = O(\delta_n + \epsilon_{\text{RSC},n} + c \exp \left(-c' \log p \right))
\]
with probability at least \( 1 - \epsilon_{\text{RSC},n} - c \exp \left(-c' \log p \right) - n^{-1} \), where \( c, c' > 0 \) are constants that do not depend on \( n, p, s_{\theta,0} \) or \( s_{k,0} \).

5 | SIMULATION STUDIES

Through extensive simulations, we illustrate the finite-sample performance of our methods: SparKLIE+ (Section 5.1) and empirical bootstrap sketching (Section 5.2).

5.1 | Inference for a single edge via Gaussian approximation

In Experiments 1 and 2, we look at the performance of statistical inference procedures based on Gaussian approximation when an edge has been fixed as a target of inferential interest.
**Experiment 1.** We check the coverage of the 95% CI \( \hat{\theta}_k \pm z_{0.975} \hat{\sigma}_k / \sqrt{n} \), where \( k \) is a fixed edge of interest and \( z_{0.975} \) is the 0.975-quantile of \( \mathcal{N}(0, 1) \). Here, SparKLIE+1 and +2 are compared with two other procedures: an oracle procedure with the knowledge of \( \text{supp}(\theta^*) \) and a naïve re-estimation procedure that re-fits the model based on the estimated support \( \text{supp}(\hat{\theta}) \), where \( \hat{\theta} \) is a sparse KLIEP estimate. See Supplement H.1 for precise definitions.

The results were obtained using Procedure 1 with Procedures 4 and 5 in Supplement G.1 for Steps 1 and 2, respectively, and with the universal penalty levels, as explained in Remark 1 in Section 3.1. However, we remark that even with the vanilla sparse KLIEP procedure (4) in Step 1, we have found the performance of Procedure 1 to be robust to the choice of \( \lambda_0 \). See Remark 2, as well as Supplement I.3.

The data are pairs of samples of i.i.d. observations from a pair of Ising models \( \gamma_x \) and \( \gamma_y \). Eight pairs of \( \gamma_x \) and \( \gamma_y \) are compared, arising from all possible combinations of the number of nodes (\( m = 25 \) or 50), the topology of \( \gamma_x \) (a chain or a ternary tree), and two choices of \( \theta^* \) from which \( \gamma_y = \gamma_x - \theta^* \) is obtained. Each differential network has five non-zero edges, one of which has been fixed as the target of inference. For illustration, see Figures 4–7 in Supplement H.2.

Table 1 gives the proportions of successful coverage out of 1000 independent replications at the nominal confidence level of 95%. In spite of the small sample sizes, the coverages of 95% CIs based on either of the two SparKLIE+ estimators are close to the nominal level, and on par with the performance of the oracle procedure across all the data generating processes considered. By contrast, we see that the naïve re-fitted estimator can undercover by as much as \( \approx 13\% \).

In Supplement H.4, we further provide normal Q–Q plots (Figures 8–11) and empirical estimates of the biases (Table 3) for the four estimators. These reveal that the inferior performance of the naïve re-fitted estimator can be attributed to the larger bias.

In Experiment 2 in Supplement I.1, we study the power of SparKLIE+1 and +2 for testing the null hypothesis \( H_0: \theta_k^* = 0 \), where \( k \) is a fixed edge of interest.

| \( \gamma_x \) | \( \gamma_y \) | \( m \) | \( n_x \) | \( n_y \) | Oracle | Naïve | SparKLIE+1 | +2 |
|-----------------|-----------------|-----|-----|-----|------|------|----------|---|
| Chain           | 1               | 25  | 150 | 300 | 0.960 | 0.850 | 0.934    | 0.945 |
|                 |                 | 50  | 300 | 600 | 0.946 | 0.822 | 0.943    | 0.948 |
|                 | 2               | 25  | 150 | 300 | 0.962 | 0.907 | 0.948    | 0.948 |
|                 |                 | 50  | 300 | 600 | 0.962 | 0.839 | 0.953    | 0.955 |
| Ternary tree    | 1               | 25  | 150 | 300 | 0.972 | 0.925 | 0.932    | 0.958 |
|                 |                 | 50  | 300 | 600 | 0.976 | 0.874 | 0.973    | 0.979 |
|                 | 2               | 25  | 150 | 300 | 0.972 | 0.946 | 0.957    | 0.977 |
|                 |                 | 50  | 300 | 600 | 0.968 | 0.913 | 0.952    | 0.977 |

*Note:* The results are averages over 1000 independent replications.
5.2 | Global inference with empirical bootstrap quantile estimates

In Experiments 3 and 4, we look at the performance of Procedure 3 for making inferences about the entire differential network $\theta^*$.

**Experiment 3.** We check that Procedure 3 produces consistent estimates of the quantiles $c_{T,1-a}$ of $T = \max_k \sqrt{n} \left| \hat{\theta}_k - \theta^* \right|$. Here, we focus on the setting $\gamma = \gamma_x = \gamma_y$, that is, $\theta^* = 0$. We generate a pair of samples of the same size $n_x = n_y = 500$ from the same Ising model with the parameter $\gamma$. The parameter $\gamma$ was generated as a disjoint union of $m/5$ chains of length 5 for $m \in \{25, 50, 100\}$. The non-zero edge weights were drawn i.i.d. from one of the three distributions: sign = 1, Unif(0.2, 0.4); sign = -1, Unif(-0.4, -0.2); or sign = 0, Unif(-0.4, -0.2) $\cup$ (0.2, 0.4).

For each draw of samples from $\gamma_x$ and $\gamma_y$, we use Procedure 3 with $n_b = 1000$ bootstrap replicates to estimate $\hat{c}_{T,1-a}$, and record $\{T \leq \hat{c}_{T,1-a}\}$ for each $1 - a = 0.05, \ldots, 0.95$. Then, the results are averaged across 1000 independent draws of the pair of samples. If Procedure 3 is consistent, $\{T \leq \hat{c}_{T,1-a}\} \approx \{T \leq c_{T,1-a}\}$, and hence the average over independent replicates would be close to $1 - a$. This is indeed what we see in Figure 1.

In *Experiment 4* in Supplement I.2, we study the power of the level-$\alpha$ test obtained by inverting the simultaneous confidence region $\hat{\theta}_k \pm \sqrt{n} \sqrt{\hat{c}_{T,1-a}}$ for testing the null hypothesis $H_0: \theta_k^* = 0$ for all $k$.

6 | REAL DATA EXAMPLE: ALERTNESS AND MOTOR CONTROL, AN fMRI STUDY

We apply Procedures 1 and 3 to analyse a new fMRI data set, made available courtesy of Dr. Jade Thai and Dr. Christelle Langley at the University of Bristol. The data set comes from a pilot study involving a multiple sclerosis (MS) subject and a healthy control (HC) with the purpose of exploring the relationship between alertness and motor control. It consists of two time series, one for the healthy control and one for the MS subject. The measurements were taken while the participants were performing one of three types of tasks: a sensorimotor task (T1), an intrinsic alertness task (T2), and an extrinsic alertness task (T3). For details concerning the study design and data post-processing, see Supplement J.

We model the fMRI measurements as independent observations from six Gaussian graphical models, where the groups are given by the disease status and the task type. For example, the measurements collected while the HC subject performed T1 are modelled as

$$ f_{HC,T1}(x) = \det(G_{HC,T1}/(2\pi))^{1/2} \exp\left\{ -(x - \mu_{HC,T1})^\top G_{HC,T1}(x - \mu_{HC,T1})/2 \right\}. $$

Since we are interested in the difference in the graph structure, we work with the data after centering by the group means. The sample sizes are given in Table 2.

For either the HC or the MS subject, we study the pairwise differences for the tasks. Specifically, while simultaneously controlling the type I error rate at $\alpha = 0.05$, we would like to learn the structure of six differential networks:

$$ \Delta_1^* = G_{HC, T1} - G_{HC, T2}, \quad \Delta_2^* = G_{HC, T1} - G_{HC, T3}, \quad \Delta_3^* = G_{HC, T2} - G_{HC, T3}, $$

$$ \Delta_4^* = G_{MS, T1} - G_{MS, T2}, \quad \Delta_5^* = G_{MS, T1} - G_{MS, T3}, \quad \Delta_6^* = G_{MS, T2} - G_{MS, T3}. $$
This is naturally a multiple comparisons problem well-suited to Procedure 3. The six differential networks $\hat{\Delta}_g$, $g = 1, \ldots, 6$, were estimated using Procedure 1 with Procedures 4 and 5 in Supplement G.1.
for Steps 1 and 2, respectively, and with the universal penalty levels, as explained in Remark 1 in Section 3.1. The test statistic \( T_0 = \max_{g=1,\ldots,6} \max_{1 \leq u \leq v \leq 25} |\hat{\Delta}_{g,uv}| \) was used to test the null hypothesis \( H_0: \Delta_g^* = 0 \) for all \( g = 1, \ldots, 6 \) at level 0.05 based on the estimate \( \hat{c}_{T_0,0.95} \) obtained from Procedure 3. The test found no edges to be statistically significant. However, the conclusion is based on a pilot study from two individuals, and more data are needed.

7 | DISCUSSION

We have developed novel methods for making statistically valid comparisons of general Markov networks based on i.i.d. observations from each. To our knowledge, this is the first work that allows one to conduct provably valid inference using a direct estimate of the network difference for general Markov networks in high-dimensional settings. This means that our methods can deal with dense networks as long as their difference is sparse. Also, our framework can easily handle non-Gaussian data. Furthermore, our theory does not require the conditions that are typically necessary to guarantee consistent support recovery, increasing applicability of our proposal. In addition, we develop the bootstrap sketching procedures to estimate the quantiles of extreme statistics accurately and in a computationally efficient manner even at large \( p \).

As remarked by a reviewer, it is natural to ask whether it is possible to use other divergences to derive similar procedures. For closely related varieties, such as the reverse and the symmetric KL, the answer is clearly yes. For arbitrary divergences, however, exact analogues may not exist. The derivation of KLIEP uses more than just the properties of a divergence. Indeed, the logarithm in KL plays an essential role in linearizing the ratio \( f_x / (r_{\theta} f_y) \), yielding a population-level loss that involves expectations of only known functions of \( \theta \). In addition the loss is convex in \( \theta \), leading to a computationally attractive procedure. Using other divergences to measure discrepancy between \( f_x \) and \( r_{\theta} f_y \) would, to the best of our knowledge, lead to an estimator that is not convex in \( \theta \). Establishing statistical properties of such an estimator is beyond the scope of this paper.

It can be checked that the special case of the reverse KL reduces to KLIEP with the role of \( f_x \) and \( f_y \) swapped. The effect of switching the samples was discussed in Remark 5 in Section 4.2. The symmetric KL leads to a procedure that minimizes the sum of the KLIEP and the reversed KLIEP loss functions. The theory developed in this paper extends in an obvious way to the symmetrized procedure. This means that the conditions that were previously imposed on only one of \( f_x \) and \( f_y \) now need to hold for both, reducing the applicability of our methods. Moreover, although the change is not expected to alter the order of error bounds, the constants are expected to be larger, and this is likely to result in a more brittle approximation at the same sample sizes, as corroborated by empirical evidence (Supplement I.3).

In addition to the approach followed in this paper, where the density ratio is estimated by minimizing the divergence between one density and the product of the density ratio and another density, alternative approaches have been considered in the literature. For example, Nguyen et al. (2010) estimate the density ratio by maximizing a lower bound on an \( f \)-divergence. Kanamori et al. (2009) estimate a density ratio by minimizing a squared loss between the true density ratio and the model of a density ratio. Developing inferential results for the parameters of differential networks obtained by such approaches is an interesting topic for future research.

Although we never place explicit assumptions on the form of dependence, some restraint is necessary in practice for good performance. This can already be seen from the results in Section
4: the bounds deteriorate rapidly as \( p \) increases to accommodate higher-order dependencies. This is why we chose to focus exclusively on pairwise models in our simulations and real data analysis. It is of future interest to develop an efficient search procedure to include only the relevant higher-order terms.

Finally, although it is a huge advantage of our methods that they can be used to compare general Markov networks, it may be possible to obtain more sample efficient procedures for particular models by utilizing distribution-specific properties. For example, it is of interest to develop inferential procedures for the network difference of Gaussian or Gaussian copula models.

**ACKNOWLEDGEMENTS**

Song Liu was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was completed in part with resources provided by the University of Chicago Research Computing Center. The fMRI data set collection is funded by MS Research Treatment and Education whose Registered Charity Number is 1043280.

**ORCID**

Byol Kim  [http://orcid.org/0000-0002-8993-4509](http://orcid.org/0000-0002-8993-4509)

Song Liu  [https://orcid.org/0000-0003-1414-9095](https://orcid.org/0000-0003-1414-9095)

Mladen Kolar  [https://orcid.org/0000-0001-7353-3404](https://orcid.org/0000-0001-7353-3404)

**REFERENCES**

Banerjee, O., El Ghaoui, L. & d’Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation. *The Journal of Machine Learning Research*, 9(3), 485–516.

Barber, R.F. & Kolar, M. (2018) Rocket: Robust confidence intervals via kendall’s tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B), 3422–3450.

Belilovsky, E., Varoquaux, G. & Blaschko, M.B. (2016) Testing for differences in gaussian graphical models: applications to brain connectivity. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I. & Garnett, R. (Eds.) *Advances in neural information processing systems* 29. Red Hook: Curran Associates Inc, pp. 595–603.

Belloni, A. & Chernozhukov, V. (2013) Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C. & Kato, K. (2018) High-dimensional econometrics and regularized GMM. arxiv: 1806.01888.

Belloni, A., Chernozhukov, V. & Hansen, C.B. (2013) Inference on treatment effects after selection amongst high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.

Belloni, A., Chernozhukov, V. & Wang, L. (2014) Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics*, 42(2), 757–788.

Belloni, A., Chernozhukov, V. & Wei, Y. (2016) Post-selection inference for generalized linear models with many controls. *The Journal of Business and Economic Statistics*, 34(4), 606–619.

Cai, T.T., Li, H., Ma, J. & Xia, Y. (2019) Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika*, 106(2), 401–416.

Cai, T.T., Liu, W. & Luo, X. (2011) A constrained \( \ell_1 \) minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494), 594–607.

Chen, X. (2018) Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *The Annals of Statistics*, 46(2), 642–678.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2013) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6), 2786–2819.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2015) Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162(1–2), 47–70.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2017) Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4), 2309–2352.
Chernozhukov, V., Hansen, C. & Spindler, M. (2015) Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics, 7*(1), 649–688.

Chiquet, J., Grandvalet, Y. & Ambroise, C. (2011) Inferring multiple graphical structures. *Statistics and Computing, 21*(4), 537–553.

Danaher, P., Wang, P. & Witten, D.M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(2), 373–397.

Deng, H. & Zhang, C.-H. (2020) Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *The Annals of Statistics, 48*(6), 3643–3671.

Dezeure, R., Bühlmann, P. & Zhang, C.-H. (2017) High-dimensional simultaneous inference with the bootstrap. *TEST, 26*(4), 685–719.

Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G. & West, M. (2004) Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis, 90*(1), 196–212.

Drton, M. & Maathuis, M.H. (2017) Structure learning in graphical modeling. *Annual Review of Statistics and its Application, 4*(1), 365–393.

Fazayeli, F. & Banerjee, A. (2016) Generalized direct change estimation in Ising model structure. In: Balcan, M.F. & Weinberger, K.Q. (Eds.) *Proceedings of the 33rd international conference on machine learning*, Volume 48 of *Proceedings of machine learning research*, New York, New York, USA. pp. 2281–2290.

Friedman, J.H., Hastie, T.J. & Tibshirani, R.J. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441.

Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2011) Joint estimation of multiple graphical models. *Biometrika, 98*(1), 1–15.

Hammersley, J.M. & Clifford, P. (1971) Markov fields on finite graphs and lattices. *Unpublished manuscript*, 46.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In: Altman, R.B., Dunker, A.K., Hunter, L. & Klein, T.E. (Eds.) *Proceedings of the 6th pacific symposium on biocomputing, PSB 2001, Hawaii, USA, January 3–7, 2001*, pp. 422–433.

Janková, J. & van de Geer, S. (2015) Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics, 9*(1), 1205–1229.

Janková, J. & van de Geer, S.A. (2017) Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST, 26*(1), 143–162.

Javanmard, A. & Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research, 15*(1), 2869–2909.

Kanamori, T., Hido, S. & Sugiyama, M. (2009) A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research, 10*, 1391–1445.

Koller, D. & Friedman, N. (2009) *Probabilistic graphical models: principles and techniques*. Cambridge, MA: MIT Press.

Lauritzen, S.L. (1996) *Graphical models*. New York: The Clarendon Press, Oxford University Press.

Liu, W. (2017) Structural similarity and difference testing on multiple sparse Gaussian graphical models. *The Annals of Statistics, 45*(6), 2680–2707.

Liu, S., Quinn, J.A., Gutmann, M.U., Suzuki, T. & Sugiyama, M. (2014) Direct learning of sparse changes in Markov networks by density ratio estimation. *Neural Computation, 26*(6), 1169–1197.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M. & Fukumizu, K. (2017) Support consistency of direct sparse-change learning in Markov networks. *The Annals of Statistics, 45*(3), 959–990.

Lu, J., Kolar, M. & Liu, H. (2018) Post-regularization inference for time-varying nonparanormal graphical models. *Journal of Machine Learning Research, 18*(203), 1–78.

MacKay, D.J.C. (2003) *Information theory, inference and learning algorithms*. New York: Cambridge University Press.

Majumdar, S. & Michailidis, G. (2018) Joint estimation and inference for data integration problems based on multiple multi-layered gaussian graphical models. arxiv: 1803.03348.

Ma, C., Lu, J. & Liu, H. (2017) Inter-subject analysis: inferring sparse interactions with dense intra-graphs.

Ma, J. & Michailidis, G. (2016) Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research, 17*(166), 1–48.
Meinshausen, N. (2015) Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5), 923–945.

Mohan, K., London, P., Fazel, M., Witten, D.M. & Lee, S.-I. (2014) Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15, 445–488.

Nguyen, X., Wainwright, M.J. & Jordan, M.I. (2010) Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11), 5847–5861.

Ning, Y. & Liu, H. (2017) A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1), 158–195.

Ravikumar, P., Wainwright, M.J., Raskutti, G. & Yu, B. (2011) High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935–980.

Ren, Z., Sun, T., Zhang, C.-H. & Zhou, H.H. (2015) Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3), 991–1026.

Shojaie, A. (2021) Differential network analysis: a statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2), e1508 16.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P.V. & Kawanabe, M. (2008) Direct importance estimation with model selection and its application to covariate shift adaptation. In: Platt, J.C., Koller, D., Singer, Y. & Roweis, S.T. (Eds.) *Advances in neural information processing systems 20*. Red Hook: Curran Associates Inc, pp. 1433–1440.

Sugiyama, M., Suzuki, T. & Kanamori, T. (2012) *Density ratio estimation in machine learning*. Cambridge: Cambridge University Press With a foreword by Thomas G. Dietterich.

Sun, T. & Zhang, C.-H. (2013) Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14, 3385–3418.

Supekar, K., Menon, V., Rubin, D., Musen, M. & Greicius, M.D. (2008) Network analysis of intrinsic functional brain connectivity in alzheimer’s disease. *PLoS Computational Biology*, 4(6), e1000100.

van de Geer, S.A., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.

van der Vaart, A.W. (1998) *Asymptotic statistics*. Cambridge: Cambridge University Press vol. 3 of *Cambridge series in statistical and probabilistic mathematics*.

Wainwright, M.J. & Jordan, M.I. (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2), 1–305.

Wang, J. & Kolar, M. (2016) Inference for high-dimensional exponential family graphical models. In: Gretton, A. & Robert, C.C. (Eds.) *Proceedings of AISTATS*, vol. 51. pp. 751–760.

Xia, Y., Cai, T. & Cai, T.T. (2015) Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2), 247–266.

Xue, K. & Yao, F. (2020) Distribution and correlation-free two-sample test of high-dimensional means. *The Annals of Statistics*, 48(3), 1304–1328.

Xu, P. & Gu, Q. (2016) Semiparametric differential graph models. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I. & Garnett, R. (Eds.) *Advances in neural information processing systems 29*. Red Hook: Curran Associates Inc, pp. 1064–1072.

Yang, E., Ravikumar, P., Allen, G.I. & Liu, Z. (2015) On graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16, 3813–3847.

Yuan, M. (2010) High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11, 2261–2286.

Zhang, C.-H. & Zhang, S.S. (2013) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 217–242.

Zhao, S.D., Cai, T.T. & Li, H. (2014) Direct estimation of differential networks. *Biometrika*, 101(2), 253–268.
SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kim B, Liu S, Kolar M. Two-sample inference for high-dimensional Markov networks. *J R Stat Soc Series B*. 2021;83:939–962. [https://doi.org/10.1111/rssb.12446](https://doi.org/10.1111/rssb.12446)