LETTER **Communicated by Yongjia Song**

Direct Learning of Sparse Changes in Markov Networks by Density Ratio Estimation

Song Liu
song@sg.cs.titech.ac.jp
Tokyo Institute of Technology, Meguro, Tokyo 152-8552, Japan

John A. Quinn
jquinn@cit.ac.ug
Makerere University, Kampala, Uganda

Michael U. Gutmann
michael.gutmann@helsinki.fi
University of Helsinki, Finland, FI-00014, Finland

Taiji Suzuki
s-taiji@is.titech.ac.jp
Masashi Sugiyama
sugi@cs.titech.ac.jp
Tokyo Institute of Technology, Meguro, Tokyo 152-8552, Japan

We propose a new method for detecting changes in Markov network structure between two sets of samples. Instead of naively fitting two Markov network models separately to the two data sets and figuring out their difference, we directly learn the network structure change by estimating the ratio of Markov network models. This density-ratio formulation naturally allows us to introduce sparsity in the network structure change, which highly contributes to enhancing interpretability. Furthermore, computation of the normalization term, a critical bottleneck of the naive approach, can be remarkably mitigated. We also give the dual formulation of the optimization problem, which further reduces the computation cost for large-scale Markov networks. Through experiments, we demonstrate the usefulness of our method.

1 Introduction

Changes in interactions between random variables are interesting in many real-world phenomena. For example, genes may interact with each other in

An earlier version of this work was presented at the European Conference on Machine Learning and Principles, and Practice of Knowledge Discovery in Databases (ECML/PKDD2013), September 23–27, 2013.

Neural Computation **26, 1169–1197** (2014)
doi:10.1162/NECO_a_00589

different ways when external stimuli change, co-occurrence between words may appear or disappear when the domains of text corpora shift, and correlation among pixels may change when a surveillance camera captures anomalous activities. Discovering such changes in interactions is a task of great interest in machine learning and data mining because it provides useful insights into underlying mechanisms in many real-world applications.

In this letter, we consider the problem of detecting changes in conditional independence among random variables between two sets of data. Such conditional independence structure can be expressed using an undirected graphical model called a Markov network (MN) (Bishop, 2006; Wainwright & Jordan, 2008; Koller & Friedman, 2009), where nodes and edges represent variables and their conditional dependencies, respectively. As a simple and widely applicable case, the pairwise MN model has been thoroughly studied (Ravikumar, Wainwright, & Lafferty, 2010; Lee, Ganapathi, & Koller, 2007). Following this line, we also focus on the pairwise MN model as a representative example.

A naive approach to change detection in MNs is the two-step procedure of first estimating two MNs separately from two sets of data by maximum likelihood estimation (MLE) and then comparing the structure of the learned MNs. However, MLE is often computationally intractable due to the normalization factor included in the density model. Therefore, gaussianity is often assumed in practice for computing the normalization factor analytically (Hastie, Tibshirani, & Friedman, 2001), though this gaussian assumption is highly restrictive in practice. We may utilize importance sampling (Robert & Casella, 2005) to numerically compute the normalization factor, but an inappropriate choice of the instrumental distribution may lead to an estimate with high variance (Wasserman, 2010); for more discussions on sampling techniques (see Gelman, 1995, and Hinton, 2002). Hyvärinen (2005) and Gutmann and Hyvärinen (2012) have explored an alternative approach to avoid computing the normalization factor that is not based on MLE. However, the two-step procedure has the conceptual weakness that structure change is not directly learned. This indirect nature causes a crucial problem. Suppose that we want to learn a sparse structure change. For learning sparse changes, we may utilize ℓ_1-regularized MLE (Banerjee, El Ghaoui, & d'Aspremont, 2008; Friedman, Hastie, & Tibshirani, 2008; Lee et al., 2007), which produces sparse MNs, and thus the change between MNs also becomes sparse. However, this approach does not work if each MN is dense but only change is sparse.

To mitigate this indirect nature, the fused lasso (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005) is helpfull, where two MNs are simultaneously learned with a sparsity-inducing penalty on the difference between two MN parameters (Zhang & Wang, 2010; Danaher, Wang, & Witten, 2013). Although this fused-lasso approach allows us to learn sparse structure change naturally, the restrictive gaussian assumption is still necessary to obtain the solution in a computationally tractable way.

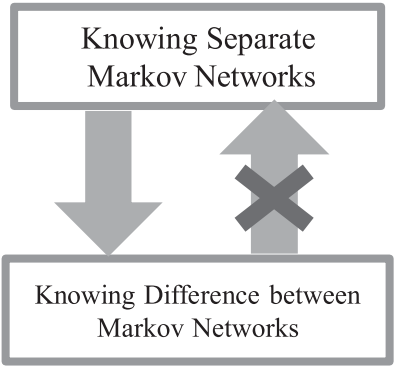


Figure 1: Rationale of direct structural change learning. Finding the difference between two MNs is a more specific task than finding the entire structures of those two networks, and hence should be possible to learn with fewer data.

The nonparanormal assumption (Liu, Lafferty, & Wasserman, 2009; Liu, Han, Yuan, Lafferty, & Wasserman, 2012) is a useful generalization of the gaussian assumption. A nonparanormal distribution is a semiparametric gaussian copula where each gaussian variable is transformed by a monotone nonlinear function. Nonparanormal distributions are much more flexible than gaussian distributions thanks to the feature-wise nonlinear transformation, while the normalization factors can still be computed analytically. Thus, the fused lasso method combined with nonparanormal models would be one of the state-of-the-art approaches to change detection in MNs. However, the method is still based on separate modeling of two MNs, and its computation for more general nongaussian distributions is challenging.

In this letter, we propose a more direct approach to structural change learning in MNs based on density ratio estimation (DRE) (Sugiyama, Suzuki, & Kanamori, 2012a). Our method does not separately model two MNs, but directly models the change in two MNs. This idea follows Vapnik's principle (Vapnik, 1998): "If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem." This principle was used in the development of support vector machines (SVMs): rather than modeling two classes of samples, an SVM directly learns a decision boundary that is sufficient for performing pattern recognition. In the current context, estimating two MNs is more general than detecting changes in MNs (see Figure 1). By directly detecting changes in MNs, we can also halve the number of parameters, from two MNs to one MN difference.

Another important advantage of our DRE-based method is that the normalization factor can be approximated efficiently, because the normalization term in a density ratio function takes the form of the expectation over a data distribution and thus can be simply approximated by the sample average without additional sampling. Through experiments on gene expression and Twitter data analysis, we demonstrate the usefulness of our proposed approach.

The remainder of this letter is structured as follows. In section 2, we formulate the problem of detecting structural changes and review currently available approaches. We propose our DRE-based structural change detection method in section 3. Results of illustrative and real-world experiments are reported in sections 4 and 5, respectively. Finally, we conclude our work and show the future direction in section 6.

2 Problem Formulation and Related Methods

In this section, we formulate the problem of change detection in Markov network structure and review existing approaches.

2.1 Problem Formulation. Consider two sets of independent samples drawn separately from two probability distributions P and Q on $\mathbb{R}^d$:

$$\{\boldsymbol{x}_i^P\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} P \quad \text{and} \quad \{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{i.i.d.}}{\sim} Q.$$

We assume that P and Q belong to the family of Markov networks (MNs) consisting of univariate and bivariate factors.[1] That is, their respective probability densities p and q are expressed as

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{u,v=1,u\geq v}^{d} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{f}\left(x^{(u)}, x^{(v)}\right)\right), \tag{2.1}$$

where $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^{\top}$ is the d-dimensional random variable, $\top$ denotes the transpose, $\boldsymbol{\theta}_{u,v}$ is the parameter vector for the elements $x^{(u)}$ and $x^{(v)}$, and

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_{1,1}^{\top}, \ldots, \boldsymbol{\theta}_{d,1}^{\top}, \boldsymbol{\theta}_{2,2}^{\top}, \ldots, \boldsymbol{\theta}_{d,2}^{\top}, \ldots, \boldsymbol{\theta}_{d,d}^{\top}\right)^{\top}$$

[1] Note that the proposed algorithm itself can be applied to any MNs containing more than two elements in each factor.

is the entire parameter vector. $\boldsymbol{f}(x^{(u)}, x^{(v)})$ is a bivariate vector-valued basis function. $Z(\boldsymbol{\theta})$ is the normalization factor defined as

$$Z(\boldsymbol{\theta}) = \int \exp\left(\sum_{u,v=1,u\geq v}^{d} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{f}\left(x^{(u)}, x^{(v)}\right) \right) \mathrm{d}\boldsymbol{x}.$$

$q(\boldsymbol{x}; \boldsymbol{\theta})$ is defined in the same way.

Given two densities that can be parameterized using $p(\boldsymbol{x}; \boldsymbol{\theta}^P)$ and $q(\boldsymbol{x}; \boldsymbol{\theta}^Q)$, our goal is to discover the changes in parameters from P to Q, that is, $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$.

2.2 Sparse Maximum Likelihood Estimation and Graphical Lasso. Maximum likelihood estimation (MLE) with group ℓ_1-regularization has been widely used for estimating the sparse structure of MNs (Schmidt & Murphy, 2010; Ravikumar et al., 2010; Lee et al., 2007):

$$\max_{\boldsymbol{\theta}} \left[\frac{1}{n_p} \sum_{i=1}^{n_p} \log p\left(\boldsymbol{x}_i^P; \boldsymbol{\theta}\right) - \lambda \sum_{u,v=1,u\geq v}^{d} \|\boldsymbol{\theta}_{u,v}\| \right], \tag{2.2}$$

where $\|\cdot\|$ denotes the ℓ_2-norm. As λ increases, $\|\boldsymbol{\theta}_{u,v}\|$ may drop to 0. Thus, this method favors an MN that encodes more conditional independencies among variables.

Computation of the normalization term $Z(\boldsymbol{\theta})$ in equation 2.1 is often computationally intractable when the dimensionality of $\boldsymbol{x}$ is high. To avoid this computational problem, the gaussian assumption is often imposed (Friedman, et al., 2008; Meinshausen & Bühlmann, 2006). More specifically, the following zero-mean gaussian model is used:

$$p(\boldsymbol{x}; \boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} \boldsymbol{x}^{\top} \boldsymbol{\Theta} \boldsymbol{x} \right),$$

where $\boldsymbol{\Theta}$ is the inverse covariance matrix (aka the precision matrix) and $\det(\cdot)$ denotes the determinant. Then $\boldsymbol{\Theta}$ is learned as

$$\max_{\boldsymbol{\Theta}} \left[\log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{\Theta} S^P) - \lambda \|\boldsymbol{\Theta}\|_1 \right],$$

where S^P is the sample covariance matrix of $\{\boldsymbol{x}_i^P\}_{i=1}^n$. $\|\boldsymbol{\Theta}\|_1$ is the ℓ_1-norm of $\boldsymbol{\Theta}$, the absolute sum of all elements. This formulation has been studied intensively in Banerjee et al. (2008), and a computationally efficient algorithm, the graphical lasso (Glasso) has been proposed (Friedman, et al., 2008).

Sparse changes in conditional independence structure between P and Q can be detected by comparing two MNs estimated separately using sparse

MLE. However, this approach implicitly assumes that two MNs are sparse, which is not necessarily true even if the change is sparse.

2.3 Fused-Lasso Method. To more naturally handle sparse changes in conditional independence structure between P and Q, a method based on fused-lasso (Tibshirani et al., 2005) has been developed (Zhang & Wang, 2010). This method directly sparsifies the difference between parameters.

The original method conducts feature-wise neighborhood regression (Meinshausen & Bühlmann, 2006) jointly for P and Q, which can be conceptually understood as maximizing the local conditional gaussian likelihood jointly on each feature (Ravikumar et al., 2010). A slightly more general form of the learning criterion may be summarized as

$$\max_{\boldsymbol{\theta}_s^P, \boldsymbol{\theta}_s^Q} \left[\ell_s^P\left(\boldsymbol{\theta}_s^P\right) + \ell_s^Q\left(\boldsymbol{\theta}_s^Q\right) - \lambda_1\left(\left\|\boldsymbol{\theta}_s^P\right\|_1 + \left\|\boldsymbol{\theta}_s^Q\right\|_1\right) - \lambda_2\left\|\boldsymbol{\theta}_s^P - \boldsymbol{\theta}_s^Q\right\|_1 \right],$$

where $\ell_s^P(\boldsymbol{\theta})$ is the log-conditional likelihood for the sth element $x^{(s)} \in \mathbb{R}$ given the rest $\boldsymbol{x}^{(-s)} \in \mathbb{R}^{d-1}$:

$$\ell_s^P(\boldsymbol{\theta}) = \frac{1}{n_P} \sum_{i=1}^{n_P} \log p\left(x_i^{(s)P} \middle| \boldsymbol{x}_i^{(-s)P}; \boldsymbol{\theta}\right).$$

$\ell_s^Q(\boldsymbol{\theta})$ is defined in the same way as $\ell_s^P(\boldsymbol{\theta})$.

Since the Flasso-based method directly sparsifies the change in MN structure, it can work well even when each MN is not sparse. However, using models other than gaussian is difficult because of the normalization issue described in section 2.2.

2.4 Nonparanormal Extensions. In the above methods, gaussianity is required in practice to compute the normalization factor efficiently, a highly restrictive assumption. To overcome this restriction, it has become popular to perform structure learning under the nonparanormal settings (Liu et al., 2009, 2012), where the gaussian distribution is replaced by a semiparametric gaussian copula.

A random vector $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ is said to follow a nonparanormal distribution if there exists a set of monotone and differentiable functions, $\{h_i(x)\}_{i=1}^d$, such that $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x^{(1)}), \ldots, h_d(x^{(d)}))^\top$ follows the gaussian distribution. Nonparanormal distributions are much more flexible than gaussian distributions thanks to the nonlinear transformation $\{h_i(x)\}_{i=1}^d$, while the normalization factors can still be computed in an analytical way.

However, the nonparanormal transformation is restricted to be elementwise, which is still restrictive to express complex distributions.

2.5 Maximum Likelihood Estimation for Nongaussian Models by Importance Sampling. A numerical way to obtain the MLE solution under general nongaussian distributions is importance sampling.

Suppose that we try to maximize the log likelihood:[2]

$$\begin{aligned}\ell_{\mathrm{MLE}}(\boldsymbol{\theta}) &= \frac{1}{n_P}\sum_{i=1}^{n_P}\log p\left(\boldsymbol{x}_i^P;\boldsymbol{\theta}\right)\\ &= \frac{1}{n_P}\sum_{i=1}^{n_P}\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^{\top}\boldsymbol{f}\left(x_i^{(u)P},x_i^{(v)P}\right)\\ &\quad -\log\int\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^{\top}\boldsymbol{f}\left(x^{(u)},x^{(v)}\right)\right)\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.3}$$

The key idea of importance sampling is to compute the integral by the expectation over an easy-to-sample instrumental density $p'(\boldsymbol{x})$ (e.g., gaussian) weighted according to the importance $1/p'(\boldsymbol{x})$. More specifically, when we use independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i'\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$, the last term of equation 2.3 can be approximately computed as follows:

$$\begin{aligned}&\log\int\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^{\top}\boldsymbol{f}(x^{(u)},x^{(v)})\right)\mathrm{d}\boldsymbol{x}\\ &\quad=\log\int p'(\boldsymbol{x})\frac{\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^{\top}\boldsymbol{f}(x^{(u)},x^{(v)})\right)}{p'(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\\ &\quad\approx\log\frac{1}{n'}\sum_{i=1}^{n'}\frac{\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^{\top}\boldsymbol{f}\left(x_i'^{(u)},x_i'^{(v)}\right)\right)}{p'(\boldsymbol{x}_i')}.\end{aligned}$$

We refer to this implementation of Glasso as IS-Glasso below.

However, importance sampling tends to produce an estimate with large variance if the instrumental distribution is not carefully chosen. Although it is often suggested to use a density whose shape is similar to the function to be integrated but with thicker tails as p', it is not straightforward in practice to decide which p' to choose, especially when the dimensionality of $\boldsymbol{x}$ is high (Wasserman, 2010).

[2]From here on, we simplify $\sum_{u,v=1,u\geq v}^{d}$ as $\sum_{u\geq v}$.

We can also consider an importance-sampling version of the Flasso method (which we refer to as IS-Flasso),[3]

$$\max_{\boldsymbol{\theta}^P,\boldsymbol{\theta}^Q}\left[\ell^P_{\mathrm{MLE}}(\boldsymbol{\theta}^P)+\ell^Q_{\mathrm{MLE}}(\boldsymbol{\theta}^Q)-\lambda_1(\|\boldsymbol{\theta}^P\|^2+\|\boldsymbol{\theta}^Q\|^2)-\lambda_2\sum_{u\geq v}\left\|\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v}\right\|\right],$$

where both $\ell^P_{\mathrm{MLE}}(\boldsymbol{\theta}^P)$ and $\ell^Q_{\mathrm{MLE}}(\boldsymbol{\theta}^Q)$ are approximated by importance sampling for nongaussian distributions. However, in the same way as IS-Glasso, the choice of instrumental distributions is not straightforward.

3 Direct Learning of Structural Changes via Density Ratio Estimation

The Flasso method can more naturally handle sparse changes in MNs than separate sparse MLE. However, the method is still based on separate modeling of two MNs, and its computation for general high-dimensional nongaussian distributions is challenging. In this section, we propose to directly learn structural changes based on density ratio estimation (Sugiyama et al., 2012a). Our approach does not involve separate modeling of each MN and allows us to approximate the normalization term efficiently for any distributions.

3.1 Density Ratio Formulation for Structural Change Detection. Our key idea is to consider the ratio of p and q:

$$\frac{p(\boldsymbol{x};\boldsymbol{\theta}^P)}{q(\boldsymbol{x};\boldsymbol{\theta}^Q)}\propto\exp\left(\sum_{u\geq v}\left(\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v}\right)^\top\boldsymbol{f}\left(x^{(u)},x^{(v)}\right)\right).$$

Here $\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v}$ encodes the difference between P and Q for factor $\boldsymbol{f}(x^{(u)},x^{(v)})$, that is, $\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v}$ is zero if there is no change in the factor $\boldsymbol{f}(x^{(u)},x^{(v)})$.

Once we consider the ratio of p and q, we do not have to estimate $\boldsymbol{\theta}^P_{u,v}$ and $\boldsymbol{\theta}^Q_{u,v}$; instead estimating their difference $\boldsymbol{\theta}_{u,v}=\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v}$ is sufficient for change detection:

$$r(\boldsymbol{x};\boldsymbol{\theta})=\frac{1}{N(\boldsymbol{\theta})}\exp\left(\sum_{u\geq v}\boldsymbol{\theta}^\top_{u,v}\boldsymbol{f}\left(x^{(u)},x^{(v)}\right)\right),\tag{3.1}$$

[3]For implementation simplicity, we maximize the joint likelihood of p and q instead of its feature-wise conditional likelihood. We also switch the first penalty term from ℓ_1 to ℓ_2.

where

$$N(\boldsymbol{\theta}) = \int q(\boldsymbol{x}) \exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{f}\left(x^{(u)}, x^{(v)}\right)\right) \mathrm{d}\boldsymbol{x}.$$

The normalization term $N(\boldsymbol{\theta})$ guarantees[4]

$$\int q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1.$$

Thus, in this density ratio formulation, we are no longer modeling p and q separately, but we model the change from p to q directly. This direct nature would be more suitable for change detection purposes according to Vapnik's principle that encourages avoidance of solving more general problems as an intermediate step (Vapnik, 1998). This direct formulation also allows us to halve the number of parameters from both $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ to only $\boldsymbol{\theta}$. Furthermore, the normalization factor $N(\boldsymbol{\theta})$ in the density ratio formulation can be easily approximated by the sample average over $\{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$, because $N(\boldsymbol{\theta})$ is

[4]If the model $q(\boldsymbol{x}; \boldsymbol{\theta}^Q)$ is correctly specified—there exists $\boldsymbol{\theta}^{Q*}$ such that $q(\boldsymbol{x}; \boldsymbol{\theta}^{Q*}) = q(\boldsymbol{x})$—then $N(\boldsymbol{\theta})$ can be interpreted as importance sampling of $Z(\boldsymbol{\theta}^P)$ via instrumental distribution $q(\boldsymbol{x})$. Indeed, since

$$Z(\boldsymbol{\theta}^P) = \int q(\boldsymbol{x}) \frac{\exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^{P\ \top} \boldsymbol{f}(x^{(u)}, x^{(v)})\right)}{q(\boldsymbol{x}; \boldsymbol{\theta}^{Q*})} \mathrm{d}\boldsymbol{x},$$

where $q(\boldsymbol{x}; \boldsymbol{\theta}^{Q*}) = q(\boldsymbol{x})$, we have

$$N(\boldsymbol{\theta}^P - \boldsymbol{\theta}^{Q*}) = \frac{Z(\boldsymbol{\theta}^P)}{Z(\boldsymbol{\theta}^{Q*})} = \int q(\boldsymbol{x}) \exp\left(\sum_{u \geq v} \left(\boldsymbol{\theta}_{u,v}^P - \boldsymbol{\theta}_{u,v}^{Q\ *}\right)^{\top} \boldsymbol{f}\left(x^{(u)}, x^{(v)}\right)\right) \mathrm{d}\boldsymbol{x}.$$

This is exactly the normalization term $N(\boldsymbol{\theta})$ of the ratio $p(\boldsymbol{x}; \boldsymbol{\theta}^P)/q(\boldsymbol{x}; \boldsymbol{\theta}^{Q*})$. However, we note that the density ratio estimation method we use in this letter is consistent with the optimal solution in the model even without the correct model assumption (Kanamori, Suzuki, & Sugiyama, 2010). An alternative normalization term,

$$N'(\boldsymbol{\theta}, \boldsymbol{\theta}^Q) = \int q(\boldsymbol{x}; \boldsymbol{\theta}^Q) r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x},$$

may also be considered, as in the case of MLE. However, this alternative form requires an extra parameter $\boldsymbol{\theta}^Q$, which is not our main interest.

the expectation over $q(\boldsymbol{x})$:

$$N(\boldsymbol{\theta}) \approx \frac{1}{n_Q} \sum_{i=1}^{n_Q} \exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}\left(x_i^{(u)Q}, x_i^{(v)Q}\right) \right).$$

3.2 Direct Density-Ratio Estimation. Density ratio estimation has been recently introduced to the machine learning community and has proven to be useful in a wide range of applications (Sugiyama et al., 2012a). Here, we concentrate on the density ratio estimator called the Kullback-Leibler importance estimation procedure (KLIEP) for log-linear models (Sugiyama et al., 2008; Tsuboi, Kashima, Hido, Bickel, & Sugiyama, 2009).

For a density ratio model $r(\boldsymbol{x}; \boldsymbol{\theta})$, the KLIEP method minimizes the Kullback-Leibler divergence from $p(\boldsymbol{x})$ to $\widehat{p}(\boldsymbol{x}) = q(\boldsymbol{x})r(\boldsymbol{x}; \boldsymbol{\theta})$:

$$\begin{aligned} \mathrm{KL}[p \| \widehat{p}] &= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})r(\boldsymbol{x}; \boldsymbol{\theta})} \mathrm{d}\boldsymbol{x} \\ &= \text{Const.} - \int p(\boldsymbol{x}) \log r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.2}$$

Note that our density-ratio model, equation 3.1, automatically satisfies the nonnegativity and normalization constraints:

$$r(\boldsymbol{x}; \boldsymbol{\theta}) \geq 0 \quad \text{and} \quad \int q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1.$$

In practice, we maximize the empirical approximation of the second term in equation 3.2:

$$\begin{aligned} \ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) &= \frac{1}{n_P} \sum_{i=1}^{n_P} \log r\left(\boldsymbol{x}_i^P; \boldsymbol{\theta}\right) \\ &= \frac{1}{n_P} \sum_{i=1}^{n_P} \sum_{u \geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}\left(x_i^{(u)P}, x_i^{(v)P}\right) \\ &\quad - \log\left(\frac{1}{n_Q} \sum_{i=1}^{n_Q} \exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}\left(x_i^{(u)Q}, x_i^{(v)Q}\right) \right) \right). \end{aligned}$$

Because $\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta})$ is concave with respect to $\boldsymbol{\theta}$, its global maximizer can be numerically found by standard optimization techniques such as gradient ascent or quasi-Newton methods. The gradient of ℓ_{KLIEP} with respect to $\boldsymbol{\theta}_{u,v}$

is given by

$$\nabla_{\boldsymbol{\theta}_{u,v}} \ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) = \frac{1}{n_P} \sum_{i=1}^{n_P} \boldsymbol{f}\left(\boldsymbol{x}_i^{(u)P}, \boldsymbol{x}_i^{(v)P}\right) - \frac{\frac{1}{n_Q} \sum_{i=1}^{n_Q} \exp\left(\sum_{u' \geq v'} \boldsymbol{\theta}_{u',v'}^{\top} \boldsymbol{f}\left(x_i^{(u')Q}, x_i^{(v')Q}\right)\right) \boldsymbol{f}\left(x_i^{(u)Q}, x_i^{(v)Q}\right)}{\frac{1}{n_Q} \sum_{j=1}^{n_Q} \exp\left(\sum_{u'' \geq v''} \boldsymbol{\theta}_{u'',v''}^{\top} \boldsymbol{f}(x_j^{(u'')Q}, x_j^{(v'')Q})\right)},$$

which can be computed in a straightforward manner for any feature vector $\boldsymbol{f}(x^{(u)}, x^{(v)})$.

3.3 Sparsity-Inducing Norm. To find a sparse change between P and Q, we propose regularizing the KLIEP solution with a sparsity-inducing norm $\sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\|$. Note that the MLE approach sparsifies both $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ so that the difference $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$ is also sparsified, while we directly sparsify the difference $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$; thus our method can still work well even if $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ are dense.

In practice, we may use the following elastic-net penalty (Zou & Hastie, 2005) to better control overfitting to noisy data,

$$\max_{\boldsymbol{\theta}} \left[\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) - \lambda_1 \|\boldsymbol{\theta}\|^2 - \lambda_2 \sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\| \right], \tag{3.3}$$

where $\|\boldsymbol{\theta}\|^2$ penalizes the magnitude of the entire parameter vector.

3.4 Dual Formulation for High-Dimensional Data. The solution of the optimization problem, equation 3.3, can be easily obtained by standard sparse optimization methods. However, in the case where the input dimensionality d is high (often the case in our setup), the dimensionality of parameter vector $\boldsymbol{\theta}$ is large, and thus obtaining the solution can be computationally expensive. Here, we derive a dual optimization problem (Boyd & Vandenberghe, 2004), which can be solved more efficiently for high-dimensional $\boldsymbol{\theta}$ (see Figure 2).

As detailed in appendix, the dual optimization problem is given as

$$\min_{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{n_Q})^{\top}} \sum_{i=1}^{n_Q} \alpha_i \log \alpha_i + \frac{1}{\lambda_1} \sum_{u \geq v} \max(0, \|\boldsymbol{\xi}_{u,v}\| - \lambda_2)^2$$

$$\text{subject to } \alpha_1, \ldots, \alpha_{n_Q} \geq 0 \quad \text{and} \quad \sum_{i=1}^{n_Q} \alpha_i = 1, \tag{3.4}$$

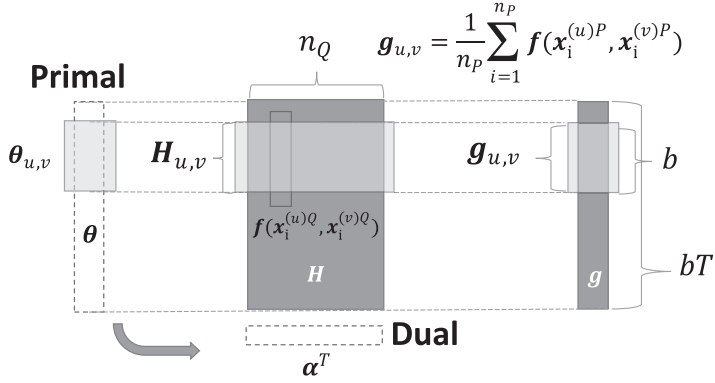


Figure 2: Schematics of primal and dual optimization. b denotes the number of basis functions and T the number of factors. Because we are considering pairwise factors, $T = \mathcal{O}(d^2)$ for input dimensionality d.

where

$$\begin{aligned}
\boldsymbol{\xi}_{u,v} &= \boldsymbol{g}_{u,v} - \boldsymbol{H}_{u,v}\boldsymbol{\alpha}, \\
\boldsymbol{H}_{u,v} &= \left[f\left(x_1^{(u)Q}, x_1^{(v)Q}\right), \ldots, f\left(x_{n_Q}^{(u)Q}, x_{n_Q}^{(v)Q}\right) \right], \\
\boldsymbol{g}_{u,v} &= \frac{1}{n_P} \sum_{i=1}^{n_P} f\left(x_i^{(u)P}, x_i^{(v)P}\right).
\end{aligned}$$

The primal solution can be obtained from the dual solution as

$$\boldsymbol{\theta}_{u,v} = \begin{cases} \dfrac{1}{\lambda_1}\left(1 - \dfrac{\lambda_2}{\|\boldsymbol{\xi}_{u,v}\|}\right)\boldsymbol{\xi}_{u,v} & \text{if} \|\boldsymbol{\xi}_{u,v}\| > \lambda_2, \\ \mathbf{0} & \text{if} \|\boldsymbol{\xi}_{u,v}\| \leq \lambda_2. \end{cases} \tag{3.5}$$

Note that the dimensionality of the dual variable $\boldsymbol{\alpha}$ is equal to n_Q, while that of $\boldsymbol{\theta}$ is quadratic with respect to the input dimensionality d because we are considering pairwise factors. Thus, if d is not small and n_Q is not very large (often the case in our experiments shown later), solving the dual optimization problem would be computationally more efficient. Furthermore, the dual objective (and its gradient) can be computed efficiently in parallel for each (u, v), a useful property when handling large-scale MNs. Note that the dual objective is differentiable everywhere, while the primal objective is not.

4 Numerical Experiments

In this section, we compare the performance of the proposed KLIEP-based method, the Flasso method, and the Glasso method for gaussian models, nonparanormal models, and nongaussian models. Results are reported on data sets with three different underlying distributions: multivariate gaussian, nonparanormal, and nongaussian "diamond" distributions. We also investigate the computation time of the primal and dual formulations as a function of the input dimensionality. The Matlab implementation of the primal and dual methods are available online at http://sugiyama-www.cs.titech.ac.jp/∼song/SCD.html.

4.1 Gaussian Distribution. First, we investigate the performance of each method under gaussianity. Consider a 40-node sparse gaussian MN, where its graphical structure is characterized by precision matrix $\mathbf{\Theta}^P$ with diagonal elements equal to 2. The off-diagonal elements are randomly chosen[5] and set to 0.2, so that the overall sparsity of $\mathbf{\Theta}^P$ is 25%. We then introduce changes by randomly picking 15 edges and reducing the corresponding elements in the precision matrix by 0.1. The resulting precision matrices $\mathbf{\Theta}^P$ and $\mathbf{\Theta}^Q$ are used for drawing samples as

$$\{\boldsymbol{x}_i^P\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, (\mathbf{\Theta}^P)^{-1}) \quad \text{and} \quad \{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, (\mathbf{\Theta}^Q)^{-1}),$$

where $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Data sets of size $n = n_P = n_Q = 50, 100$ are tested.

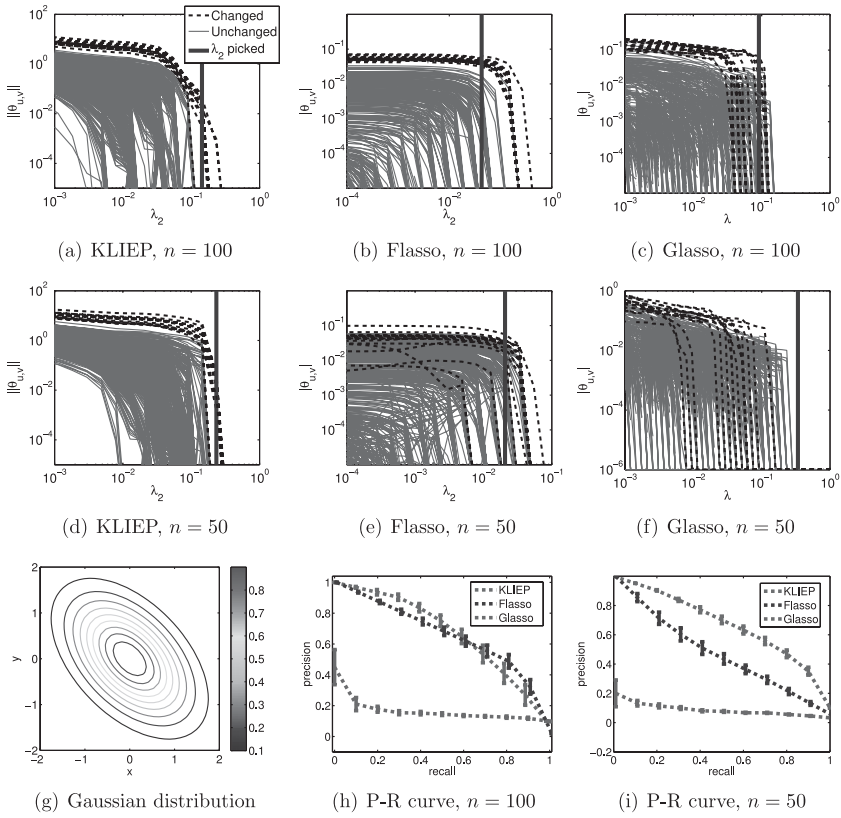
We compare the performance of the KLIEP, Flasso, and Glasso methods. Because all methods use the same gaussian model, the difference in performance is caused only by the difference in estimation methods. We repeat the experiments 20 times with randomly generated data sets and report the results in Figure 3.

The top six graphs are examples of regularization paths.[6] The dashed lines represent changed edges in the ground truth, while the solid lines represent unchanged edges. The top row is for $n = 100$, while the middle row is for $n = 50$. The bottom three graphs are the data-generating distribution and averaged precision-recall (P-R) curves with standard error over 20 runs. The P-R curves are plotted by varying the group-sparsity control parameter λ_2 with $\lambda_1 = 0$ in KLIEP and Flasso, and by varying the sparsity control parameters as $\lambda = \lambda^P = \lambda^Q$ in Glasso.

In the regularization path plots, solid vertical lines show the regularization parameter values picked based on holdout data $\{\widetilde{\boldsymbol{x}}_i^P\}_{i=1}^{3000} \overset{\text{i.i.d.}}{\sim} P$ and $\{\widetilde{\boldsymbol{x}}_i^Q\}_{i=1}^{3000} \overset{\text{i.i.d.}}{\sim} Q$ as follows:

[5] We set $\Theta_{u,v} = \Theta_{v,u}$ for not breaking the symmetry of the precision matrix.

[6] Paths of univariate factors are omitted for clear visibility.



(a) KLIEP, $n = 100$ (b) Flasso, $n = 100$ (c) Glasso, $n = 100$

(d) KLIEP, $n = 50$ (e) Flasso, $n = 50$ (f) Glasso, $n = 50$

(g) Gaussian distribution (h) P-R curve, $n = 100$ (i) P-R curve, $n = 50$

Figure 3: Experimental results on the gaussian data set.

- **KLIEP:** The holdout log likelihood (HOLL) is maximized:

$$\frac{1}{\widetilde{n}_P}\sum_{i=1}^{\widetilde{n}_P}\log\frac{\exp\left(\sum_{u\geq v}\widehat{\boldsymbol{\theta}}_{u,v}^{\top}\boldsymbol{f}(\widetilde{x}_i^{(u)P},\widetilde{x}_i^{(v)P})\right)}{\frac{1}{\widetilde{n}_Q}\sum_{j=1}^{\widetilde{n}_Q}\exp\left(\sum_{u'\geq v'}\widehat{\boldsymbol{\theta}}_{u',v'}^{\top}\boldsymbol{f}(\widetilde{x}_j^{(u')Q},\widetilde{x}_j^{(v')Q})\right)}.$$

- **Flasso:** The sum of feature-wise conditional HOLLs for $p(x^{(s)}|\boldsymbol{x}^{(-s)};\boldsymbol{\theta}_s)$ and $q(x^{(s)}|\boldsymbol{x}^{(-s)};\boldsymbol{\theta}_s)$ over all nodes is maximized:

$$\frac{1}{\widetilde{n}_P}\sum_{i=1}^{\widetilde{n}_P}\sum_{s=1}^{d}\log p(\widetilde{x}_i^{(s)P}|\widetilde{\boldsymbol{x}}_i^{(-s)P};\widehat{\boldsymbol{\theta}}_s^{P})$$

$$+\frac{1}{\widetilde{n}_Q}\sum_{i=1}^{\widetilde{n}_Q}\sum_{s=1}^{d}\log q(\widetilde{x}_i^{(s)Q}|\widetilde{\boldsymbol{x}}_i^{(-s)Q};\widehat{\boldsymbol{\theta}}_s^{Q}).$$

- **Glasso:** The sum of HOLLs for $p(\boldsymbol{x}; \boldsymbol{\theta})$ and $q(\boldsymbol{x}; \boldsymbol{\theta})$ is maximized:

$$\frac{1}{\widetilde{n}_P} \sum_{i=1}^{\widetilde{n}_P} \log p(\widetilde{\boldsymbol{x}}_i^P; \widehat{\boldsymbol{\theta}}^P) + \frac{1}{\widetilde{n}_Q} \sum_{i=1}^{\widetilde{n}_Q} \log q(\widetilde{\boldsymbol{x}}_i^Q; \widehat{\boldsymbol{\theta}}^Q).$$

When $n = 100$, KLIEP and Flasso clearly distinguish changed (dashed lines) and unchanged (solid lines) edges in terms of parameter magnitude. However, when the sample size is halved to $n = 50$, the separation is visually rather unclear in the case of Flasso. In contrast, the paths of changed and unchanged edges are still almost disjoint in the case of KLIEP. The Glasso method performs rather poorly in both cases. A similar tendency can also be observed in the P-R curve plot: when the sample size is $n = 100$, KLIEP and Flasso work equally well, but KLIEP gains its lead when the sample size is reduced to $n = 50$. Glasso does not perform well in both cases.

4.2 Nonparanormal Distribution. We post-process the gaussian data set used in section 4.1 to construct nonparanormal samples. More specifically, we apply the power function,

$$h_i^{-1}(x) = \mathrm{sign}(x)|x|^{\frac{1}{2}},$$

to each dimension of $\boldsymbol{x}^P$ and $\boldsymbol{x}^Q$, so that $\boldsymbol{h}(\boldsymbol{x}^P) \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{\Theta}^P)^{-1})$ and $\boldsymbol{h}(\boldsymbol{x}^Q) \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{\Theta}^Q)^{-1})$.

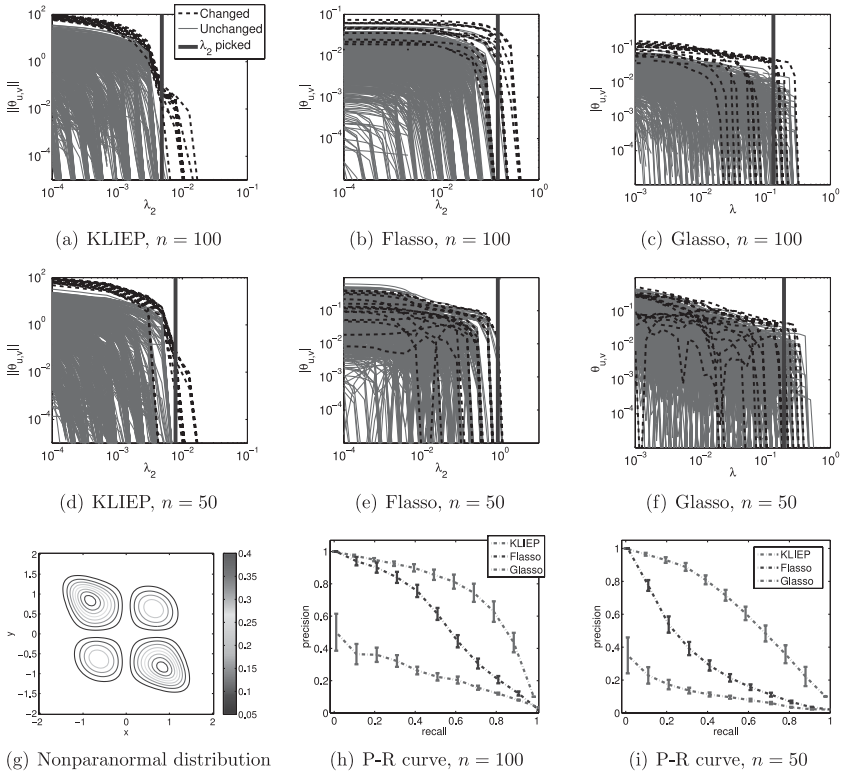
To cope with the nonlinearity in the KLIEP method, we use the power nonparanormal basis functions with power $k = 2$, 3, and 4:

$$\boldsymbol{f}(x_i, x_j) = (\mathrm{sign}(x_i)|x_i|^k, \mathrm{sign}(x_j)|x_j|^k, 1)^\top.$$

Model selection of k is performed together with the regularization parameter by HOLL maximization. For Flasso and Glasso, we apply the nonparanormal transform as described in Liu et al. (2009) before the structural change is learned.

The experiments are conducted on 20 randomly generated data sets with $n = 50$ and 100, respectively. The regularization paths, data-generating distribution, and averaged P-R curves are plotted in Figure 4. The results show that Flasso clearly suffers from performance degradation compared with the gaussian case, perhaps because the number of samples is too small for the complicated nonparanormal distribution. Due to the two-step estimation scheme, the performance of Glasso is poor. In contrast, KLIEP separates changed and unchanged edges clearly for both $n = 50$ and $n = 100$. The P-R curves also show the same tendency.

4.3 "Diamond" Distribution with No Pearson Correlation. In the experiments in section 4.2, though samples are nongaussian, the Pearson



(a) KLIEP, $n = 100$ (b) Flasso, $n = 100$ (c) Glasso, $n = 100$

(d) KLIEP, $n = 50$ (e) Flasso, $n = 50$ (f) Glasso, $n = 50$

(g) Nonparanormal distribution (h) P-R curve, $n = 100$ (i) P-R curve, $n = 50$

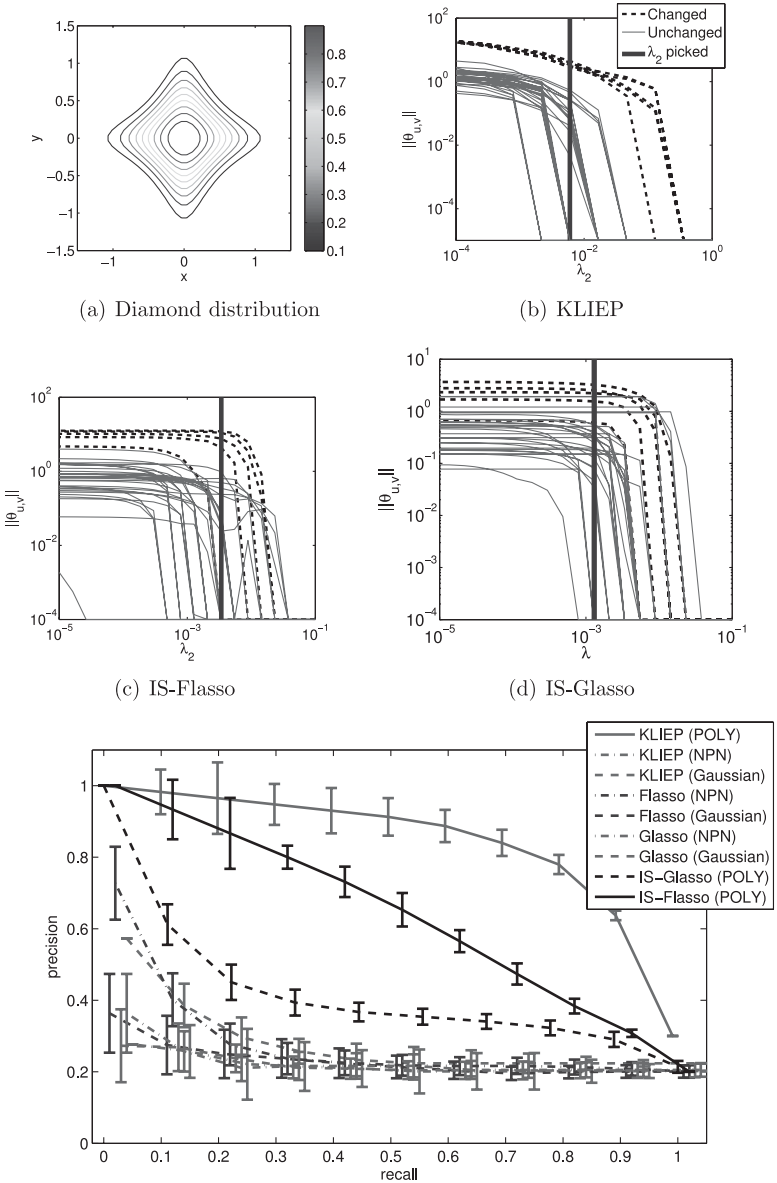
Figure 4: Experimental results on the nonparanormal data set.

correlation is not zero. Therefore, methods assuming gaussianity can still capture some linear correlation between random variables. Here, we consider a more challenging case with a diamond-shaped distribution within the exponential family that has zero Pearson correlation between variables. Thus, the methods assuming gaussianity cannot extract any information in principle from this data set.

The probability density function of the diamond distribution is defined as follows (see Figure 5a):

$$p(\boldsymbol{x}) \propto \exp\left(-\sum_{i=1}^{d} 2x_i^2 - \sum_{(i,j):A_{i,j}\neq 0} 20x_i^2 x_j^2\right), \tag{4.1}$$

where the adjacency matrix $\boldsymbol{A}$ describes the MN structure. Note that this distribution cannot be transformed into a gaussian distribution by any nonparanormal transformations.



(a) Diamond distribution (b) KLIEP

(c) IS-Flasso (d) IS-Glasso

Figure 5: Experimental results on the diamond data set. NPN and POLY denote the nonparanormal and polynomial models, respectively. Note that the precision rate of 100% recall for a random guess is approximately 20%.

We set $d = 9$ and $n_P = n_Q = 5000$. $\boldsymbol{A}^P$ is randomly generated with 35% sparsity, while $\boldsymbol{A}^Q$ is created by randomly removing edges in $\boldsymbol{A}^P$ so that the sparsity level is dropped to 15%. Samples from the above distribution are drawn by using a slice sampling method (Neal, 2003). Since generating samples from high-dimensional distributions is nontrivial and time-consuming, we focus on a relatively low-dimensional case. To avoid sampling error, which may mislead the experimental evaluation, we also increase the sample size, so that the erratic points generated by accident will not affect the overall population.

In this experiment, we compare the performance of KLIEP, Flasso, and Glasso with the gaussian model, the power nonparanormal model, and the polynomial model:

$$\boldsymbol{f}(x_i, x_j) = (x_i^k, x_j^k, x_i x_j^{k-1}, \ldots, x_i^{k-1} x_j, x_i^{k-1}, x_j^{k-1}, \ldots, x_i, x_j, 1)^\top \text{ for } i \neq j.$$

The univariate polynomial transform is defined as $\boldsymbol{f}(x_i, x_i) = \boldsymbol{f}(x_i, 0)$. We test $k = 2, 3, 4$ and choose the best one in terms of HOLL. The Flasso and Glasso methods for the polynomial model are computed by importance sampling, that is, we use the IS-Flasso and IS-Glasso methods (see section 2.5). Since these methods are computationally very expensive, we test only $k = 4$, which we found to be a reasonable choice. We set the instrumental distribution p' as the standard normal $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and use sample $\{\boldsymbol{x}'_i\}_{i=1}^{70000} \sim p'$ for approximating integrals. p' is purposely chosen so that it has a similar bell shape to the target densities but with a larger variance on each dimension.

The averaged P-R curves over 20 data sets are shown in Figure 5e. KLIEP with the polynomial model significantly outperforms all the other methods, while the IS-Glasso and especially IS-Flasso give better results than the KLIEP, Flasso, and Glasso methods with the gaussian and nonparanormal models. This means that the polynomial basis function is indeed helpful in handling completely nongaussian data. However, as discussed in section 2.2, it is difficult to use such a basis function in Glasso and Flasso because of the computational intractability of the normalization term. Although IS-Glasso can approximate integrals, the result shows that such approximation of integrals does not lead to very good performance. In comparison, the result of the IS-Flasso method is much improved thanks to the coupled sparsity regularization, but it is still not comparable to KLIEP.

The regularization paths of KLIEP with the polynomial model illustrated in Figure 5b show the usefulness of the proposed method in change detection under nongaussianity. We also give regularization paths obtained by the IS-Flasso and IS-Glasso methods on the same data set in Figures 5c and 5d, respectively. The graphs show that both methods do not separate changed and unchanged edges well, though the IS-Flasso method works slightly better.

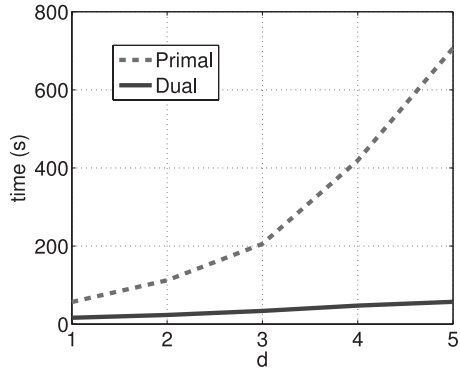


Figure 6: Comparison of computation time for solving primal and dual optimization problems.

4.4 Computation Time: Dual Versus Primal Optimization Problems. Finally, we compare the computation time of the proposed KLIEP method when solving the dual optimization problem, equation 3.4, and the primal optimization problem, equation 3.3. Both the optimization problems are solved by using the same convex optimizer *minFunc*.[7] The data sets are generated from two gaussian distributions constructed in the same way as in section 4.1. we draw 150 samples separately from two distributions with dimension $d = 40, 50, 60, 70, 80$. We then perform change detection by computing the regularization paths using 20 choices of λ_2 ranging from 10^{-4} to 10^0 and fix $\lambda_1 = 0.1$. The results are plotted in Figure 6.

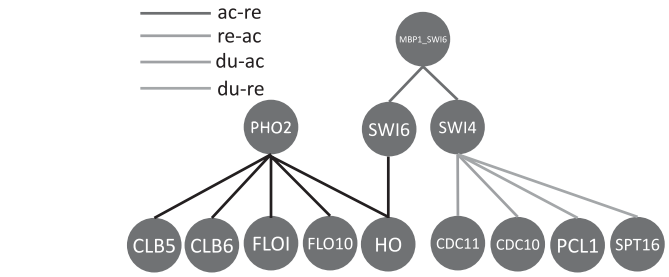
It can be seen from the graph that as the dimensionality increases, the computation time for solving the primal optimization problem is sharply increased, while that for solving the dual optimization problem grows only moderately: when $d = 80$, the computation time for obtaining the primal solution is almost 10 times more than that required for obtaining the dual solution. Thus, the dual formulation is computationally much more efficient than the primal formulation.

5 Applications

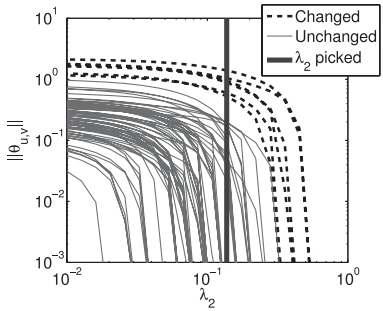
In this section, we report the experimental results on a synthetic gene expression data set and a Twitter data set.

5.1 Synthetic Gene Expression Data Set. A gene regulatory network encodes interactions between DNA segments. However, the way genes

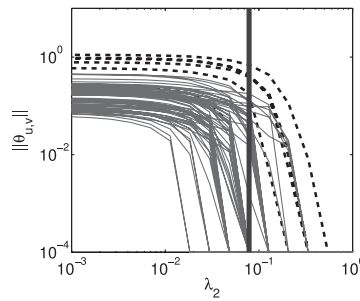
[7] http://www.di.ens.fr/∼mschmidt/Software/minFunc.html.



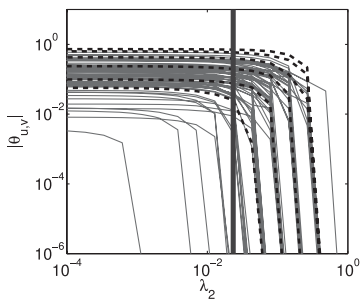
(a) Gene regulatory network



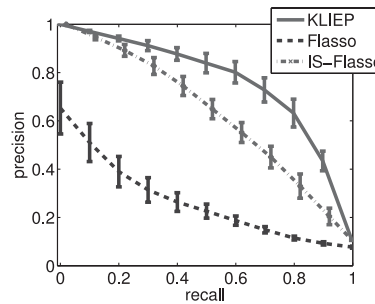
(b) KLIEP



(c) IS-Flasso



(d) Flasso



(e) P-R curve

Figure 7: Experiments on synthetic gene expression data sets.

interact may change due to environmental or biological stimuli. In this experiment, we focus on detecting such changes. We use *SynTReN*, a generator of gene regulatory networks used for benchmark validation of bioinformatics algorithms (Van den Bulcke et al., 2006).

We first choose a subnetwork containing 13 nodes from an existing signaling network in *Saccharomyces cerevisiae* (shown in Figure 7a). Three types

of interactions are modeled: activation (ac), deactivation (re), and dual (du). Fifty samples are generated in the first stage, after which we change the types of interactions in six edges and generate 50 samples again. Four types of changes are considered: ac $\rightarrow$ re, re $\rightarrow$ ac, du $\rightarrow$ ac, and du $\rightarrow$ re.

We use KLIEP and IS-Flasso with the polynomial transform function for $k \in \{2, 3, 4\}$. The regularization parameter λ_1 in KLIEP and Flasso is tested with choices $\lambda_1 \in \{0.1, 1, 10\}$. We set the instrumental distribution p' as the standard normal $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and use sample $\{\boldsymbol{x}'_i\}_{i=1}^{70000} \sim p'$ for approximating integrals in IS-Flasso.

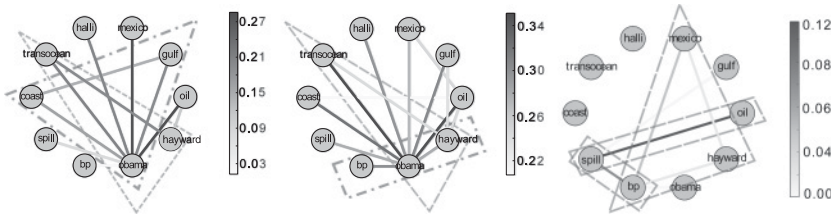
The regularization paths on one example data set for KLIEP, IS-Flasso, and the plain Flasso with the gaussian model are plotted in Figures 7b, 7c, and 7d, respectively. Averaged P-R curves over 20 simulation runs are shown in Figure 7e. We can see clearly from the KLIEP regularization paths shown in Figure 7b that the magnitude of estimated parameters on the changed pairwise interactions is much higher than that of the unchanged edges. IS-Flasso also achieves rather clear separation between changed and unchanged interactions, though a few unchanged interactions drop to zero at the final stage. Flasso gives many false alarms by assigning nonzero values to the unchanged edges, even after some changed edges hit zeros.

Reflecting a similar pattern, the P-R curves plotted in Figure 7e show that the proposed KLIEP method has the best performance among all three methods. We can also see that the IS-Flasso method achieves significant improvement over the plain Flasso method with the gaussian model. The improvement from Flasso to IS-Flasso shows that the use of the polynomial basis is helpful on this data set, and the improvement from IS-Flasso to KLIEP shows that the direct estimation can further boost the performance.

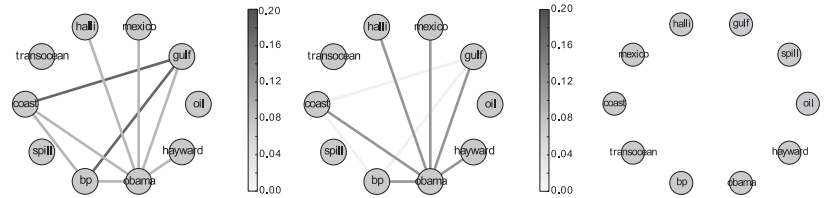
5.2 Twitter Storytelling. Finally, we use KLIEP with the polynomial transform function for $k \in \{2, 3, 4\}$ and Flasso as event detectors from Twitter. More specifically, we choose the Deepwater Horizon oil spill as the target event and hope that our method can recover some story lines from Twitter as the news events develop.[8] Counting the frequencies of 10 keywords (BP, oil, spill, Mexico, gulf, coast, Hayward, Halliburton, Transocean, and Obama), we obtain a data set by sampling four times per day from February 1, 2010, to October 15, 2010, resulting in 1061 data samples.

We segment the data into two parts: the first 300 samples collected before the day of the oil spill (April 20, 2010) are regarded as conforming to a 10-dimensional joint distribution Q, while the second set of samples that are in an arbitrary 50-day window after the oil spill accident happened is regarded as following distribution P. Thus, the MN of Q encodes the original conditional independence of frequencies between 10 keywords, while the underlying MN of P has changed since an event occurred. We

[8] http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill.



(a) April 17th–June 5th, KLIEP (b) June 6th–July 25th, KLIEP (c) July 26th–Sept. 14th, KLIEP



(d) April 17th–June 5th, Flasso (e) June 6th–July 25th, Flasso (f) July 26th–Sept. 14th, Flasso

Figure 8: Change graphs captured by the proposed KLIEP method (top) and the Flasso method (bottom). The date range beneath each figure indicates when P was sampled, while Q is fixed to dates from February 1 to April 20. Notable structures shared by the graph of both methods are surrounded by the dash-dotted lines. Unique structures that appear only in the graph of the proposed KLIEP method are surrounded by the dashed lines.

expect that unveiling changes in MNs between P and Q can recover the drift of popular topic trends on Twitter in terms of the dependency among keywords.

The detected change graphs (i.e., the graphs with only detected changing edges) on 10 keywords are illustrated in Figure 8. The edges are selected at a certain value of λ_2 indicated by the maximal cross-validated log likelihood (CVLL). Since the edge set that is picked by CVLL may not be sparse in general, we sparsify the graph based on the permutation test as follows. We randomly shuffle the samples between P and Q and repeatedly run change detection algorithms 100 times; then we observe detected edges by CVLL. Finally, we select the edges that are detected using the original nonshuffled data set and remove those that were detected in the shuffled data sets more than five times (i.e., the significance level 5%). For KLIEP, k is also tuned by using CVLL. In Figure 8, we plot detected change graphs generated using samples of P starting from April 17, July 6, and July 26, respectively.

The initial explosion happened on April 20, 2010. Both methods discover dependency changes between keywords. Generally, KLIEP captures more conditional independence changes between keywords than the Flasso method, especially when comparing Figures 8c and 8f. At the first two stages

(see Figures 8a, 8b, 8d, and 8e) the keyword *Obama* is very well connected with other keywords in the results given by both methods. Indeed, at the early development of this event, he the president lies in the center of the news stories, and his media exposure peaks after his visit to the Louisiana coast (May 2, May 28, and June 5) and his meeting with BP CEO Tony Hayward on June 16. Notably, both methods highlight the "gulf-obama-coast" triangle in Figures 8a and 8d and the "bp-obama-hayward" chain in Figures 8b and 8e.

However, there are some important differences worth mentioning. First, the Flasso method misses the "transocean-hayward-obama" triangle in Figures 8d and 8e. Transocean is the contracted operator in the Deepwater Horizon platform, where the initial explosion happened. On Figure 8c, the chain "bp-spill-oil" may indicate that the phrase "bp spill" or "oil spill" has been publicly recognized by the Twitter community since then, while the "hayward-bp-mexico" triangle, although relatively weak, may link to the event that Hayward stepped down as CEO on July 27.

It is also noted that Flasso cannot find any changed edges in Figure 8f, perhaps due to the gaussian restriction.

6 Discussion, Conclusion, and Future Work

In this letter, we proposed a direct approach to learning sparse changes in MNs by density ratio estimation. Rather than fitting two MNs separately to data and comparing them to detect a change, we estimated the ratio of the probability densities of two MNs where changes can be naturally encoded as sparsity patterns in estimated parameters. This direct modeling allows us to halve the number of parameters and approximate the normalization term in the density ratio model by a sample average without sampling. We also showed that the number of parameters to be optimized can be further reduced with the dual formulation, which is highly useful when the dimensionality is high. Through experiments on artificial and real-world data sets, we demonstrated the usefulness of the proposed method over state-of-the-art methods, including nonparanormal-based methods, and sampling-based methods.

Our important future work is to theoretically elucidate the advantage of the proposed method, beyond Vapnik's principle of solving the target problem directly. The relation to score matching (Hyvärinen, 2005), which avoids computing the normalization term in density estimation, is also an interesting issue to be further investigated. Considering higher-order MN models such as the hierarchical log-linear model (Schmidt & Murphy, 2010) is a promising direction for extension.

In the context of change detection, we are mainly interested in the situation where p and q are close to each other (if p and q are completely different, it is straightforward to detect changes). When p and q are similar, density

ratio estimation for $p(\boldsymbol{x})/q(\boldsymbol{x})$ or $q(\boldsymbol{x})/p(\boldsymbol{x})$ performs similarly. However, given the asymmetry of density ratios, the solutions for $p(\boldsymbol{x})/q(\boldsymbol{x})$ or $q(\boldsymbol{x})/p(\boldsymbol{x})$ are generally different. The choice of the numerator and denominator in the ratio is left for future investigation.

Detecting changes in MNs is the main target of this letter. Estimating the difference or divergence between two probability distributions has been studied under a more general context in the statistics and machine learning communities (Amari & Nagaoka, 2000; Eguchi & Copas, 2006; Wang, Kulkarni, & Verdú, 2009; Sugiyama, Suzuki, & Kanamori, 2012b; Sugiyama, Liu, et al., 2013). In fact, the estimation of the Kullback-Leibler divergence (Kullback & Leibler, 1951) is related to the KLIEP-type density ratio estimation method (Nguyen, Wainwright, & Jordan, 2010), and the estimation of the Pearson divergence (Pearson, 1900) is related to the squared-loss density ratio estimation method (Kanamori, Hido, & Sugiyama, 2009). However, the density-ratio-based divergences tend to be sensitive to outliers. To overcome this problem, a divergence measure based on relative density ratios was introduced, and its direct estimation method was developed (Yamada, Suzuki, Kanamori, Hachiya, & Sugiyama, 2013). L^2-distance is another popular difference measure between probability density functions. L^2-distance is symmetric, unlike the Kullback-Leibler divergence and the Pearson divergence, and its direct estimation method has been investigated recently (Sugiyama, Suzuki, et al., 2013; Kim & Scott, 2010).

Change detection in time series is a related topic. A straightforward approach is to evaluate the difference (dissimilarity) between two consecutive segments of time-series signals. Various methods have been developed to identify the difference by fitting two models to two segments of time series separately, for example, the singular spectrum transform (Moskvina & Zhigljavsky, 2003; Ide & Tsuda, 2007), subspace identification (Kawahara, Yairi, & Machida, 2007), and the method based on the one-class support vector machine (Desobry, Davy, & Doncarli, 2005). In the same way as this letter, direct modeling of the change has also been explored for change detection in time-series (Kawahara & Sugiyama, 2012; Liu, Yamada, Collier, & Sugiyama, 2013; Sugiyama, Suzuki, et al., 2013).

Appendix: Derivation of the Dual Optimization Problem

First, we rewrite the optimization problem, equation 3.3 as

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \left[\log\left(\sum_{i=1}^{n_Q} \exp\left(w_i\right) \right) - \boldsymbol{\theta}^\top \boldsymbol{g} + \frac{\lambda_1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \lambda_2 \sum_{u\geq v} \|\boldsymbol{\theta}_{u,v}\| - C \right] \quad \text{(A.1)}$$

$$\text{subject to } \boldsymbol{w} = \boldsymbol{H}^\top\boldsymbol{\theta},$$

where

$$\begin{aligned}
\boldsymbol{w} &= (w_1, \ldots, w_{n_Q})^\top, \\
\boldsymbol{H} &= (\boldsymbol{H}_{1,1}^\top, \ldots, \boldsymbol{H}_{d,1}^\top, \boldsymbol{H}_{2,2}^\top, \ldots, \boldsymbol{H}_{d,2}^\top, \ldots, \boldsymbol{H}_{d,d}^\top)^\top, \\
\boldsymbol{H}_{u,v} &= [\boldsymbol{f}(x_1^{(u)Q}, x_1^{(v)Q}), \ldots, \boldsymbol{f}(x_{n_Q}^{(u)Q}, x_{n_Q}^{(v)Q})], \\
\boldsymbol{g} &= (\boldsymbol{g}_{1,1}^\top, \ldots, \boldsymbol{g}_{d,1}^\top, \boldsymbol{g}_{2,2}^\top, \ldots, \boldsymbol{g}_{d,2}^\top, \ldots, \boldsymbol{g}_{d,d}^\top)^\top, \\
\boldsymbol{g}_{u,v} &= \frac{1}{n_P} \sum_{i=1}^{n_P} \boldsymbol{f}(x_i^{(u)P}, x_i^{(v)P}), \\
C &= \log n_Q.
\end{aligned}$$

With Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_Q})^\top$, the Lagrangian of equation A.1 is given as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}) &= \min_{\boldsymbol{w},\boldsymbol{\theta}} \left[\log \sum_{i=1}^{n_Q} \exp(w_i) - \boldsymbol{\theta}^\top \boldsymbol{g} + \frac{\lambda_1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right. \\
&\qquad \left. + \lambda_2 \sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\| - (\boldsymbol{w} - \boldsymbol{H}^\top \boldsymbol{\theta})^\top \boldsymbol{\alpha} \right] - C \\
&= \min_{\boldsymbol{w}} \left[\log \sum_{i=1}^{n_Q} \exp(w_i) - \boldsymbol{w}^\top \boldsymbol{\alpha} \right] \\
&\qquad + \min_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top (\boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{g}) + \frac{\lambda_1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \lambda_2 \sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\| \right] - C \\
&= \min_{\boldsymbol{w}} \psi_1(\boldsymbol{w}) + \min_{\boldsymbol{\theta}} \psi_2(\boldsymbol{\theta}) - C. \qquad \text{(A.2)}
\end{aligned}$$

A few lines of algebra can show that $\psi_1(\boldsymbol{w})$ reaches the minimum $-\sum_{i=1}^{n_Q} \alpha_i \log \alpha_i$ at

$$\alpha_i = \frac{\exp(w_i)}{\sum_{i=1}^{n_Q} \exp(w_i)}, \quad i = 1, \ldots, n_Q.$$

Note that extra constraints are implied from the above equation:

$$\alpha_1, \ldots, \alpha_{n_Q} \geq 0 \quad \text{and} \quad \sum_{i=1}^{n_Q} \alpha_i = 1.$$

Since $\psi_2(\boldsymbol{\theta})$ is not differentiable at $\boldsymbol{\theta}_{u,v} = \mathbf{0}$, we can only obtain its subgradient:

$$\nabla_{\boldsymbol{\theta}_{u,v}} \psi_2(\boldsymbol{\theta}) = -\boldsymbol{\xi}_{u,v} + \lambda_1 \boldsymbol{\theta} + \lambda_2 \nabla_{\boldsymbol{\theta}_{u,v}} \|\boldsymbol{\theta}_{u,v}\|,$$

where

$$\boldsymbol{\xi}_{u,v} = \boldsymbol{g}_{u,v} - \boldsymbol{H}_{u,v}\boldsymbol{\alpha},$$

$$\nabla_{\boldsymbol{\theta}_{u,v}} \|\boldsymbol{\theta}_{u,v}\| = \begin{cases} \dfrac{\boldsymbol{\theta}_{u,v}}{\|\boldsymbol{\theta}_{u,v}\|} & \text{if } \boldsymbol{\theta}_{u,v} \neq \mathbf{0}, \\ \{\boldsymbol{y} | \|\boldsymbol{y}\| \leq 1\} & \text{if } \boldsymbol{\theta}_{u,v} = \mathbf{0}. \end{cases}$$

By setting $\nabla_{\boldsymbol{\theta}_t} \psi_2(\boldsymbol{\theta}) = \mathbf{0}$, we can obtain the solution to this minimization problem by equation 3.5.

Substituting the solutions of the above two minimization problems with respect to $\boldsymbol{\theta}$ and $\boldsymbol{w}$ into equation A.2, we obtain the dual optimization problem, equation 3.4.

Acknowledgments

S.L. is supported by the JST PRESTO program and the JSPS fellowship. J.Q. is supported by the JST PRESTO program. M.U.G. is supported by the Finnish Centre-of-Excellence in Computational Inference Research COIN (251170). T.S. is partially supported by MEXT Kakenhi 25730013, and the Aihara Project, the FIRST program from JSPS, initiated by CSTP. M.S. is supported by the JST CREST program and AOARD.

References

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry.* Providence, RI: American Mathematical Society, and New York: Oxford University Press.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, *9*, 485–516.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization.* Cambridge: Cambridge University Press.

Danaher, P., Wang, P., & Witten, D. M. (2013). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 373–397.

Desobry, F., Davy, M., & Doncarli, C. (2005). An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, *53*(8), 2961–2974.

Eguchi, S., & Copas, J. (2006). Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis*, *97*(9), 2034–2040.

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

Gelman, A. (1995). Method of moments using Monte Carlo simulation. *Journal of Computational and Graphical Statistics*, *4*(1), 36–54.

Gutmann, M. U., & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, *13*, 307–361.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–1800.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*, 695–709.

Ide, T., & Tsuda, K. (2007). Change-point detection using Krylov subspace learning. In *Proceedings of the SIAM International Conference on Data Mining* (pp. 515–520). Pittsburgh, PA: Society of Industrial and Applied Mathematics.

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

Kanamori, T., Suzuki, T., & Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E93-A*(4), 787–798.

Kawahara, Y., & Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, *5*(2), 114–127.

Kawahara, Y., Yairi, T., & Machida, K. (2007). Change-point detection in time-series data based on subspace identification. In *Proceedings of the 7th IEEE International Conference on Data Mining* (pp. 559–564). Piscataway, NJ: IEEE.

Kim, J., & Scott, C. (2010). L_2 kernel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(10), 1822–1831.

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

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lee, S.-I., Ganapathi, V., & Koller, D. (2007). Efficient structure learning of Markov networks using l_1-regularization. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems, 19* (pp. 817–824). Cambridge, MA: MIT Press.

Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). The nonparanormal skeptic. In *Proceedings of the 29th International Conference on Machine Learning.* Madison, WI: Omnipress.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, *10*, 2295–2328.

Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, *43*, 72–83.

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

Moskvina, V., & Zhigljavsky, A. (2003). Change-point detection algorithm based on the singular-spectrum analysis. *Communications in Statistics: Simulation and Computation*, *32*, 319–352.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*(3), 705–741.

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.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, *38*(3), 1287–1319.

Robert, C. P., & Casella, G. (2005). *Monte Carlo statistical methods.* New York: Springer-Verlag.

Schmidt, M. W., & Murphy, K. P. (2010). Convex structure learning in log-linear models: Beyond pairwise potentials. *Journal of Machine Learning Research: Proceedings Track*, *9*, 709–716.

Sugiyama, M., Liu, S., du Plessis, M. C., Yamanaka, M., Yamada, M., Suzuki, T., & Kanamori, T. (2013). Direct divergence approximation between probability distributions and its applications in machine learning. *Journal of Computing Science and Engineering*, *7*(2), 99–111.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012a). *Density ratio estimation in machine learning.* Cambridge: Cambridge University Press.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012b). Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, *64*(5), 1009–1044.

Sugiyama, M., Suzuki, T., Kanamori, T., du Plessis, M. C., Liu, S., & Takeuchi, I. (2013). Density-difference estimation. *Neural Computation*, *25*(10), 2734–2775.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*(4), 699–746.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., & Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, *17*, 138–155.

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., . . . Marchal, K. (2006). SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, *7*(1), 43.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: Wiley.

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, Q., Kulkarni, S. R., & Verdú, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, *55*(5), 2392–2405.

Wasserman, L. (2010). *All of statistics: A concise course in statistical inference.* New York: Springer.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., & Sugiyama, M. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, *25*(5), 1324–1370.

Zhang, B., & Wang, Y. (2010). Learning structural changes of gaussian graphical models in controlled experiments. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (pp. 701–708). Corvallis, OR: AUAI Press.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*(2), 301–320.

Received September 27, 2013; accepted December 31, 2013.