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

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

Song Liu[1], John A. Quinn[2], Michael U. Gutmann[3], and Masashi Sugiyama[1]

[1] Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro, Tokyo 152-8552, Japan.
{song@sg., sugi@}cs.titech.ac.jp
[2] Makerere University, P.O. Box 7062, Kampala, Uganda.
jquinn@cit.ac.ug
[3] University of Helsinki and HIIT, Finland, P.O. Box 68, FI-00014, Finland.
michael.gutmann@helsinki.fi

Abstract. 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, which is a critical computational bottleneck of the naive approach, can be remarkably mitigated. Through experiments on gene expression and Twitter data analysis, we demonstrate the usefulness of our method.

1 Introduction

Changes in the structure of interactions between random variables are interesting in many real-world phenomena. For example, genes may interact with each other in different ways when external stimuli change, co-occurrence between words may disappear/appear 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 paper, 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 as an undirected graphical model called a *Markov network* (MN) [1,2,3], where nodes and edges represent variables and their conditional dependency. As a simple and widely applicable case, the 2nd-order pairwise MN model has been thoroughly studied recently [4,5]. 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 learned MNs. However,

Fig. 1. The rationale of direct structural change learning.

MLE is often computationally expensive due to the normalization factor included in the density model. There are estimation methods which do not rely on knowing the normalization factor [6], but Gaussianity is often assumed for computing the normalization factor analytically [7]. However, this Gaussian assumption is highly restrictive in practice.

Another conceptual weakness of the above two-step procedure is that structure change is not directly learned. This indirect nature causes a problem, for example, if we want to learn a sparse structure change. For learning sparse changes, we may utilize ℓ_1-regularized MLE [8,9,5], which produces sparse MNs and thus the change between MNs also becomes sparse. However, this approach does not work if MNs are rather dense but change is sparse.

To mitigate this indirect nature, the *fused lasso* [10] is useful, where two MNs are simultaneously learned with a sparsity-inducing penalty on the difference between two MN parameters [11]. 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 efficient way.

A *nonparanormal* assumption [12,13] is a useful generalization of the Gaussian assumption. A nonparanormal distribution is a *semi-parametric Gaussian copula* where each Gaussian variable is transformed by a non-linear function. Nonparanormal distributions are much more flexible than Gaussian distributions thanks to the feature-wise non-linear transformation, while the normalization factors can still be computed analytically.

Thus, the fused-lasso method combined with nonparanormal models would be the state-of-the-art approach to change detection in MNs. However, the fused-lasso method is still based on separate modeling of two MNs, and its computation for more general non-Gaussian distributions is challenging.

In this paper, we propose a more direct approach to structural change learning in MNs based on *density ratio estimation* (DRE) [14]. Our method does not separately model two MNs, but directly models the *change* in two MNs. This idea follows Vapnik's principle [15]:

> 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, 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 (Figure 1). This direct approach means that we halve the number of parameters, from two MNs to one MN-difference.

Furthermore, the normalization factor in our DRE-based method can be approximated efficiently, because the normalization term in a density ratio function takes the form of an expectation and thus it can be simply approximated by sample averages without sampling.

The remainder of this paper is structured as follows. In Section 2, we formulate the problem of detecting structural changes and review currently available approaches. We then propose our DRE-based structural change detection method in Section 3. Results of illustrative and real-world experiments are reported in Section 4 and Section 5, respectively. Finally, we conclude our work and show future directions 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 samples drawn separately from two probability distributions P and Q on $\mathbb{R}^d$:

$$\{\boldsymbol{x}_i^P\}_{i=1}^{n_P} \overset{\text{iid}}{\sim} p(\boldsymbol{x}) \text{ and } \{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{iid}}{\sim} q(\boldsymbol{x}).$$

We assume that p and q belong to the family of *Markov networks* (MNs) consisting of univariate and bivariate factors:

$$p(\boldsymbol{x};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \exp\left(\sum_{i=1}^{d} \boldsymbol{\alpha}_i^\top \boldsymbol{g}_i(x_i) + \sum_{i,j=1,i>j}^{d} \boldsymbol{\alpha}_{i,j}^\top \boldsymbol{g}_{i,j}(x_i, x_j)\right), \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$, $\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_{i,j}$ are parameters, $\boldsymbol{g}_i, \boldsymbol{g}_{i,j}$ are univariate and bivariate vector-valued basis functions, and $Z(\boldsymbol{\alpha})$ is the normalization factor. $q(\boldsymbol{x};\boldsymbol{\alpha})$ is defined in the same way.

For notational simplicity, we unify both univariate and bivariate factors as

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_t \boldsymbol{\theta}_t^\top \boldsymbol{f}_t(\boldsymbol{x})\right), \quad \text{where} \quad Z(\boldsymbol{\theta}) = \int \exp\left(\sum_t \boldsymbol{\theta}_t^\top \boldsymbol{f}_t(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}.$$

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

Our goal is to detect the change in conditional independence between random variables between P to Q.

2.2 Sparse MLE and Graphical Lasso

Maximum likelihood estimation (MLE) with group ℓ_1-regularization has been widely used for estimating the sparse structure of MNs [16,4,5]:

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\boldsymbol{x}_i^P; \boldsymbol{\theta}) - \lambda \sum_t \|\boldsymbol{\theta}_t\|, \tag{2}$$

where $\|\cdot\|$ denotes the ℓ_2-norm. As λ increases, $\boldsymbol{\theta}_t$ for pairwise factors may drop to 0. Thus, this method favors an MN that encodes more conditional independencies among variables. For computing the normalization term $Z(\boldsymbol{\theta})$ in Eq.(1), sampling techniques such as Markov-chain Monte-Carlo (MCMC) and importance sampling are usually employed. However, obtaining a reasonable value by these methods becomes computationally more expensive as the dimension d grows.

To avoid this computational problem, the Gaussian assumption is often imposed [9,17]. If we consider a zero-mean Gaussian distribution, the following $p(\boldsymbol{x}; \boldsymbol{\Theta})$ can be used to replace the density model in Eq.(2):

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

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

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

where $\boldsymbol{S}^P$ is the sample covariance matrix of $\{\boldsymbol{x}_i^P\}_{i=1}^n$. $\|\boldsymbol{\Theta}\|_1$ is the ℓ_1-norm of $\boldsymbol{\Theta}$, i.e., the absolute sum of all elements. This formulation has been studied intensively in [8], and a computationally efficient solution called the *graphical lasso* [9] has been proposed.

Sparse changes in conditional independence structure between P and Q can be detected by comparing two MNs separately estimated 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* [10] has been developed [11]. This method jointly maximizes the conditional likelihood in a feature-wise manner for P and Q with a sparsity penalty on the *difference* between parameters. More specifically, for each element x_s $(s = 1, \ldots, d)$ of $\boldsymbol{x}$,

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

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

$$\ell_s^P(\boldsymbol{\theta}) = \sum_{i=1}^{n_P} \log p(x_{i,s}^P | \boldsymbol{x}_{i,-s}^P; \boldsymbol{\theta}).$$

$\ell_s^Q(\boldsymbol{\theta})$ is defined in the same way as $\ell_s^P(\boldsymbol{\theta})$. In this fused-lasso method, Gaussianity is usually assumed to cope with 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, which is a highly restrictive assumption.

To overcome this restriction, it has become popular to perform structure learning under the *nonparanormal* settings [12,13], where the Gaussian distribution is replaced by a *semi-parametric Gaussian copula*. $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ 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 non-linear transformation $\{h_i(x)\}_{i=1}^d$, while the normalization factors can still be computed in an analytical way.

3 Direct Learning of Structural Changes via Density Ratio Estimation

The fused-lasso method can more naturally handle sparse changes in MNs than separate sparse MLE, and its nonparanormal extension is more flexible than the Gaussian counterpart. However, the fused-lasso method is still based on separate modeling of two MNs, and its computation for more general non-Gaussian distributions is challenging.

In this section, we propose to directly learn structural changes based on *density ratio estimation* [14], which does not involve separate modeling of each MN and which allows us to approximate the normalization term efficiently.

3.1 Density Ratio Formulation for Structural Change Detection

Our key idea is to consider the ratio of p and q:

$$\frac{p(\boldsymbol{x}; \boldsymbol{\theta}^P)}{q(\boldsymbol{x}; \boldsymbol{\theta}^Q)} \propto \exp\left(\sum_t (\boldsymbol{\theta}_t^P - \boldsymbol{\theta}_t^Q)^\top \boldsymbol{f}_t(\boldsymbol{x})\right).$$

Here $\boldsymbol{\theta}_t^P - \boldsymbol{\theta}_t^Q$ encodes the difference between P and Q for factor $\boldsymbol{f}_t$, i.e., $\boldsymbol{\theta}_t^P - \boldsymbol{\theta}_t^Q$ is zero if there is no change in the t-th factor.

Once we consider the ratio of p and q, we actually do not have to estimate $\boldsymbol{\theta}_t^P$ and $\boldsymbol{\theta}_t^Q$; instead an estimate of their difference $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^P - \boldsymbol{\theta}_t^Q$ is sufficient for change detection:

$$r(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{N(\boldsymbol{\theta})} \exp\left(\sum_t \boldsymbol{\theta}_t^\top f_t(\boldsymbol{x})\right), \quad \text{where} \quad N(\boldsymbol{\theta}) = \int q(\boldsymbol{x}) \exp\left(\sum_t \boldsymbol{\theta}_t^\top f_t(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}. \tag{3}$$

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 each 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 [15]. 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 sample average over $\{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{iid}}{\sim} q(\boldsymbol{x})$, because $N(\boldsymbol{\theta})$ is the expectation over $q(\boldsymbol{x})$:

$$N(\boldsymbol{\theta}) \approx \frac{1}{n_Q} \sum_{i=1}^{n_Q} \exp\left(\sum_t \boldsymbol{\theta}_t^\top f_t(\boldsymbol{x}_i^Q)\right).$$

3.2 Direct Density-Ratio Estimation

Density ratio estimation (DRE) methods have been recently introduced to the machine learning community [14] and are proven to be useful in a wide range of applications. Here, we concentrate on a DRE method called the *Kullback-Leibler importance estimation procedure* (KLIEP) for a log-linear model [18,19].

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

$$\mathrm{KL}[p \| \widehat{p}] = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{x} = \text{Const.} - \int p(\boldsymbol{x}) \log r(\boldsymbol{x};\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{x}. \tag{4}$$

Note that our density-ratio model (3) automatically satisfies the non-negativity and normalization constraints:

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

[4] 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. However, the expectation with respect to a model distribution can be computationally expensive as in the case of MLE, and this alternative form requires an extra parameter $\boldsymbol{\theta}^Q$ which is not our main interest. It is noteworthy that the use of $N(\boldsymbol{\theta})$ as a normalization factor guarantees the consistency of density ratio estimation [18].

In practice, we maximize the empirical approximation of the second term in the right-hand side of Eq.(4):

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

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 $\boldsymbol{\theta}_t$ is given by

$$\nabla_{\boldsymbol{\theta}_t}\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) = \frac{1}{n_P}\sum_{i=1}^{n_P}\boldsymbol{f}_t(\boldsymbol{x}_i^P) - \frac{\frac{1}{n_Q}\sum_{i=1}^{n_Q}\exp\left(\sum_t \boldsymbol{\theta}_t^\top \boldsymbol{f}_t(\boldsymbol{x}_i^Q)\right)\boldsymbol{f}_t(\boldsymbol{x}_i^Q)}{\frac{1}{n_Q}\sum_{j=1}^{n_Q}\exp\left(\sum_t \boldsymbol{\theta}_t^\top f_t(\boldsymbol{x}_j^Q)\right)}.$$

3.3 Sparsity-Inducing Norm

To find a sparse change in P and Q, we may regularize our KLIEP solution with a sparsity-inducing norm $\sum_t \|\boldsymbol{\theta}_t\|$. Note that the motivation for introducing sparsity in KLIEP is different from MLE. In the case of MLE, both $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ are sparsified and then consequently the difference $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$ is also sparsified. On the other hand, in our case, only the difference $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$ is sparsified; 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 [20] to better control overfitting to noisy data:

$$\max_{\boldsymbol{\theta}}\left[\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) - \lambda_1\|\boldsymbol{\theta}\|^2 - \lambda_2\sum_t\|\boldsymbol{\theta}_t\|\right],$$

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

4 Numerical Experiments

In this section, we compare the proposed KLIEP-based method with the Fused-lasso (Flasso) method [11] and the Graphical-lasso (Glasso) method [9]. Results are reported on datasets with three different underlying distributions: multivariate Gaussian, nonparanormal, and a non-Gaussian "diamond" distribution.

4.1 Setup

Performance Metrics: by taking the advantage of knowing the ground truth of structural changes in artificial experiments, we measure the performance of change detection methods using the *precision-recall (P-R) curve*. For KLIEP and Flasso, a precision and recall curve can be plotted by varying the group-sparsity control parameter λ_2; we fix $\lambda_1 = 0$ because the artificial datasets are noise-free. For Glasso, we vary the sparsity control parameters as $\lambda = \lambda^P = \lambda^Q$.

Model Selection: for KLIEP, we use the log-likelihood of an estimated density ratio on a hold-out dataset, which we refer to as *hold-out log-likelihood* (HOLL). More precisely, given two sets of hold-out data $\{\widetilde{\boldsymbol{x}}_i^P\}_{i=1}^{\widetilde{n}_P} \overset{\text{iid}}{\sim} P$, $\{\widetilde{\boldsymbol{x}}_i^Q\}_{i=1}^{\widetilde{n}_Q} \overset{\text{iid}}{\sim} Q$ for $\widetilde{n}_P = \widetilde{n}_Q = 3000$, we use the following quantity:

$$\ell_{\text{HOLL}} = \frac{1}{\widetilde{n}_P} \sum_{i=1}^{\widetilde{n}_P} \log \frac{\exp\left(\sum_t \widehat{\boldsymbol{\theta}}_t^\top f_t(\widetilde{\boldsymbol{x}}_i^P)\right)}{\frac{1}{\widetilde{n}_Q}\sum_{j=1}^{\widetilde{n}_Q} \exp\left(\sum_t \widehat{\boldsymbol{\theta}}_t^\top f_t(\widetilde{\boldsymbol{x}}_j^Q)\right)}.$$

In case such a hold-out dataset is not available, the *cross-validated log-likelihood* (CVLL) may be used instead.

For the Glasso and Flasso methods, we perform model selection by adding the hold-out/cross-validated likelihoods on $p(\boldsymbol{x};\boldsymbol{\theta})$ and $q(\boldsymbol{x};\boldsymbol{\theta})$ together:

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

Basis Function: we consider two types of f_t: a power nonparanormal f_{npn} and a polynomial transform f_{poly}.

The pairwise nonparanormal transform with power k is defined as

$$f_{\text{npn}}(x_i, x_j) := [\text{sign}(x_i)x_i^k \text{sign}(x_j)x_j^k, 1].$$

This transforms the original data by the power of k, so that the transformed data are jointly Gaussian (see Section 4.3). The univariate nonparanormal transform is defined as $f_{\text{npn}}(x_i) := f_{\text{npn}}(x_i, x_i)$.

The polynomial transform up to degree of k is defined as:

$$f_{\text{poly}}(x_i, x_j) := [x_i^k, x_j^k, x_i x_j^{k-1}, \ldots, x_i^{k-1}x_j, x_i^{k-1}, x_j^{k-1}, \ldots, x_i, x_j, 1].$$

The univariate polynomial transform is defined as $f_{\text{poly}}(x_i) := f_{\text{poly}}(x_i, 0)$.

4.2 Multivariate Gaussian

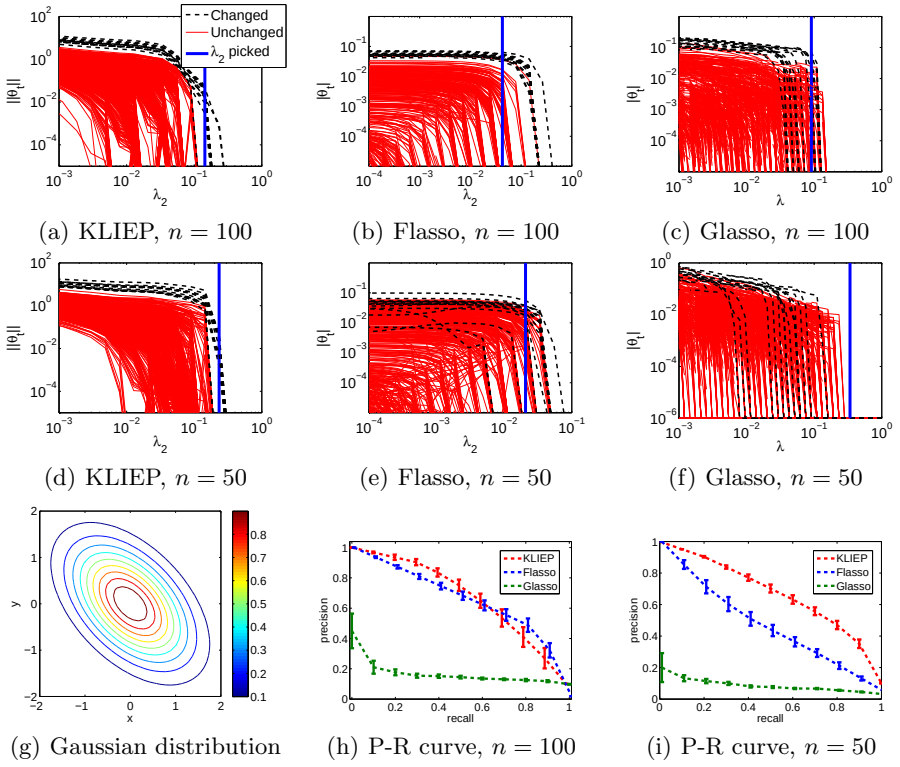
First, we investigate the performance of each learning method under Gaussianity.

Consider a 40-node sparse Gaussian MN, where its graphical structure is characterized by precision matrix $\boldsymbol{\Theta}^P$ 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 $\boldsymbol{\Theta}^P$ 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 $\boldsymbol{\Theta}^P$ and $\boldsymbol{\Theta}^Q$ are used for drawing samples as

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

Datasets of size $n = 50$ and $n = 100$ are tested.

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



(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, $n = 50$

Fig. 2. Experimental results on the multivariate Gaussian dataset.

We repeat the experiments 20 times with randomly generated datasets and report the results in Figure 2. The top 6 graphs are examples of regularization paths (black and red color represents the ground truth) and the bottom 3 graphs are the data generating distribution and averaged P-R curves with standard error. The top row is for $n = 100$ while the middle row is for $n = 50$. The regularization parameters picked by the model selection procedures described in Section 4.1 are marked with blue vertical lines. In this experiment, the Gaussian model (the nonparanormal basis function with power $k = 1$) is used for KLIEP. Because the Gaussian model is also used in Flasso and Glasso, the difference in performance is caused only by the difference of estimation methods.

When $n = 100$, KLIEP and Flasso clearly distinguish changed (black) and unchanged (red) edges in terms of parameter magnitude. However, when the sample size is halved, 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 be observed also in the averaged P-R curve. When the sample size is 100, KLIEP and Flasso work equally well, but KLIEP gains its lead when the sample size is reduced. Glasso does not perform well in both cases.

4.3 Nonparanormal

We post-process the dataset used in Section 4.2 to construct nonparanormal samples: simply, 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 $\boldsymbol{h}(\boldsymbol{x}^P) \sim \mathcal{N}(\boldsymbol{0}, (\boldsymbol{\Theta}^P)^{-1})$ and $\boldsymbol{h}(\boldsymbol{x}^Q) \sim \mathcal{N}(\boldsymbol{0}, (\boldsymbol{\Theta}^Q)^{-1})$.

In order to cope with the non-linearity, we apply the nonparanormal basis function with power 2, 3 and 4 in KLIEP and choose the one that maximizes the peak HOLL value. For Flasso and Glasso, we apply the nonparanormal transform described in [12] before the structural change is learned.

The experiments are conducted on 20 randomly generated datasets with $n = 50$ and 100, respectively. The regularization paths, data generating distribution, and averaged P-R curves are plotted in Figure 3. The results show that Flasso clearly suffers from the 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 still clearly for both $n = 50$ and $n = 100$. The P-R curves also show the same tendency.

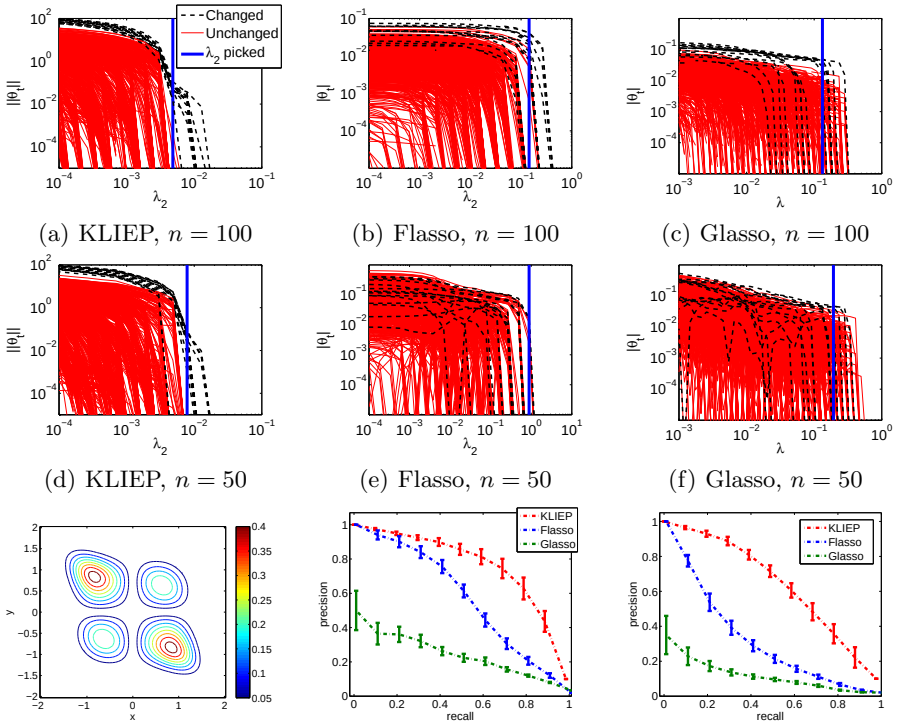
4.4 "Diamond" Distribution with No Pearson Correlation

In the previous experiment, though samples are non-Gaussian, the *Pearson correlation* is not zero. Therefore, methods assuming Gaussianity can still capture the linear correlation between random variables. In this experiment, we consider a more challenging case with a diamond-shaped distribution within the exponential family that has zero Pearson correlation coefficient between dependent variables. Thus, the methods assuming Gaussianity (i.e., Glasso and Flasso) can not extract any information in principle from this dataset.

The probability density function of the diamond distribution is defined as follows (Figure 4(a)):

$$p(\boldsymbol{x}) \propto \exp\left(-\sum_i 2x_i^2 - \sum_{(i,j):A_{i,j}\neq 0} 20x_i^2 x_j^2\right), \tag{5}$$

where the adjacency matrix $\boldsymbol{A}$ describes an MN structure. Note that this distribution can not be transformed into a Gaussian distribution by any nonparanormal transformations. Samples from the above distribution are drawn by using a *slice sampling* method [21]. However, since generating samples from a high-dimensional distribution is non-trivial and time-consuming, we focus on a relatively low-dimensional case to avoid sampling errors which may mislead the experimental evaluation.



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

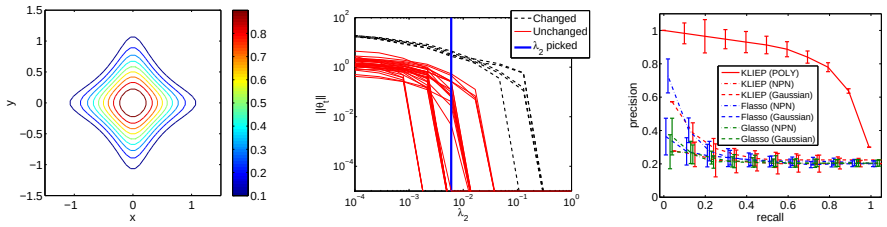
(d) KLIEP, $n = 50$ (e) Flasso, $n = 50$ (f) Glasso, $n = 50$

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

Fig. 3. Experimental results on the nonparanormal dataset.

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%.

In this experiment, we compare the performance of all three methods with their available transforms: KLIEP ($f_{\mathrm{poly}}, k = 2, 3, 4$), KLIEP ($f_{\mathrm{npn}}, k = 2, 3, 4$), KLIEP ($f_{\mathrm{npn}}, k = 1$; same as the Gaussian model), Flasso (nonparanormal), Flasso (Gaussian), Glasso (nonparanormal) and Glasso (Gaussian). The averaged P-R curves are shown in Figure 4(c). As expected, except KLIEP (f_{poly}), all other methods do not work properly. This means that the polynomial kernel is indeed very helpful in handling completely non-Gaussian data. However, as discussed in Section 2.2, it is difficult to use such a kernel in the MLE-based approaches (Glasso and Flasso) because computationally demanding sampling is involved in evaluating the normalization term. The regularization path of KLIEP (f_{poly}) illustrated in Figure 4(b) shows the usefulness of the proposed method in change detection under non-Gaussianity.



(a) Diamond distribution (b) KLIEP (c) P-R curve

Fig. 4. Experimental results on the diamond dataset. "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%.

5 Applications

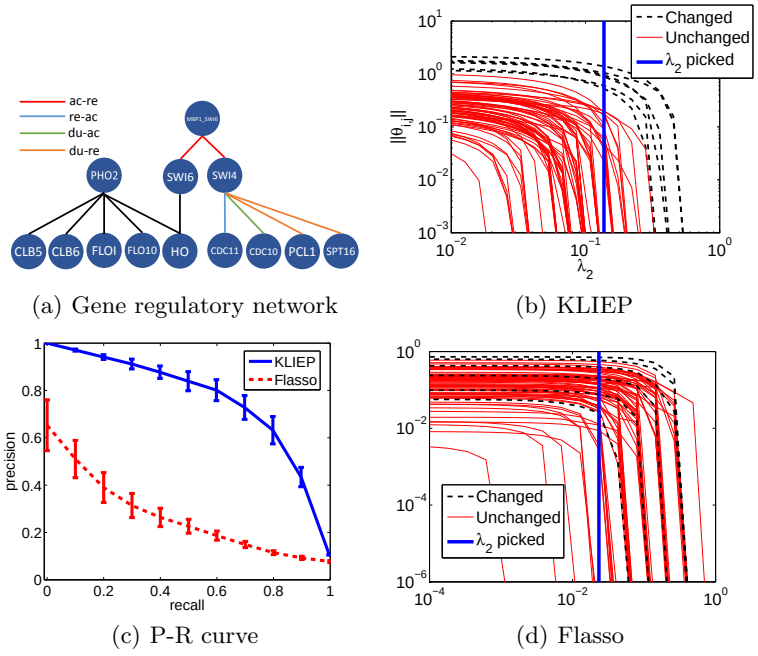
In this section, experiments are conducted on a synthetic gene expression dataset and on a Twitter dataset, respectively. We consider only the KLIEP and Flasso methods here. For KLIEP, the polynomial transform function with $k \in \{2, 3, 4\}$ is used. The parameter λ_1 in KLIEP and Flasso is tested with choices $\lambda_1 \in \{0.1, 1, 10\}$. The performance reported for the experiments in Section 5.1 and 5.2 are obtained using the models selected by HOLL and 5-fold CVLL (see Section 4.1), respectively.

5.1 Synthetic Gene Expression Dataset

A gene regulatory network encodes interactions between DNA segments. However, the way genes interact may change due to environmental or biological stimuli. In this experiment, we focus on detecting such changes. We use *SynTReN*, which is a generator of gene regulatory networks used as the benchmark validation of bioinformatics algorithms [22].

To test the applicability of the proposed method, we first choose a subnetwork containing 13 nodes from an existing signalling network in *Saccharomyces cerevisiae* (shown in Figure 5(a)). Three types of interactions are modelled: activation (ac), deactivation (re), and dual (du). 50 samples are generated in the first stage, after which we change the types of interactions in 6 edges, and generate 50 samples again. Four types of changes are considered in such case: ac $\rightarrow$ re, re $\rightarrow$ ac, du $\rightarrow$ ac, and du $\rightarrow$ re.

The regularization paths for KLIEP and Flasso are plotted in Figures 5(b) and 5(d). Averaged precision-recall curves over 20 simulation runs are shown in Figure 5(c). Clearly from the example of KLIEP regularization paths shown in Figure 5(d), the magnitude of estimated parameters on the changed pairwise interactions is much higher than that of the unchanged ones, and hits zero only at the final stage. On the other hand, Flasso gives many false alarms by assigning non-zero parameters to the unchanged interactions, even after some changed ones hit zeros. Reflecting a similar pattern, the P-R curves plot in Figure 5(c)



(a) Gene regulatory network (b) KLIEP

(c) P-R curve (d) Flasso

Fig. 5. Experiments on synthetic gene expression datasets.

show that the proposed KLIEP method achieves significant improvement over the Flasso method.

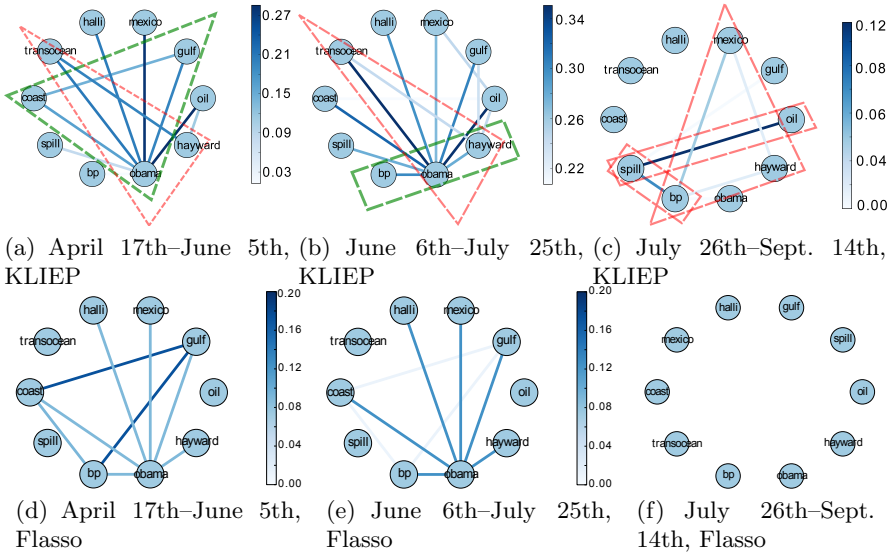
5.2 Twitter Story Telling

In this experiment, we use KLIEP and Flasso as event detectors from Twitter. More specifically, we choose the *Deepwater Horizon oil spill* [6] as the target event, and we hope that our method can recover some story lines from Twitter as the news event develops. Counting the frequencies of 10 keywords (BP, oil, spill, Mexico, gulf, coast, Hayward, Halliburton, Transocean, and Obama), we obtain a dataset by sampling 1061 times (4 per day), from February 1st, 2010 to October 15th, 2010.

To conduct our experiments, we segment the data into two parts. The first 300 samples collected before the day of oil spill (April 20th, 2010) are regarded as conforming to a 10-dimensional joint distribution Q, while the second set of samples that are drawn in an arbitrary 50-day window approximately after the event happened is regarded as following distribution P.

The MN of Q encodes the original conditional independence of frequencies between 10 keywords, and the underlying MN of P has changed since an event occurred. Thus, unveiling a change in MNs between P and Q may recover popular topic trends on Twitter in terms of the dependency among keywords.

[6] `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

Fig. 6. 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 1st to April 17th. Notable structures shared by the graph of both methods are surrounded by the green dashed lines. Unique structures that only appear in the graph of the proposed KLIEP method are surrounded by the red dashed lines.

The detected change graphs (i.e. the graphs with only detected changing edges) on 10 keywords are illustrated in Figure 6. The edges are selected at a certain value of λ_2 indicated by the maximal 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 for 100 times; then we observe detected edges by CVLL. Finally, we select the edges that are detected using the original non-shuffled dataset and remove those that were detected in the shuffled datasets for more than 5 times. In Figure 6, we plot detected change graphs which are generated using samples of P starting from April 17th, July 6th, and July 26th.

The initial explosion happened on April 20th, 2010. Both methods discover dependency changes between keywords. Generally speaking, KLIEP captures more conditional independence changes between keywords than the Flasso method, especially when comparing Figure 6(c) and Figure 6(f). At the first two stages (Figures 6(a), 6(b), 6(d) and 6(e)), 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 lies in the center of the news stories, and his media exposure peaks after his visit to the Louisiana coast (May 2nd, May 28nd, and June 5th) and his meeting with BP CEO Tony Hayward on June 16th.

Notably, both methods highlight the "gulf-obama-coast" triangle in Figures 6(a) and 6(d) and the "bp-obama-hayward" chain in Figures 6(b) and 6(e).

However, there are some important differences worth mentioning. First, the Flasso method misses the "transocean-hayward-obama" triangle in Figures 6(d) and 6(e). Transocean is the contracted operator in the Deepwater Horizon platform, where the initial explosion happened. On Figure 6(c), 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 steped down from the CEO position on July 27th.

6 Conclusion and Future Work

In this paper, 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 two MNs where changes can be naturally encoded as sparsity patterns in estimated parameters. Through experiments on artificial and real-world datasets, we demonstrated the usefulness of the proposed method.

Compared with the conventional two-stage MLE approach, a notable advantage of our method is that the normalization term in the density ratio model can be approximated by a sample average without sampling. This considerably loosens the restriction on applicable distributions. Moreover, thanks to its direct modeling nature with density ratios, the number of parameters is halved.

We only considered MNs with pairwise factors in this paper. However, such a model may be misspecified when higher order interactions exist. For example, combination with the idea *hierarchical log-linear model* presented in [16] may lead to a promising solution to this problem, which will be investigated in our future work.

Acknowledgement

SL is supported by the JST PRESTO program and the JSPS fellowship, JQ is supported by the JST PRESTO program, and MS is supported by the JST CREST program. MUG is supported by the Finnish Centre-of-Excellence in Computational Inference Research COIN (251170).

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York, NY, USA (2006)
2. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning **1**(1-2) (2008) 1–305
3. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)

4. Ravikumar, P., Wainwright, M.J., Lafferty, J.D.: High-dimensional ising model selection using ℓ_1-regularized logistic regression. The Annals of Statistics **38**(3) (2010) 1287–1319
5. Lee, S.I., Ganapathi, V., Koller, D.: Efficient structure learning of Markov networks using l_1-regularization. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2007) 817–824
6. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research **13** (2012) 307–361
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, USA (2001)
8. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. Journal of Machine Learning Research **9** (March 2008) 485–516
9. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3) (2008) 432–441
10. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(1) (2005) 91–108
11. Zhang, B., Wang, Y.: Learning structural changes of Gaussian graphical models in controlled experiments. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010). (2010) 701–708
12. Liu, H., Lafferty, J., Wasserman, L.: The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. The Journal of Machine Learning Research **10** (2009) 2295–2328
13. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: The nonparanormal skeptic. In: Proceedings of the 29th International Conference on Machine Learning (ICML2012). (2012)
14. Sugiyama, M., Suzuki, T., Kanamori, T.: Density Ratio Estimation in Machine Learning. Cambridge University Press, Cambridge, UK (2012)
15. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York, NY, USA (1998)
16. Schmidt, M.W., Murphy, K.P.: Convex structure learning in log-linear models: Beyond pairwise potentials. Journal of Machine Learning Research - Proceedings Track **9** (2010) 709–716
17. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. The Annals of Statistics **34**(3) (2006) 1436–1462
18. Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M.: Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics **60**(4) (2008) 699–746
19. Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., Sugiyama, M.: Direct density ratio estimation for large-scale covariate shift adaptation. Journal of Information Processing **17** (2009) 138–155
20. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B **67**(2) (2005) 301–320
21. Neal, R.M.: Slice sampling. The Annals of Statistics **31**(3) (2003) 705–741
22. Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., Marchal, K.: SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics **7**(1) (2006) 43