

Structural Change Detection by Sparse Density Ratio Estimation

Masashi Sugiyama

Graduate School of Frontier Sciences,
The University of Tokyo
sugi@cs.titech.ac.jp
<http://www.ms.k.u-tokyo.ac.jp>

Abstract

The objective of *change detection* is to investigate whether change exists between two data sets $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$. In this paper, we explore methods of *structural change detection*, which are aimed at analyzing change in the *dependency structure* between elements of d -dimensional variable $\mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top$.

1 Sparse Maximum Likelihood Estimation

Let us consider a *Gaussian Markov network*, which is a d -dimensional Gaussian model with expectation zero:

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

where not the variance-covariance matrix, but its inverse called the *precision matrix* is parameterized by Θ . If Θ is regarded as an *adjacency matrix*, the Gaussian Markov network can be visualized as a *graph* (see Figure 1). An advantage of this precision-based parameterization is that the connectivity governs conditional independence. For example, in the

Gaussian Markov network illustrated in the left-hand side of Figure 1, $x^{(1)}$ and $x^{(2)}$ are connected via $x^{(3)}$. This means that $x^{(1)}$ and $x^{(2)}$ are conditionally independent given $x^{(3)}$.

Suppose that $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$ are drawn independently from the Gaussian Markov networks with precision matrices Θ and Θ' , respectively. Then analyzing $\Theta - \Theta'$ allows us to identify change in Markov network structure (see Figure 1 again).

A sparse estimate of Θ may be obtained by maximum likelihood estimation with the ℓ_1 -constraint:

$$\max_{\Theta} \sum_{i=1}^n \log q(\mathbf{x}_i; \Theta) \quad \text{subject to } \|\Theta\|_1 \leq R^2,$$

where $R \geq 0$ is the radius of the ℓ_1 -ball. This method is also referred to as the *graphical lasso* [2].

The derivative of $\log q(\mathbf{x}; \Theta)$ with respect to Θ is given by

$$\frac{\partial \log q(\mathbf{x}; \Theta)}{\partial \Theta} = \frac{1}{2} \Theta^{-1} - \frac{1}{2} \mathbf{x} \mathbf{x}^\top,$$

where the following formulas are used for its derivation:

$$\frac{\partial \log \det(\Theta)}{\partial \Theta} = \Theta^{-1} \quad \text{and} \quad \frac{\partial \mathbf{x}^\top \Theta \mathbf{x}}{\partial \Theta} = \mathbf{x} \mathbf{x}^\top.$$

A MATLAB code of a gradient-projection algorithm of ℓ_1 -constraint maximum likelihood estimation for Gaussian Markov networks is given in Figure 2, where projection onto the ℓ_1 -ball is computed by the method developed in [1].

For the true precision matrices

$$\Theta = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix} \quad \text{and} \quad \Theta' = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

sparse maximum likelihood estimation gives

$$\hat{\Theta} = \begin{pmatrix} 1.382 & 0 & 0.201 \\ 0 & 1.788 & 0 \\ 0.201 & 0 & 1.428 \end{pmatrix} \quad \text{and} \quad \hat{\Theta}' = \begin{pmatrix} 1.617 & 0 & 0 \\ 0 & 1.711 & 0 \\ 0 & 0 & 1.672 \end{pmatrix}.$$

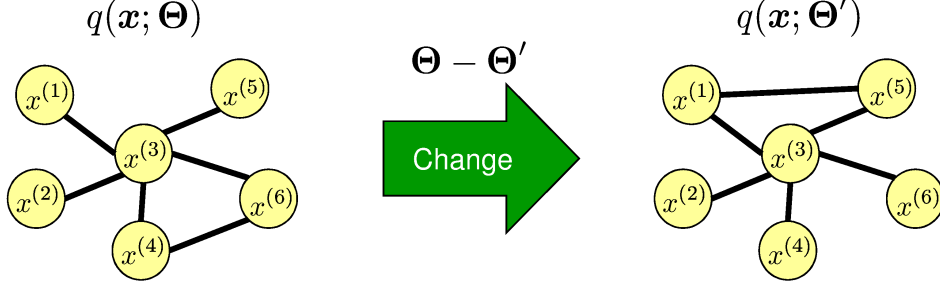


Figure 1: Structural change in Gaussian Markov networks.

Thus, the true sparsity patterns of Θ and Θ' (in off-diagonal elements) can be successfully recovered. Since

$$\Theta - \Theta' = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \hat{\Theta} - \hat{\Theta}' = \begin{pmatrix} -0.235 & 0 & 0.201 \\ 0 & 0.077 & 0 \\ 0.201 & 0 & -0.244 \end{pmatrix},$$

change in sparsity patterns (in off-diagonal elements) can be correctly identified.

On the other hand, when the true precision matrices are

$$\Theta = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \quad \text{and} \quad \Theta' = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

sparse maximum likelihood estimation gives

$$\hat{\Theta} = \begin{pmatrix} 1.303 & 0.348 & 0 \\ 0.348 & 1.157 & 0.240 \\ 0 & 0.240 & 1.365 \end{pmatrix} \quad \text{and} \quad \hat{\Theta}' = \begin{pmatrix} 1.343 & 0 & 0.297 \\ 0 & 1.435 & 0.236 \\ 0.297 & 0.236 & 1.156 \end{pmatrix}.$$

Thus, the true sparsity patterns of Θ and Θ' can still be successfully recovered. However, since

$$\Theta - \Theta' = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \hat{\Theta} - \hat{\Theta}' = \begin{pmatrix} -0.040 & 0.348 & -0.297 \\ 0.348 & -0.278 & 0.004 \\ -0.297 & 0.004 & 0.209 \end{pmatrix},$$

change in sparsity patterns was not correctly identified. This shows that, when a non-zero unchanged edge exists, say $\Theta_{k,k'} = \Theta'_{k,k'} > 0$ for some

```

TT=[2 0 1; 0 2 0; 1 0 2];
%TT=[2 0 0; 0 2 0; 0 0 2];
%TT=[2 1 0; 1 2 1; 0 1 2];
%TT=[2 0 1; 0 2 1; 1 1 2];
d=3; n=50; x=TT^(-1/2)*randn(d,n); S=x*x'/n;
T0=eye(d); C=5; e=0.1;
for o=1:100000
    T=T0+e*(inv(T0)-S);
    T(:)=L1BallProjection(T(:),C);
    if norm(T-T0)<0.00000001, break, end
    T0=T;
end
T, TT

```

```

function w=L1BallProjection(x,C)

u=sort(abs(x),'descend'); s=cumsum(u);
r=find(u>(s-C)./(1:length(u))',1,'last');
w=sign(x).*max(0,abs(x)-max(0,(s(r)-C)/r));

```

Figure 2: MATLAB code of a gradient-projection algorithm of ℓ_1 -constraint maximum likelihood estimation for Gaussian Markov networks. The bottom function should be saved as “L1BallProjection.m”.

k and k' , it is difficult to identify this unchanged edge because $\hat{\Theta}_{k,k'} \approx \hat{\Theta}'_{k,k'}$ does not necessarily hold by separate sparse maximum likelihood estimation from $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$.

2 Sparse Density Ratio Estimation

As illustrated above, sparse maximum likelihood estimation can perform poorly in structural change detection. Another limitation of sparse maximum likelihood estimation is the Gaussian assumption. A Gaussian Markov network can be extended to a non-Gaussian model as

$$q(\mathbf{x}; \boldsymbol{\theta}) = \frac{\bar{q}(\mathbf{x}; \boldsymbol{\theta})}{\int \bar{q}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}},$$

where, for a *feature vector* $\mathbf{f}(x, x')$,

$$\bar{q}(\mathbf{x}; \boldsymbol{\theta}) = \exp \left(\sum_{k \geq k'} \boldsymbol{\theta}_{k,k'}^\top \mathbf{f}(x^{(k)}, x^{(k')}) \right).$$

This model is reduced to the Gaussian Markov network if

$$\mathbf{f}(x, x') = -\frac{1}{2}xx',$$

while higher-order correlations can be captured by considering higher-order terms in the feature vector. However, applying sparse maximum likelihood estimation to non-Gaussian Markov networks is not straightforward in practice because the normalization term $\int \bar{q}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$ is often computationally intractable.

To cope with these limitations, let us handle the change in parameters, $\boldsymbol{\theta}_{k,k'} - \boldsymbol{\theta}'_{k,k'}$, directly via the following density ratio function:

$$\frac{q(\mathbf{x}; \boldsymbol{\theta})}{q(\mathbf{x}; \boldsymbol{\theta}')} \propto \exp \left(\sum_{k \geq k'} (\boldsymbol{\theta}_{k,k'} - \boldsymbol{\theta}'_{k,k'})^\top \mathbf{f}(x^{(k)}, x^{(k')}) \right).$$

Based on this expression, let us consider the following density ratio model:

$$r(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\exp \left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^\top \mathbf{f}(x^{(k)}, x^{(k')}) \right)}{\int p'(\mathbf{x}) \exp \left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^\top \mathbf{f}(x^{(k)}, x^{(k')}) \right) d\mathbf{x}},$$

where $\boldsymbol{\alpha}_{k,k'}$ is the difference of parameters:

$$\boldsymbol{\alpha}_{k,k'} = \boldsymbol{\theta}_{k,k'} - \boldsymbol{\theta}'_{k,k'},$$

Then let us learn the parameters $\{\boldsymbol{\alpha}_{k,k'}\}_{k \geq k'}$ by *group-sparse* maximum

```

Tp=[2 0 1; 0 2 0; 1 0 2]; Tq=[2 0 0; 0 2 0; 0 0 2];
Tp=[2 1 0; 1 2 1; 0 1 2]; Tq=[2 0 1; 0 2 1; 1 1 2];
d=3; n=50; xp=Tp^(-1/2)*randn(d,n); Sp=xp*xp'/n;
xq=Tq^(-1/2)*randn(d,n); A0=eye(d); C=1; e=0.1;
for o=1:1000000
    U=exp(sum((A0*xq).*xq));
    A=A0-e*(( repmat(U,[d 1]).*xq)*xq'/sum(U)-Sp);
    A(:)=L1BallProjection(A(:),C);
    if norm(A-A0)<0.00000001, break, end
    A0=A;
end
-2*A, Tp-Tq

```

Figure 3: MATLAB code of a gradient-projection algorithm of ℓ_1 -constraint Kullback-Leibler density ratio estimation for Gaussian Markov networks. “L1BallProjection.m” is given in Figure 2.

likelihood estimation [6, 5, 3]:

$$\begin{aligned}
& \min_{\{\alpha_{k,k'}\}_{k \geq k'}} \log \frac{1}{n'} \sum_{i'=1}^{n'} \exp \left(\sum_{k \geq k'} \alpha_{k,k'}^\top \mathbf{f}(x_{i'}^{(k)}, x_{i'}^{(k')}) \right) \\
& \quad - \frac{1}{n} \sum_{i=1}^n \sum_{k \geq k'} \alpha_{k,k'}^\top \mathbf{f}(x_i^{(k)}, x_i^{(k')}) \\
& \quad \text{subject to } \sum_{k \geq k'} \|\alpha_{k,k'}\| \leq R^2,
\end{aligned}$$

where $R \geq 0$ controls the sparseness of the solution. Support consistency of this sparse density ratio estimator has been theoretically investigated in [4].

A MATLAB code of a gradient-projection algorithm of sparse Kullback-Leibler density ratio estimation for Gaussian Markov networks is given in Figure 3. For the true precision matrices

$$\Theta - \Theta' = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

sparse Kullback-Leibler density ratio estimation gives

$$\begin{pmatrix} 0 & 0 & 1.000 \\ 0 & 0 & 0 \\ 1.000 & 0 & 0 \end{pmatrix}.$$

This implies that change in sparsity patterns can be correctly identified.

Even when non-zero unchanged edges exist as

$$\Theta - \Theta' = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} - \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

sparse Kullback-Leibler density ratio estimation gives

$$\begin{pmatrix} 0 & 0.707 & -0.293 \\ 0.707 & 0 & 0 \\ -0.293 & 0 & 0 \end{pmatrix}.$$

Thus, change in Markov network structure can still be correctly identified.

References

- [1] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1 -ball for learning in high dimensions. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 272–279. Omnipress, 2008.
- [2] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [3] S. Liu, J. Quinn, M. U. Gutmann, and M. Sugiyama. Direct learning of sparse changes in Markov networks by density ratio estimation. *Neural Computation*, 26(6):1169–1197, 2014.
- [4] S. Liu, T. Suzuki, and M. Sugiyama. Support consistency of direct sparse-change learning in Markov networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI2015)*, pages xxx–xxx, Austin, Texas, USA, Jan. 25–29 2015. The AAAI Press.

- [5] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.
- [6] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Büna, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Graduate School of Frontier Sciences, The University of Tokyo
Tokyo 113-0033
JAPAN
E-mail address: sugi@k.u-tokyo.ac.jp

東京大学・大学院新領域創成科学研究科 杉山 将