ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 4, *pp.* 1187–1206

Change point estimation in high dimensional Markov random-field models

Sandipan Roy,

University College London, UK

Yves Atchadé

University of Michigan, Ann Arbor, USA

and George Michailidis

University of Florida, Gainesville, USA

[Received May 2014. Final revision July 2016]

Summary. The paper investigates a change point estimation problem in the context of high dimensional Markov random-field models. Change points represent a key feature in many dynamically evolving network structures. The change point estimate is obtained by maximizing a profile penalized pseudolikelihood function under a sparsity assumption. We also derive a tight bound for the estimate, up to a logarithmic factor, even in settings where the number of possible edges in the network far exceeds the sample size. The performance of the estimator proposed is evaluated on synthetic data sets and is also used to explore voting patterns in the US Senate in the 1979–2012 period.

Keywords: Change point analysis; High dimensional inference; Markov random fields; Network analysis; Profile pseudolikelihood

1. Introduction

Networks are capable of capturing dependence relationships and have been extensively employed in diverse scientific fields including biology, economics and the social sciences. A rich literature has been developed for static networks leveraging advances in estimating sparse graphical models. However, increasing availability of data sets that evolve over time has accentuated the need for developing models for time varying networks. Examples of such data sets include time course gene expression data and voting records of legislative bodies.

In this work, we consider modelling the underlying network through a Markov random field that exhibits a change in its structure at some point in time. Specifically, suppose that we have T observations $\{X^{(t)}, 1 \leqslant t \leqslant T\}$ over p-variables with $X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ and $X_j^{(t)} \in \mathbf{X}$, for some finite set $\mathbf{X}$. Further, we assume that there is a time point $\tau_* = \lceil \alpha_* T \rceil \in \{1, \ldots, T-1\}$, with $\alpha_* \in (0,1)$, such that $\{X^{(t)}, 1 \leqslant t \leqslant \tau_*\}$ is an independent and identically distributed (IID) sequence from a distribution $g_{\theta_*^{(1)}}(\cdot)$ parameterized by a real symmetric matrix $\theta_*^{(1)}$, whereas the remaining observations $\{X^{(t)}, \tau_* + 1 \leqslant t \leqslant T\}$ form also an independent and identically distributed sequence from a distribution $g_{\theta_*^{(2)}}(\cdot)$ parameterized by another real symmetric matrix

Address for correspondence: George Michailidis, Department of Statistics and Informatics Institute, University of Florida, 205 Griffin-Floyd Hall, PO Box 118545, Gainesville, FL 32611-8545, USA.
E-mail: gmichail@umich.edu

1369–7412/17/791187

$\theta_*^{(2)}$. We assume that the two distributions $g_{\theta_*^{(1)}}(\cdot)$ and $g_{\theta_*^{(2)}}(\cdot)$ belong to a parametric family of Markov random-field distributions given by

$$g_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{j=1}^{p} \theta_{jj} B_0(x_j) + \sum_{1 \leqslant k < j \leqslant p} \theta_{jk} B(x_j, x_k) \right\}, \qquad x \in \mathbf{X}^p, \tag{1}$$

for a non-zero function $B_0 : \mathbf{X} \rightarrow \mathbb{R}$, and a non-zero symmetric function $B : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ which encodes the interactions between the nodes. The term $Z(\boldsymbol{\theta})$ is the corresponding normalizing constant. Thus, the observations over time come from a Markov random field that exhibits a change in its structure at time τ_* and the matrices $\theta_*^{(1)}$ and $\theta_*^{(2)}$ encode the conditional independence structure between the p random variables respectively before and after the change point.

The objective is to estimate the change point τ_*, as well as the *sparse* network structures $\theta_*^{(1)}$ and $\theta_*^{(2)}$. Although the problem of identifying a change point has a long history in statistics (see Bai (2010), Carlstein (1988), Hinkley (1970), Loader (1996), Lan *et al.* (2009), Muller (1992), Raimondo (1998) and references therein), its use in a high dimensional network problem is novel and motivated by the US Senate voting record application that is discussed in Section 6. In a low dimensional setting, the results that are obtained for the change point depend on the regime that is considered; specifically, if there is a fixed shift, then the asymptotic distribution of the change point is given by the minimizer of a compound Poisson process (see Kosorok (2008)), whereas, if the shift decreases to 0 as a function of the sample size, the distribution corresponds to that of Brownian motion with triangular drift (see Bhattacharya (1987) and Muller (1992)).

The methodology that is developed in this paper is useful in other areas, where similar problems occur. Examples include biological settings, where a gene regulatory network may exhibit a significant change at a particular dose of a drug treatment, or in finance, where major economic announcements may disrupt financial networks.

Estimation of time invariant networks from independent and identically distributed data based on the Markov random-field model has been a very active research area (see for example Banerjee *et al.* (2008), Höfling and Tibshirani (2009), Ravikumar *et al.* (2010), Xue *et al.* (2012), Guo *et al.* (2010) and references therein). Sparsity (which is an often realistic assumption in many applications in molecular biology, chemoinformatics, climate modelling, finance, etc.) plays an important role in this literature and allows the recovery of the underlying network with relatively few observations (Ravikumar *et al.*, 2010; Guo *et al.*, 2010). In case the sparsity assumption does not hold exactly for the specific sample size and number of variables under consideration in a real application, the various sparse estimation procedures that are available in the literature will nevertheless estimate the strongest statistical relationships supported by the data.

However, there is significantly less work on time varying networks (see Zhou *et al.* (2010), Kolar *et al.* (2010), Kolar and Xing (2012) etc.). The closest setting to the current paper is the work in Kolar and Xing (2012), which considered Gaussian graphical models where *each* node can exhibit multiple change points. In contrast, this paper focuses on a *single* change point impacting the global network structure of the underlying Markov random field. In general, which setting is more appropriate depends on the application. In biological applications where the focus is on particular biomolecules (e.g. genes, proteins or metabolites), nodewise change point analysis would typically be preferred, whereas in many social network applications (such as the political network example that is considered below) global structural changes in the network are of primary interest. Further, note that node level changes detected at multiple nodes can be inconsistent, noisy and difficult to reconcile to extract global structural changes.

Another key difference between these two papers is the modelling framework that is employed.

Specifically, in Kolar and Xing (2012) the number of nodes in the Gaussian graphical model is *fixed* and *smaller* than the available sample size. The high dimensional challenge comes from the possible presence of multiple change points per node, which leads to a large number of parameters to be estimated. To overcome this issue, a total variation penalty is introduced, which is a strategy that has worked well in regression modelling where the number of parameters is the same as the number of observations. In contrast, this paper assumes a high dimensional framework where the number of nodes (and hence the number of parameters of interest, namely the edges) grows with the number of time points and focuses on estimating a single change point in a general Markov random-field model.

To avoid the intractable normalizing constant issue in estimating the network structures, we employ a pseudolikelihood framework. As customary in the analysis of change point problems (Bai, 2010; Lan *et al.*, 2009), we employ a profile pseudolikelihood function to obtain the estimate $\hat{\tau}$ of the true change point τ_*. Under a sparsity assumption, and some regularity conditions that allow the number of parameters $p(p+1)$ to be much larger than the sample size T, we establish that, with high probability, $|\hat{\tau}/T-\alpha_*|=O\{\log(pT)/T\}$, as $p, T\to\infty$. In classical change point problems with a fixed magnitude change, it is well known that the maximum likelihood estimator of the change point satisfies $|\hat{\tau}/T-\alpha_*|=O_p(1/T)$ (see for example Bhattacharya (1987) and Bai (2010)). This suggests that our result is rate optimal, up to the logarithm factor $\log(T)$. Since the appearance of the initial version of this paper, we note that there has been additional work on the subject of change point estimation problems in high dimensional settings (Soh and Chandrasekaran, 2014; Leonardi and Bühlmann, 2016). Both Soh and Chandrasekaran (2014) and Leonardi and Bühlmann (2016) focus on the linear regression case, but they also consider multiple change points. However, the convergence rate for the change point parameters, even for the case of a single change point, is slower than the rate $O\{\log(pT)/T\}$ that is derived here.

The derivation of the result requires a careful handling of model misspecification in Markov random fields as explained in Section 3, which is a novel aspect that is not present when estimating a single Markov random field from IID observations. See also Atchadé (2014) for another example of misspecification in Markov random fields. Further, to speed up the computation of the change point estimator $\hat{\tau}$, we discuss a sampling strategy of the available observations, coupled with a smoothing procedure of the resulting likelihood function.

Last, but not least, we employ the methodology that is developed to analyse the US Senate voting record from 1979 to 2012. In this application, each Senate seat represents a node of the network and the voting record of these 100 Senate seats on a given bill is viewed as a realization of an underlying Markov random field that captures dependences between them. The analysis strongly points to a change point around January 1995, which was the beginning of the tenure of the 104th Congress. This change point comes in the footsteps of the November 1994 election that witnessed the Republican Party's capturing the US House of Representatives for the first time since 1956. Other analyses based on more *ad hoc* methods also point to a significant change occurring after the November 1994 election (e.g. Moody and Mucha (2013)).

The remainder of the paper is organized as follows. Modelling assumptions and the estimation framework are presented in Section 2, whereas Section 3 establishes the key technical results. Section 4 discusses computational issues and Section 5 evaluates the performance of the estimation procedure by using synthetic data. Section 6 illustrates the procedure on the US Senate voting record. Finally, proofs are deferred to the on-line supplement.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Methodology

Let $\{X^{(t)}, 1 \leqslant t \leqslant T\}$ be a sequence of independent random vectors, where $X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ is a p-dimensional Markov random field whose jth component $X_j^{(t)}$ takes values in a finite set $\mathbf{X}$. We assume that there is a time point (change point) $\tau_* \in \{1, \ldots, T-1\}$ and symmetric matrices $\theta_*^{(1)}, \theta_*^{(2)} \in \mathbb{R}^{p \times p}$, such that, for all $x \in \mathbf{X}^p$,

$$\mathbb{P}(X^{(t)} = x) = g_{\theta_*^{(1)}}(x), \qquad \text{for } t = 1, \ldots, \tau_*,$$

and

$$\mathbb{P}(X^{(t)} = x) = g_{\theta_*^{(2)}}(x), \qquad \text{for } t = \tau_* + 1, \ldots, T,$$

where g_θ is the Markov random-field distribution given in equation (1). We assume without any loss of generality that $\tau_* = \lceil \alpha_* T \rceil$, for some $\alpha_* \in (0, 1)$, where $\lceil x \rceil$ denotes the smallest integer that is larger than or equal to x. The likelihood function of the observations $\{X^{(t)}, 1 \leqslant t \leqslant T\}$ is then given by

$$L_T(\tau, \theta^{(1)}, \theta^{(2)} | X^{(1:T)}) = \prod_{t=1}^{\tau} g_{\theta^{(1)}}(X^{(t)}) \prod_{t=\tau+1}^{T} g_{\theta^{(2)}}(X^{(t)}). \tag{2}$$

We write $\mathbb{E}$ to denote the expectation operator with respect to $\mathbb{P}$. For a symmetric matrix $\theta \in \mathbb{R}^{p \times p}$, we write $\mathbb{P}_\theta$ to denote the probability distribution on $\mathbf{X}^p$ with probability mass function g_θ and $\mathbb{E}_\theta$ its expectation operator.

We are interested in estimating both the change point τ_*, as well as the parameters $\theta_*^{(1)}$ and $\theta_*^{(2)}$. Let $\mathcal{M}_p$ be the space of all $p \times p$ real symmetric matrices. We equip $\mathcal{M}_p$ with the Frobenius inner product

$$\langle \theta, \vartheta \rangle_{\mathrm{F}} \overset{\mathrm{def}}{=} \sum_{k \leqslant j} \theta_{jk} \vartheta_{jk},$$

and the associated norm

$$\|\theta\|_{\mathrm{F}} \overset{\mathrm{def}}{=} \sqrt{\langle \theta, \theta \rangle}.$$

This is equivalent to identifying $\mathcal{M}_p$ with the Euclidean space $\mathbb{R}^{p(p+1)/2}$, and this identification prevails whenever we define gradients and Hessians of functions $f : \mathcal{M}_p \to \mathbb{R}$. For $\theta \in \mathcal{M}_p$ we also define

$$\|\theta\|_1 \overset{\mathrm{def}}{=} \sum_{k \leqslant j} |\theta_{jk}|,$$

and

$$\|\theta\|_\infty \overset{\mathrm{def}}{=} \sup_{k \leqslant j} |\theta_{jk}|.$$

If $u \in \mathbb{R}^d$, for some $d \geqslant 1$, and A is an ordered subset of $\{1, \ldots, d\}$, we define

$$u_A \overset{\mathrm{def}}{=} (u_j, j \in A),$$

and u_{-j} is a short cut for $u_{\{1,\ldots,d\} \setminus \{j\}}$.

To avoid some of the computational difficulties in dealing with the normalizing constant of g_θ, we take a pseudolikelihood approach. For $\theta \in \mathcal{M}_p$ and $j \in \{1, 2, \ldots, p\}$, define

$$f_\theta^{(j)}(u|x) \overset{\mathrm{def}}{=} \mathbb{P}_\theta(X_j = u | X_{-j} = x_{-j}),$$

for $u \in \mathbf{X}$, and $x \in \mathbf{X}^p$. From the expression of the joint distribution g_θ in equation (1), we have

$$f_\theta^{(j)}(u|x) = \frac{1}{Z_\theta^{(j)}(x)} \exp\Big\{\theta_{jj} B_0(u) + \sum_{k \neq j} \theta_{jk} B(u, x_k)\Big\}, \qquad u \in \mathbf{X}, \quad x \in \mathbf{X}^p, \tag{3}$$

where

$$Z_\theta^{(j)}(x) \overset{\text{def}}{=} \int_{\mathbf{X}} \exp\Big\{\theta_{jj} B_0(z) + \sum_{k \neq j} \theta_{jk} B(z, x_k)\Big\} \mathrm{d}z. \tag{4}$$

Remark 1. The normalizing constant $Z_\theta^{(j)}(x)$ defined in equation (4) is actually a summation over $\mathbf{X}$, but for notational convenience we write it as an integral against the counting measure on $\mathbf{X}$. Furthermore, it is implicitly assumed that these normalizing constants are available in closed form, which is so for most commonly used graphical models. For instance, in the case of the Ising model that is used below, $\mathbf{X} = \{0, 1\}$, $B_0(z) = z$ and $B(z, y) = zy$, so $Z_\theta^{(j)}(x)$ is explicitly given by

$$Z_\theta^{(j)}(x) = 1 + \exp\Big(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_k\Big).$$

Next, we introduce

$$\phi(\theta, x) \overset{\text{def}}{=} -\sum_{j=1}^{p} \log\{f_\theta^{(j)}(x_j|x)\}. \tag{5}$$

The negative log-pseudolikelihood of the model (divided by T) is given by

$$l_T(\tau; \theta_1, \theta_2) \overset{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{\tau} \phi(\theta_1, X^{(t)}) + \frac{1}{T} \sum_{t=(\tau+1)}^{T} \phi(\theta_2, X^{(t)}). \tag{6}$$

For $1 \leqslant \tau < T$, and $\lambda > 0$, we define the estimators

$$\hat{\boldsymbol{\theta}}_{1,\tau}^{(\lambda)} \overset{\text{def}}{=} \underset{\theta \in \mathcal{M}_p}{\arg\min} \frac{1}{T} \sum_{t=1}^{\tau} \phi(\theta, X^{(t)}) + \lambda \|\boldsymbol{\theta}\|_1,$$

and

$$\hat{\boldsymbol{\theta}}_{2,\tau}^{(\lambda)} \overset{\text{def}}{=} \underset{\theta \in \mathcal{M}_p}{\arg\min} \frac{1}{T} \sum_{t=\tau+1}^{T} \phi(\theta, X^{(t)}) + \lambda \|\boldsymbol{\theta}\|_1.$$

We propose to estimate the change point τ_* by using a profile pseudolikelihood approach. More precisely our estimator $\hat{\tau}$ is defined as

$$\hat{\tau} = \underset{\tau \in \mathcal{T}}{\arg\min}\, l_T(\tau; \hat{\boldsymbol{\theta}}_{1,\tau}, \hat{\boldsymbol{\theta}}_{2,\tau}), \tag{7}$$

for a search domain $\mathcal{T} \subset \{1, \ldots, T\}$ of the form $\{k_l, k_l + 1, \ldots, T - k_u\}$, where, for each $\tau \in \mathcal{T}$, $\hat{\boldsymbol{\theta}}_{1,\tau} = \hat{\theta}_{1,\tau}^{(\lambda_{1,\tau})}$ and $\hat{\boldsymbol{\theta}}_{2,\tau} = \hat{\theta}_{2,\tau}^{(\lambda_{2,\tau})}$, for some positive penalty parameters $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$. Since the network estimation errors at the boundaries of the time line $\{1, \ldots, T\}$ are typically large, a restriction on the search domain is needed to guarantee the consistency of the method. This motivates the introduction of $\mathcal{T}$. We give more details on $\mathcal{T}$ below. The penalty parameters $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ also play an important role in the behaviour of the estimators, and we provide some guidelines below.

3. Theoretical results

The recovery of τ_* rests on the ability of the estimators $\hat{\boldsymbol{\theta}}_{j,\tau}$ to estimate $\theta_*^{(j)}$, $j \in \{1,2\}$ correctly. Estimators for the static version of the problem where one has IID observations from a single Markov random field have been extensively studied; see Guo *et al.* (2010), Höfling and Tibshirani (2009), Meinshausen and Bühlmann (2006), Ravikumar *et al.* (2010) and references therein for computational and theoretical details. However, in the present setting one of the estimators $\hat{\theta}_{j,\tau}$, $j \in \{1,2\}$, is derived from a misspecified model. Hence, to establish the error bound for $\|\hat{\theta}_{j,\tau} - \theta_*^{(j)}\|_2$, we borrow from the approach in Atchadé (2014). For penalty terms $\lambda_{j,\tau}$ as in equation (8) below and under some regularity assumptions, we derive a bound on the estimator errors $\|\hat{\theta}_{j,\tau} - \theta_*^{(j)}\|_2$, for all $\tau \in \mathcal{T}$. We then use this result to show that the profile pseudo-log-likelihood estimator $\hat{\tau}$ is an approximate minimizer of $\tau \mapsto l_T(\tau; \theta_*^{(1)}, \theta_*^{(2)})$ and this allows us to establish a bound on the distance between $\hat{\tau}$ and the true change point τ_*.

We assume that the penalty parameters take the following specific form:

$$\begin{aligned} \lambda_{1,\tau} &= \frac{32 c_0 \surd\{\tau \log(dT)\}}{T}, \\ \lambda_{2,\tau} &= \frac{32 c_0 \surd\{(T-\tau) \log(dT)\}}{T}, \end{aligned} \tag{8}$$

where $d \overset{\text{def}}{=} p(p+1)/2$, and

$$c_0 = \sup_{u,v \in \mathbf{X}} |B_0(u) - B_0(v)| \vee \sup_{x,u,v \in \mathbf{X}} |B(x,u) - B(x,v)|, \tag{9}$$

which serves as (an upper bound on the) standard deviation of the random variables $B_0(X)$ and $B(X,Y)$. In practice, we use $\lambda_{1,\tau} = a_1 T^{-1} c_0 \surd\{\tau \log(dT)\}$ and $\lambda_{2,\tau} = a_2 T^{-1} c_0 \surd\{(T-\tau)\log(dT)\}$, where a_1 and a_2 are chosen from the data by an analogue of the Bayesian information criterion (Schwarz, 1978).

For $j = 1, 2$, define

$$\mathcal{A}_j \overset{\text{def}}{=} \{1 \leqslant k \leqslant i \leqslant p : \theta_{*ik}^{(j)} \neq 0\},$$

and define $s_j \overset{\text{def}}{=} |\mathcal{A}_j|$ the cardinality (and hence the sparsity) of the true model parameters. We also define

$$\mathbb{C}_j \overset{\text{def}}{=} \Big\{ \boldsymbol{\theta} \in \mathcal{M}_p : \sum_{(k,i) \in \mathcal{A}_j^c} |\theta_{ik}^{(j)}| \leqslant 3 \sum_{(k,i) \in \mathcal{A}_j} |\theta_{ik}^{(j)}| \Big\}, \qquad j \in \{1,2\}, \tag{10}$$

used next in the definition of the restricted strong convexity assumption.

Assumption 1 (restricted strong convexity). For $j \in \{1,2\}$, and $X \sim g_{\theta_*^{(j)}}$, there exists $\rho_j > 0$ such that, for all $\Delta \in \mathbb{C}_j$,

$$\sum_{i=1}^{p} \mathbb{E}_{\theta_*^{(j)}} \Big[\mathrm{var}_{\theta_*^{(j)}} \Big\{ \sum_{k=1}^{p} \Delta_{ik} B_{ik}(X_i, X_k) | X_{-i} \Big\} \Big] \geqslant 2 \rho_j \|\Delta\|_2^2, \tag{11}$$

where $B_{ik}(x,y) = B_0(x)$ if $i = k$, and $B_{ik}(x,y) = B(x,y)$ if $i \neq k$.

Remark 2. Assumption 1 is an (averaged) restricted strong convexity assumption on the negative log-pseudolikelihood function $\phi(\theta, x)$. This can be seen by noting that condition (11) can also be written as

$$\Delta' \mathbb{E}[\nabla^{(2)} \phi(\theta_*^{(j)}, X^{(j)})] \Delta \geqslant 2 \rho_j \|\Delta\|_2^2, \qquad X^{(j)} \sim g_{\theta_*^{(j)}}, \quad \Delta \in \mathbb{C}_j, \quad j \in \{1,2\}.$$

These restricted strong convexity assumptions of objective functions are more pertinent in high dimensional problems and appear in one form or another in the analysis of high dimensional statistical methods (see for example Neghaban *et al.* (2010) and references therein). Note that the restricted strong convexity assumption is expressed here in expectation, unlike Neghaban *et al.* (2010) which used an 'almost sure' version. Imposing this assumption in expectation (i.e. at the population level) is more natural and is known to imply the almost sure version in many instances (see Rudelson and Zhou (2013) and lemma 4 in the on-line supplement).

We impose the following condition on the change point and the sample size.

Assumption 2 (sample size requirement). We assume that there exists $\alpha_* \in (0,1)$ such that $\tau_* = \lceil \alpha_* T \rceil \in \{1,\ldots,T-1\}$, and the sample size T satisfies

$$\min\left\{\frac{T}{2^{11}\log(pT)}, \frac{T}{48^2 \times 32^2 \log(dT)}\right\} \geqslant c_0^2 \max\left\{\frac{s_1^2}{\alpha_* \rho_1^2}, \frac{s_2^2}{(1-\alpha_*)\rho_2^2}\right\},$$

where ρ_1 and ρ_2 are as in assumption 1.

Remark 3. The constants 2^{11} and $48^2 \times 32^2$ that are required in assumption 2 will typically yield a very conservative bound on the sample size T. We believe that these large constants are mostly artefacts of our techniques and can be improved. The key point of assumption 2 is the fact that we require the sample T to be such that $T/\log(T)$ is a linear function of $\max(s_1^2, s_2^2)\log(p)$. Up to the $\log(T)$-term, this condition is in agreement with recent results on high dimensional sparse graphical model recovery.

The ability to detect the change point requires that the change from $\theta_*^{(1)}$ to $\theta_*^{(2)}$ be identifiable.

Assumption 3 (identifiability condition). Assume that $\theta_*^{(1)} \neq \theta_*^{(2)}$, and

$$\kappa \overset{\text{def}}{=} \min(\mathbb{E}_{\theta_*^{(2)}}[\phi(\theta_*^{(1)}, X) - \phi(\theta_*^{(2)}, X)], \mathbb{E}_{\theta_*^{(1)}}[\phi(\theta_*^{(2)}, X) - \phi(\theta_*^{(1)}, X)]) > 0. \tag{12}$$

Remark 4. Assumption 3 is needed for the identifiability of the change point τ_*. Since the distributions g_θ are discrete data analogues of Gaussian graphical distributions, it is informative to look at assumption 3 for Gaussian graphical distributions. Indeed, if g_θ is the density of the p-dimensional normal distribution $N(0, \theta^{-1})$ with precision matrix θ and, if we take $\phi(\theta, x) = -\log\{g_\theta(x)\}$, then it can be easily shown that

$$\kappa \geqslant \frac{1}{4L^2} \|\theta_*^{(2)} - \theta_*^{(1)}\|_2^2,$$

where L is an upper bound on the largest eigenvalue of $\theta_*^{(1)}$ and $\theta_*^{(2)}$. Hence in this case assumption 3 holds. Such a general result is more difficult to establish for discrete Markov random fields. However, it can be easily shown that assumption 3 holds if

$$\begin{aligned}(\theta_*^{(1)} - \theta_*^{(2)})' \mathbb{E}_{\theta_*^{(2)}}[\nabla^{(2)}\phi(\theta_*^{(2)}, X)](\theta_*^{(1)} - \theta_*^{(2)})' > 0,\\ (\theta_*^{(2)} - \theta_*^{(1)})' \mathbb{E}_{\theta_*^{(1)}}[\nabla^{(2)}\phi(\theta_*^{(1)}, X)](\theta_*^{(2)} - \theta_*^{(1)})' > 0.\end{aligned} \tag{13}$$

And, in the particular setting where $\theta_*^{(1)}$ and $\theta_*^{(2)}$ have similar sparsity patterns (in the sense that $\theta_*^{(2)} - \theta_*^{(1)} \in \mathbb{C}_1 \cap \mathbb{C}_2$), then expression (13) follows from assumption 1, and the discussion in remark 2.

Finally, we define the search domain as the set

$$\mathcal{T} = \mathcal{T}_+ \cup \mathcal{T}_-, \tag{14}$$

where $\mathcal{T}_+$ is defined as the set of all time points $\tau \in \{\tau_* + 1, \ldots, T\}$ such that

$$\begin{gathered} c_0 b(\tau - \tau_*) \leqslant 2\surd\{\tau \log(dT)\}, \\ 64 c_0^3 b s_1 (\tau - \tau_*) \leqslant \rho_1 \tau, \end{gathered} \tag{15}$$

and $\mathcal{T}_-$ is defined as the set of all time points $\tau \in \{1, \ldots, \tau_*\}$ such that

$$\begin{gathered} c_0 b(\tau_* - \tau) \leqslant 2\surd\{(T-\tau) \log(dT)\}, \\ 64 c_0^3 b s_2 (\tau_* - \tau) \leqslant \rho_2 (T - \tau), \end{gathered} \tag{16}$$

where

$$b \overset{\text{def}}{=} \sup_{1 \leqslant j \leqslant p} \sum_{k=1}^{p} |\theta^{(2)}_{*jk} - \theta^{(1)}_{*jk}|. \tag{17}$$

Furthermore, for all $\tau \in \mathcal{T}$,

$$\begin{gathered} \tau \geqslant \max\{2^{11}, (48 \times 32)^2\} c_0^2 \left(\frac{s_1}{\rho_1}\right)^2 \log(dT), \\ T - \tau \geqslant \max\{2^{11}, (48 \times 32)^2\} c_0^2 \left(\frac{s_2}{\rho_2}\right)^2 \log(dT). \end{gathered} \tag{18}$$

Remark 5. Note that $\mathcal{T}$ is of the form $\{k_l, k_l + 1, \ldots, \tau_*, \tau_* + 1, \ldots, T - k_u\}$, since for τ close to τ_* both expression (15) and (16), and expression (18) hold provided that T is sufficiently large.

We can then establish the key result of this paper. Set

$$M = \frac{s_1}{\rho_1}\left(1 + c_0 \frac{s_1}{\rho_1}\right) + \frac{s_2}{\rho_2}\left(1 + c_0 \frac{s_2}{\rho_2}\right).$$

Theorem 1. Consider the model posited in equation (2), and assume assumptions 1–3. Let $\hat\tau$ be the estimator that is defined in equation (7), with $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ as in equation (8), and with a search domain $\mathcal{T}$ that satisfies expressions (15), (16) and (18). Then there is a universal finite constant $a > 0$ such that, with $\delta = aMc_0^2 \log(dT)$, we have

$$\mathbb{P}\left(\left|\frac{\hat\tau}{T} - \alpha_*\right| > \frac{4\delta}{\kappa T}\right) \leqslant \frac{16}{d} + \frac{4\exp[-\{\delta/(32c_0^2 s)\}(\kappa/\|\theta_*^{(2)} - \theta_*^{(1)}\|_2^2)^2]}{1 - \exp\{-\kappa^2/(2^7 c_0^2 s \|\theta_*^{(2)} - \theta_*^{(1)}\|_2^2)\}}, \tag{19}$$

where s is the number of non-zero components of $\theta_*^{(2)} - \theta_*^{(1)}$.

Theorem 1 gives a theoretical guarantee that for large p and for sufficiently large sample size T such that $T/\log(T) = O\{\max(s_1^2, s_2^2)\log(p)\}$ and $|\hat\tau/T - \alpha_*| = O\{\log(pT)/T\}$ with high probability. For fixed parameter change point problems, the maximum likelihood estimator of the change point is known to satisfy $|\hat\tau/T - \alpha_*| = O_P(1/T)$ (see for example Bai (2010)). This shows that our result is rate optimal, up to the logarithm factor $\log(T)$. Whether we can improve the bound and remove the $\log(T)$-term hinges on the existence of an exponential bound for the maximum of weighted partial sums of sub-Gaussian random variables, as we explain in remark 1 of the on-line supplement. Whether such a bound holds is currently an open problem, to the best of our knowledge. However, note that the $\log(p)$-term that appears in theorem 1 cannot be improved in general in the large p regime.

If the signal κ that was introduced in assumption 3 satisfies

$$\kappa \geqslant \kappa_0 \|\theta_*^{(2)} - \theta_*^{(1)}\|_2^2, \tag{20}$$

then the second term on the right-hand side of inequality (19) is upper bounded by

$$\left(\frac{1}{dT}\right)^{aM\kappa_0/(32s)} \frac{1}{1-\exp\{-\kappa_0^2/(2^7 c_0^2 s)\|\theta_*^{(2)}-\theta_*^{(1)}\|_2^2\}}. \tag{21}$$

This shows that theorem 1 can also be used to analyse cases where $\|\theta_*^{(2)}-\theta_*^{(1)}\|_2^2 \downarrow 0$, as $p \to \infty$. In such cases, consistency is guaranteed provided that the term in expressions (21) converges to 0. From the right-hand side of inequality (20), we then see that the convergence rate of the estimator in such cases is changed to

$$\frac{c_0^2}{\|\theta_*^{(2)}-\theta_*^{(1)}\|_2^2} \frac{\log(dT)}{T}.$$

Another nice feature of theorem 1 is that the constant M describes the behaviour of the change point estimator as a function of the key parameters of the problem. In particular, the bound in expression (19) shows that the change point estimator improves as s_1 and s_2 (the number of non-zero entries of the matrices $\theta_*^{(1)}$ and $\theta_*^{(2)}$ respectively), or the noise term c_0 (the maximum fluctuation of B_0 and B) decrease.

4. Algorithm and implementation issues

Given a sequence of observed p-dimensional vectors $\{x^{(t)}, 1 \leqslant t \leqslant T\}$, we propose the following algorithm to compute the change point $\hat{\tau}$, as well as the estimates $(\hat{\theta}_{1,\hat{\tau}}, \hat{\theta}_{2,\hat{\tau}})$.

Algorithm 1 (basic algorithm). Input a sequence of observed p-dimensional vectors $\{x^{(t)}, 1 \leqslant t \leqslant T\}$, and $\mathcal{T} \subseteq \{1, \ldots, T\}$ the search domain.

(a) For each $\tau \in \mathcal{T}$, estimate $\hat{\theta}_{1,\tau}$ and $\hat{\theta}_{2,\tau}$ using for instance the algorithm in Höfling and Tibshirani (2009).

(b) For each $\tau \in \mathcal{T}$, plug in the estimates $\hat{\theta}_{1,\tau}$ and $\hat{\theta}_{2,\tau}$ in equation (6) and obtain the profile (negative) pseudo-log-likelihood function

$$\mathrm{Pl}(\tau) \overset{\mathrm{def}}{=} l_T(\tau; \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}).$$

(c) Identify $\hat{\tau}$ that achieves the minimum of $\mathrm{Pl}(\tau)$ over the grid $\mathcal{T}$, and use $\hat{\theta}_{1,\hat{\tau}}$ and $\hat{\theta}_{2,\hat{\tau}}$ as the estimates of $\theta_*^{(1)}$ and $\theta_*^{(2)}$ respectively.

In our implementation of the basic algorithm, we choose a search domain $\mathcal{T}$ of the form $\mathcal{T} = \{k_l, k_l+1, \ldots, T-k_l\}$, with k_l sufficiently large to ensure reasonably good estimation errors at the boundaries. Existing results (Ravikumar *et al.*, 2010; Guo *et al.*, 2010) suggest that a sample size of order $O\{s^2 \log(d)\}$ is needed, where s is the number of edges, for a good recovery of Markov random fields.

To identify the change point $\hat{\tau}$ the algorithm requires a *full scan* of all the time points in the set $\mathcal{T}$, which can be expensive when $\mathcal{T}$ is large. As a result, we propose a fast implementation that operates in two stages. In the first stage, a coarser grid $\mathcal{T}_1 \subset \mathcal{T}$ of time points is used and steps (a) and (b) of the basic algorithm are used to obtain $l_T(\tau; \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}), \tau \in \mathcal{T}_1$. Subsequently, the profile likelihood function l_T is smoothed by using a Nadaraya–Watson kernel (Nadaraya, 1965). On the basis of this smoothed version of the profile likelihood, an initial estimate of the change point is obtained. In the second stage, a new fine resolution grid $\mathcal{T}_2$ is formed around the first-stage estimate of $\hat{\tau}$. Then, the basic algorithm is used for the grid points in $\mathcal{T}_2$ to obtain the final estimate. This leads to a more practical algorithm which is summarized next.

Algorithm 2 (fast implementation algorithm). Input a sequence of observed p-dimensional vectors $\{x^{(t)}, 1 \leqslant t \leqslant T\}$, and $\mathcal{T} \subseteq \{1, \ldots, T\}$ the search domain.

(a) Find a coarser grid $\mathcal{T}_1$ of time points.
(b) For each $\tau \in \mathcal{T}_1$, use steps (a) and (b) of the basic algorithm to obtain $\mathrm{Pl}_T(\tau), \tau \in \mathcal{T}_1$.
(c) Compute the profile negative pseudo-log-likelihood over the interval $[1, T]$ by Nadaraya–Watson kernel smoothing:

$$\widetilde{\mathrm{Pl}_{1s}}(\tau) \overset{\text{def}}{=} \frac{\sum_{\tau_i \in \mathcal{T}_1} K_{h_\nu}(\tau, \tau_i) l(\tau_i; \hat{\boldsymbol{\theta}}_{1,\tau_i}, \hat{\boldsymbol{\theta}}_{2,\tau_i})}{\sum_{\tau_i \in \mathcal{T}_1} l(\tau_i; \hat{\boldsymbol{\theta}}_{1,\tau_i}, \hat{\boldsymbol{\theta}}_{2,\tau_i})}, \qquad 1 \leqslant \tau \leqslant T.$$

The first-stage change point estimate is then obtained as

$$\hat{\tau} = \underset{1<\tau<T}{\arg\min}\, \widetilde{\mathrm{Pl}_{1s}}(\tau).$$

(d) Form a second-stage grid $\mathcal{T}_2$ around the first-stage estimate $\hat{\tau}$ and, for each $\tau \in \mathcal{T}_2$, estimate $\hat{\hat{\boldsymbol{\theta}}}_{1,\tau}$ and $\hat{\hat{\boldsymbol{\theta}}}_{2,\tau}$ by using steps (a) and (b) of the basic algorithm.
(e) Construct the second-stage smoothed profile pseudolikelihood

$$\widetilde{\mathrm{Pl}_{2s}}(\tau) \overset{\text{def}}{=} \frac{\sum_{\tau_i \in \mathcal{T}_2} K_{h_\nu}(\tau, \tau_i) l(\tau_i; \hat{\hat{\boldsymbol{\theta}}}_{1,\tau_i}, \hat{\hat{\boldsymbol{\theta}}}_{2,\tau_i})}{\sum_{\tau_i \in \mathcal{T}_2} l(\tau_i; \hat{\hat{\boldsymbol{\theta}}}_{1,\tau_i}, \hat{\hat{\boldsymbol{\theta}}}_{2,\tau_i})}, \qquad \min(\mathcal{T}_2) \leqslant \tau \leqslant \max(\mathcal{T}_2).$$

The final change point estimate is then given by

$$\hat{\hat{\tau}} = \underset{\min(\mathcal{T}_2) \leqslant \tau \leqslant \max(\mathcal{T}_2)}{\arg\min} \widetilde{\mathrm{Pl}_{2s}}(\tau).$$

5. Performance assessment

5.1. Comparing algorithm 1 and algorithm 2

We start by examining the relative performance of both the basic (algorithm 1) and the fast implementation algorithms (algorithm 2). We use the so-called Ising model, i.e. when equation (1) has $B_0(x_j) = x_j$, $B(x_j, x_k) = x_j x_k$ and $\mathbf{X} \equiv \{0, 1\}$. In all the simulation settings the sample size is set to $T = 700$, and the true change point is at $\tau_* = 350$, whereas the network size p varies from 40 to 100. All the simulation results that are reported below are based on 30 replications of algorithm 1 and algorithm 2.

The data are generated as follows. We first generate two $p \times p$ symmetric adjacency matrices each having density 10%, i.e. only about 10% of the entries are different from 0. Each off-diagonal element of $\theta^{(i)}_{*jk}$ $(i = 1, 2)$ is drawn uniformly from $[-1, -0.5] \cup [0.5, 1]$ if there is an edge between nodes j and k; otherwise $\theta^{(i)}_{*jk} = 0$. All the diagonal entries are set to 0. Given the two matrices $\theta^{(1)}_*$ and $\theta^{(2)}_*$, we generate the data $\{X^{(t)}\}_{t=1}^{\tau_*} \sim \mathrm{IID} g_{\theta^{(1)}_*}$ and $\{X^{(t)}\}_{t=\tau_*+1}^{T} \sim \mathrm{IID} g_{\theta^{(2)}_*}$ by Gibbs sampling.

Different 'signal strengths' are considered, by setting the degree of similarity between $\theta^{(1)}_*$ and $\theta^{(2)}_*$ to 0%, 20% and 40%. The degree of similarity is the proportion of equal off-diagonal elements between $\theta^{(1)}_*$ and $\theta^{(2)}_*$. Thus, the difference $\|\theta^{(2)}_* - \theta^{(1)}_*\|_1$ becomes smaller for higher degrees of similarity and, as can be seen from assumption 3, the estimation problem becomes more difficult in such cases.

Table 1. Change point estimation results by using the basic algorithm, for various percentages of similarity ($p=40$)

% of similarity	$\hat{\tau}$	*Root-mean-squared error*	*Coefficient of variation*
0	355	14.77	0.03
20	362	24.65	0.06
40	375	38.49	0.08

Table 2. Specificity, sensitivity and relative error in estimating $\theta_*^{(1)}$ and $\theta_*^{(2)}$ from the basic algorithm, with various percentages of similarity ($p=40$)

% of similarity	*Specificity*		*Sensitivity*		*Relative error*	
	$\theta_*^{(1)}$	$\theta_*^{(2)}$	$\theta_*^{(1)}$	$\theta_*^{(2)}$	$\theta_*^{(1)}$	$\theta_*^{(2)}$
0	0.78	0.87	0.79	0.89	0.70	0.63
20	0.74	0.88	0.80	0.88	0.72	0.67
40	0.71	0.80	0.77	0.81	0.75	0.72

The choices of the tuning parameters $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ were made on the basis of the Bayesian information criterion (BIC) where we search $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ over a grid Λ and for each penalty parameter the λ-value that minimizes the BIC score (which is defined below) over Λ is selected. If we define λ_1^{BIC} and λ_2^{BIC} as the λ-values selected for λ_1 and λ_2 by the BIC we have

$$\lambda_1^{\text{BIC}} = \underset{\lambda\in\Lambda}{\arg\min} -\frac{2}{T}\sum_{t=1}^{\tau}\phi(\hat{\theta}_{1,\tau}^{(\lambda)}, X^{(t)}) + \log(\tau)\|\hat{\theta}_{1,\tau}^{(\lambda)}\|_0$$

and

$$\lambda_2^{\text{BIC}} = \underset{\lambda\in\Lambda}{\arg\min} -\frac{2}{T}\sum_{t=\tau+1}^{T}\phi(\hat{\theta}_{2,\tau}^{(\lambda)}, X^{(t)}) + \log(T-\tau)\|\hat{\theta}_{2,\tau}^{(\lambda)}\|_0$$

where

$$\|\theta\|_0 \overset{\text{def}}{=} \sum_{k\leqslant j} \mathbf{1}_{\{|\theta_{jk}|>0\}}.$$

For the fast algorithm (algorithm 2), the first-stage grid that was employed had a step size of 10 and ranged from 60 to 640, whereas the second-stage grid was chosen in the interval $[\hat{\tau}-30, \hat{\tau}+30]$ with a step size of 3.

We present the results for algorithm 1 in Table 1 for the case $p=40$. It can be seen that algorithm 1 performs very well for stronger signals (0% and 20% similarity), whereas there is a small deterioration for the 40% similarity setting. The results on the specificity, sensitivity and the relative error of the estimated network structures are given in Table 2. Specificity is defined as the proportion of true negative results and can also be interpreted as a type 1 error. In contrast sensitivity is the proportion of true positive results and can be interpreted as the power of the method. The results for algorithm 2 for $p=40, 60, 100$ for the change point estimates are

Table 3. Change point estimation results for various values of p and various percentages of similarity for the fast implementation algorithm†

p	*% of similarity*	$\hat{\tau}$	$\hat{\hat{\tau}}$	*Root-mean-squared error*	*Coefficient of variation*
40	0	360	360	17.89	0.04
	20	363	361	30.07	0.08
	40	375	373	47.97	0.10
60	0	357	356	23.05	0.06
	20	388	386	43.20	0.08
	40	410	408	61.45	0.09
100	0	356	355	35.93	0.10
	20	408	401	62.89	0.10
	40	424	421	85.04	0.12

†$T=700$, $s_1=s_2=10p(p+1)/2\%$ and $\tau^*=350$.

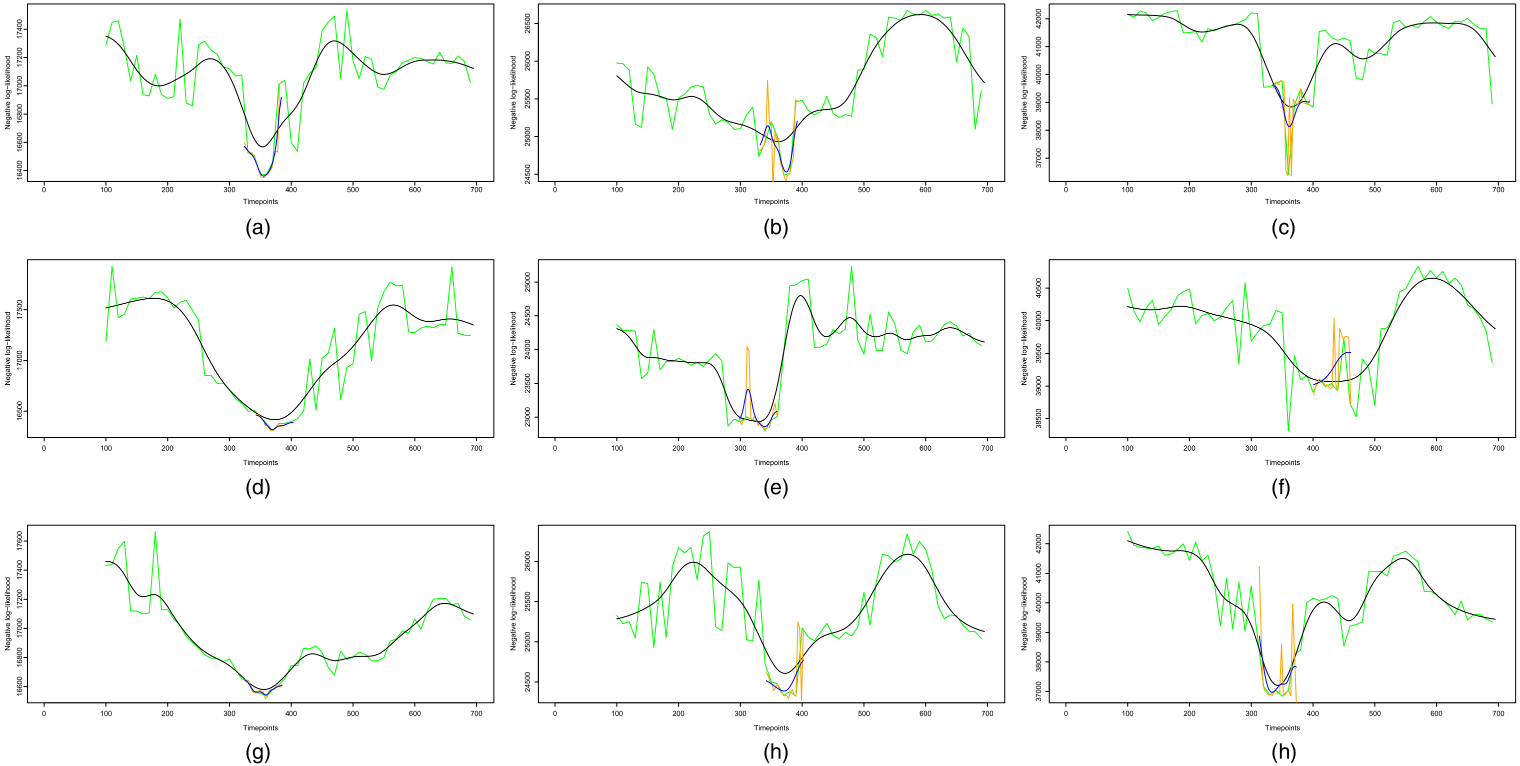
Table 4. Specificity, sensitivity and relative error of the two parameters for various values of p and various percentages of similarity for the fast implementation algorithm

p	*% of similarity*	*Specificity*		*Sensitivity*		*Relative error*	
		$\theta_*^{(1)}$	$\theta_*^{(2)}$	$\theta_*^{(1)}$	$\theta_*^{(2)}$	$\theta_*^{(1)}$	$\theta_*^{(2)}$
40	0	0.74	0.86	0.78	0.86	0.74	0.67
	20	0.74	0.81	0.76	0.82	0.73	0.71
	40	0.72	0.78	0.78	0.82	0.74	0.70
60	0	0.81	0.83	0.77	0.82	0.75	0.66
	20	0.82	0.87	0.70	0.72	0.79	0.73
	40	0.80	0.86	0.65	0.68	0.81	0.78
100	0	0.82	0.88	0.75	0.84	0.78	0.66
	20	0.81	0.87	0.66	0.70	0.81	0.78
	40	0.85	0.87	0.63	0.68	0.83	0.81

Table 5. Ratio of the computing time of one iteration of algorithm 1 and algorithm 2

p	*Ratio of computing times*
40	4.93
60	4.82
100	4.81

given in Table 3, whereas the specificity, sensitivity and relative error of the estimated network structures are given in Table 4. These results show that algorithm 2 has about 20% higher mean-squared error compared with algorithm 1. However, as pointed out in Section 4, algorithm 2 is significantly faster. In fact in this particular simulation setting, algorithm 2 is almost five times faster in a standard computing environment with four central processor unit cores. See also the



(a) (b) (c)

(d) (e) (f)

(g) (h) (h)

Fig. 1. Smoothed profile pseudo-log-likelihood functions from one run of algorithm 2 (——, non-smoothed profile pseudo-log-likelihood from stage 1 of algorithm 2; ——, smoothed version; ——, non-smoothed profile pseudo-log-likelihood functions from stage 2 of algorithm 2; ——, smoothed version): (a) 0% similarity, $p = 40$; (b) 0% similarity, $p = 60$; (c) 0% similarity, $p = 100$; (d) 20% similarity, $p = 40$; (e) 20% similarity, $p = 60$; (f) 20% similarity, $p = 100$; (g) 40% similarity, $p = 40$; (h) 40% similarity, $p = 60$; (i) 40% similarity, $p = 100$

Table 6. Positive and negative edges before and after the true change point for the two-community model

Edges	*Before change point*			*After change point*		
	Community 1	*Community 2*	*Between*	*Community 1*	*Community 2*	*Between*
Positive	50	63	0	52	21	0
Negative	0	0	10	0	0	50
Total	50	63	10	52	21	50

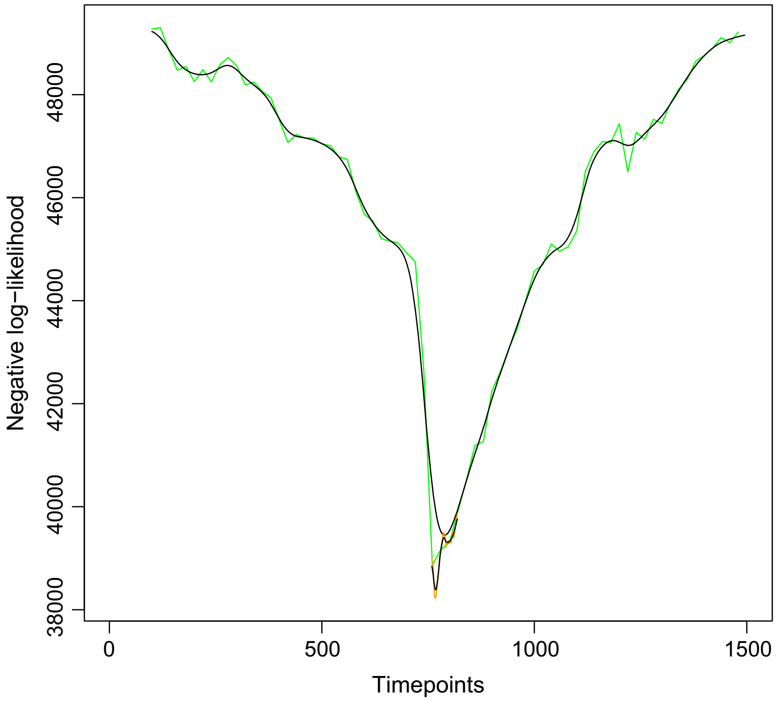


Fig. 2. Change point estimate for the two-community model with $p=50$, $T=1500$ and $\tau^*=750$

results in Table 5 which reports the ratio of the run time of a single iteration of algorithm 1 and algorithm 2.

Further, selected plots of the profile smoothed pseudo-log-likelihood functions $\widetilde{\mathrm{Pl}}_{1s}(\tau)$ and $\widetilde{\mathrm{Pl}}_{2s}(\tau)$ from the first and second stage of algorithm 2 are given in Fig. 1.

5.2. A community-based network structure

Next, we examine a setting similar to the setting that emerges from the US Senate analysis that is presented in the next section. Specifically, there are two highly 'connected' communities of size $p=50$ that are more sparsely connected before the change point but exhibit fairly strong negative association between their members after the change point. Further, the within-community connections are increased for one of them and decreased for the other after the change point. We keep the density of the two matrices encoding the network structure before and after the true change point at 10%. In the pre-change-point regime, 40% of the non-zero entries are attributed to within-group connections in community 1 (Table 6), and 50% to community 2 (Table 6), whereas the remaining 10% non-zeros represent between-group connections and are negative. Note that the within-group connections are all positive. In the post-change-point regime, the community 1 within-group connections slightly increase to 42% of the non-zero entries, whereas those of community 2 decrease to 17% of the non-zero entries. The between-group connections increase to 41% of the non-zero entries in the post-change-point regime. As before, each off-diagonal element $\theta^{(i)}_{jk}$, $i = 1, 2$, is drawn uniformly from $[-1, -0.5] \cup [0.5, 1]$ if nodes j and k are linked by an edge; otherwise $\theta^{(i)}_{*,jk} = 0$, $i = 1, 2$, and the diagonals for both the matrices are assigned as 0s. Given the two matrices $\theta^{(1)}_*$ and $\theta^{(2)}_*$, we generate data by using the 'BMN' package (Hoefling, 2010) as described earlier. The total sample size that was employed is $T = 1500$ and the true change point is at $\tau^* = 750$. We choose the first-stage grid comprising 50 points with a step size of 27 and the second-stage grid is chosen in a neighbourhood of the first-stage estimate with a step size of 3 with 20 points. We replicate the study five times and find that the estimated change point averaged over the five replications is $\hat{\tau} = 768$. Fig. 2 is the relevant figure for this two-community model. The analysis indicates that our proposed methodology can estimate the true change point sufficiently well in the presence of varying degrees of connections between two communities over two different time periods, which is a reassuring feature for the US Senate application that is presented next.

6. Application to roll call data of the US Senate

The data that are examined correspond to voting records of the US Senate covering the period from 1979 (96th Congress) to 2012 (112th Congress) and were obtained from the Web site `www.voteview.com`. Specifically, for each of the 12 129 votes that were cast during this period, the following information is recorded: the date that the vote occurred and the response to the bill or resolution under consideration—yes, no or abstain—of the 100 Senate members. Because of the length of the time period under consideration, there was significant turnover of Senate members due to retirements, loss of re-election bids, appointments to the cabinet or other administrative positions, or physical demise. To hold the number of nodes fixed to 100 (the size of membership of the US Senate at any point in time), we considered Senate seats (e.g. Michigan 1 and Michigan 2) and carefully mapped the senators to their corresponding seats, thus creating a continuous record of the voting pattern of each Senate seat.

A significant number of the 12 129 votes deal with fairly mundane procedural matters, thus resulting in nearly unanimous outcomes. Hence, only votes exhibiting conformity less than 75% (yes or no) in either direction were retained, thus resulting in an effective sample size of $T = 7949$ votes. Further, missing values due to abstentions were imputed by the value (yes or no) of that member's party majority position on that particular vote. Other imputation methods of missing values were employed:

(a) replacing all missing values by the value (yes or no) representing the winning majority on that bill and
(b) replacing the missing value of a Senator by the value that the majority of the opposite party voted on that particular bill.

The results based on these two alternative imputation methods are given in the on-line supplement.

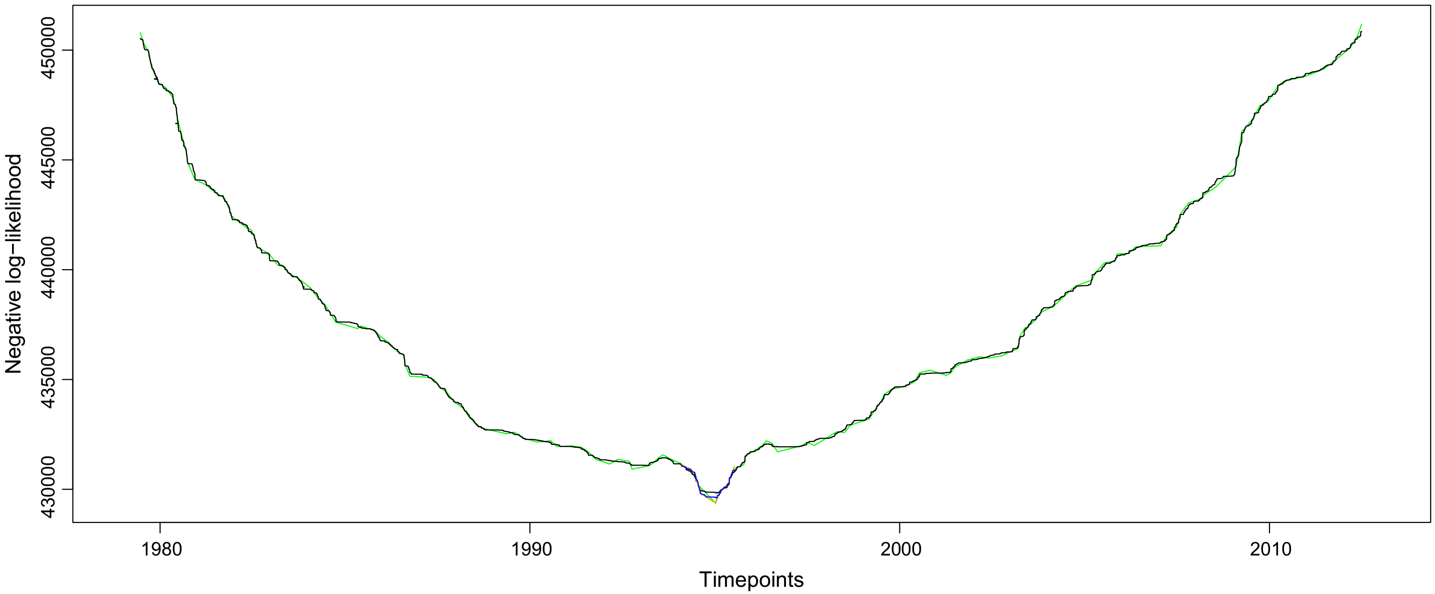
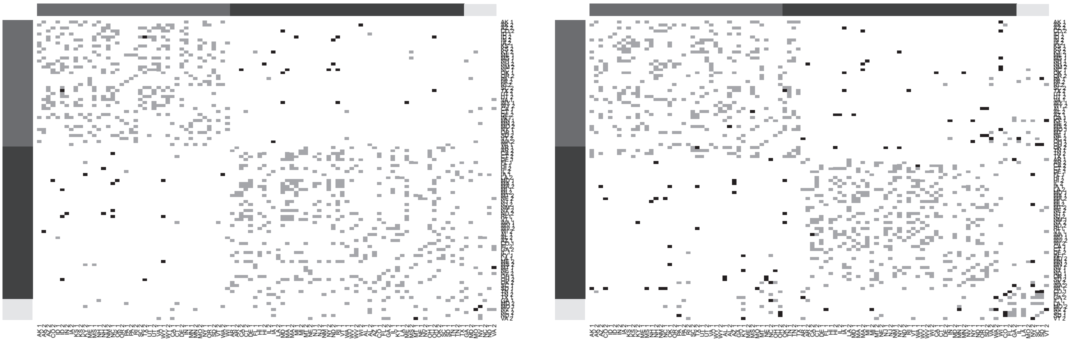


Fig. 3. Estimate of the change point for the combined US Senate data from 1979 to 2012



(a) (b)

Fig. 4. Heat map of the stable network structures (a) before and (b) after the estimated change point: ▨, Republican; ■, Democrat; □, mixed

Finally, the yes–no votes were encoded as 1–0 respectively. Under the posited model, votes are considered as IID from the same underlying distribution pre and post any change point. In reality, voting patterns are more complex and in all likelihood exhibit temporal dependence within the 2-year period that a Congress serves and probably even beyond that due to the slow turnover of Senate members. Nevertheless, the model proposed serves as a *working model* that captures essential features of the evolving voting dependence structure between Senate seats over time.

The likelihood function together with an estimate of a change point are depicted in Fig. 3 based on the fast implementation algorithm that was presented in Section 4. We choose our first-stage grid with a step size of 50 that yields 157 points excluding time points that are close to both boundaries. In the second stage, we choose a finer resolution grid with a step size of 20 in a neighbourhood of the first-stage change point estimate. The vote corresponding to the change point occurred on January 17th, 1995, at the beginning of the tenure of the 104th Congress. This change point comes in the footsteps of the November 1994 election that witnessed the Republican Party's capturing the US House of Representatives for the first time after 1956. As discussed in the political science literature, the 1994 election marked the end of the 'conservative coalition', which was a bipartisan coalition of conservative-oriented Republicans and Democrats on President Roosevelt's New Deal policies, which had often managed to control Congressional outcomes since the New Deal era. Note that other analyses based on fairly *ad hoc* methods (e.g. Moody and Mucha (2013)) also point to a significant change occurring after the November 1994 election.

Next, we examine more closely the pre- and post-change-point network structures, shown in the form of heat maps of the adjacency matrices in Fig. 4. To obtain stable estimates of the network structures, stability selection (Meinshausen and Bühlmann, 2010) was employed with edges retained if they were present in more than 90% of the 50 networks estimated from bootstrapped data. To aid interpretation, the 100 Senate seats were assigned to three categories: Democrat (dark grey bars), mixed (very light grey bars) and Republican (light grey bars). Specifically, a seat was assigned to the Democrat or Republican categories if it were held for more than 70% of the time by the corresponding party within the pre- or post-change-point periods; otherwise, it was assigned to the mixed category. This means that, if a seat was held for more than five out of the eight Congresses in the pre-change-point period and similarly six out of nine Congresses in the post period by the Democrats, then it is assigned to that category and similarly for Republican assignments; otherwise, it is categorized as mixed.

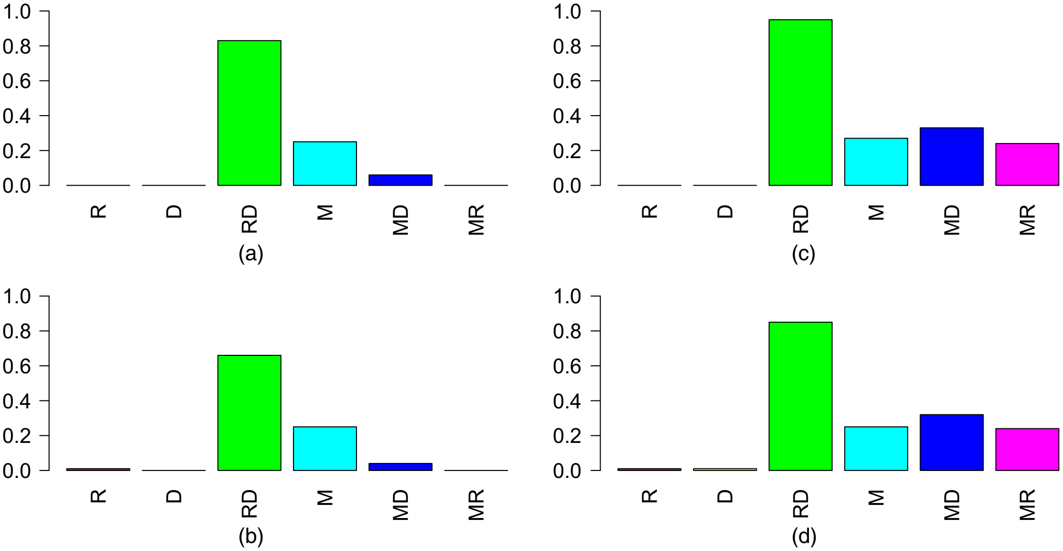


Fig. 5. Proportion of negative edges for network structures (a), (b) before and (c), (d) after the estimated change point for (a), (c) the BIC and (b), (d) stability selection with threshold 0.8

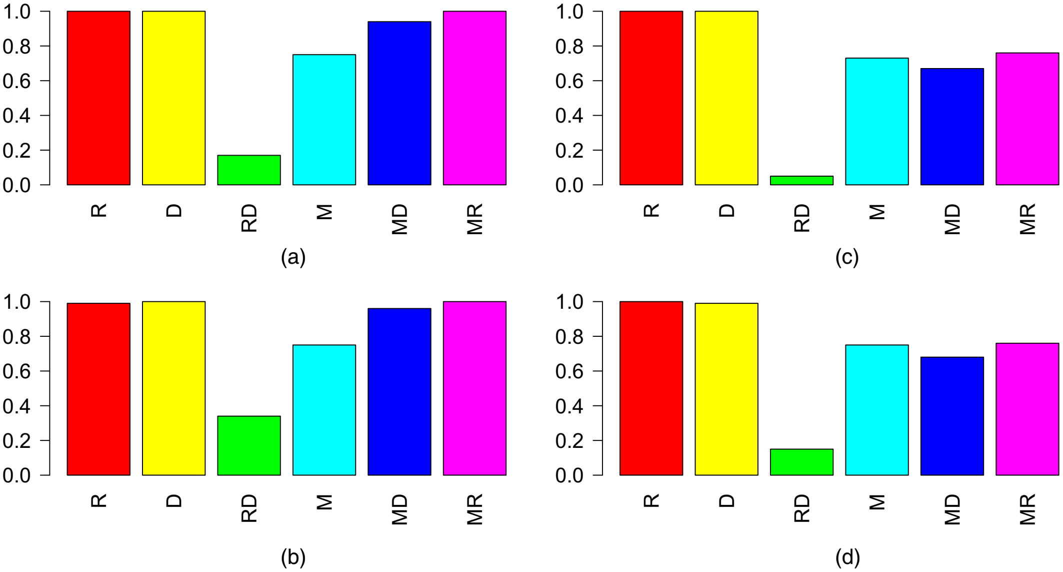


Fig. 6. Proportion of positive edges for network structures (a), (b) before and (c), (d) after the estimated change point for (a), (c) the BIC and (b), (d) stability selection with threshold 0.8

In the heat maps depicted, the orderings of the Senate seats in the pre- and post-change-point regimes are kept as similar as possible, since some of the seats changed their category membership completely across periods. Further, the light grey dots represent positive edge weights, mostly corresponding to within-categories interactions, whereas black dots represent negative edge weights, mostly between-category interactions. An emergence of a significant number of black dots can be clearly seen in the post-change-point regimes, which is indicative of sharper disagreements between political parties and thus increased polarization. Further, it

Table 7. Different network statistic values for stability selection with threshold 0.9 and 0.8

Stability selection threshold	*Network statistic*	*Before change point*			*After change point*		
		Republican	*Democrat*	*Mixed*	*Republican*	*Democrat*	*Mixed*
0.9	Centrality score	0.004	0.368	0.054	0.001	0.483	0.034
	Clustering coefficient	0.346	0.311	0.339	0.334	0.251	0.391
0.8	Centrality score	0.004	0.378	0.055	0.001	0.481	0.078
	Clustering coefficient	0.366	0.371	0.360	0.378	0.307	0.364

can be seen that in the post-change-point regime the mixed group becomes more prominent, indicating that it contributes to the emergence of a change point.

To explore the reasons behind the presence of a change point further, we provide some network statistics in Fig. 5 and Fig. 6. Specifically, Figs 5 and 6 present the proportion of positive and negative edges, before and after the estimated change point by using two different methods for selecting the penalty tuning parameters: an analogue of the BIC and threshold 0.8 for the stability selection method. The patterns shown across Figs 5 and 6 for the two different methods are very similar—high proportions of positive edges within groups and very low or almost negligible proportions of negative edges within the Republican or Democrat groups in both pre- and post-change-point periods. Further, a large proportion of negative edges can account for Republican and Democrat group interactions, which tend to increase in the post-regime. One noticeable fact is that the proportion of positive edges within the Republican and Democrat groups remain almost the same from the pre- to the post-change-point regime under both the BIC and stability selection whereas the proportion of positive edges between the two groups decreases and the proportion of negative edges between them tend to increase from the pre- to post-change-point regime for both methods. It can also be observed that the mixed and the Democrat groups exhibit a large proportion of positive edges between them in the pre-regime, as gleaned from their overlap in the corresponding heat map.

We also present some other network statistics, such as average degree, centrality scores and average clustering coefficients for the three groups Republican, Democrat and mixed in Table 7. We observe that in terms of centrality scores the Democrat group is more influential than the Republican group, in both the pre- and the post-change-point network structures, whereas in terms of clustering coefficient values the Republican group is ahead of the Democrat group and the gap increases from pre- to post-change-point regime, also reflected in the finding that the number of edges within the Republican group mostly remains the same from pre- to post-regimes, whereas for the Democrats it decreases. These results suggest that the Republicans form a tight cluster, whereas the Democrats not to the same extent.

Acknowledgements

The authors thank the Joint Editor, Associate Editor and the two referees for their insightful comments that have substantially improved this paper. The authors also thank Howard Rosenthal, New York University, and Keith Poole, University of Georgia, for maintaining the freely downloadable roll call data of the US Senate at `www.voteview.com`. This work was supported in part by National Science Foundation awards DMS-1228164 (to YA and GM) and DMS-1545277 (to GM), and National Institutes of Health award 7R21GM10171903 (GM).

References

Atchadé, F. Y. (2014) Estimation of Network Structures from partially observed markov random field. *Electron. J. Statist.*, **8**, 2242–2263.

Bai, J. (2010) Estimation of a change-point in multiple regression models. *Rev. Econ. Statist.*, **4**, 551–563.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

Bhattcharya, K. P. (1987) Maximum likelihood estimation of a change-point in the distribution of the independent random variables: General Multiparameter Case. *J. Multiv. Anal.*, **23**, 183–208.

Carlstein, E. (1988) Nonparametric change-point estimation. *Ann. Statist.*, **16**, 188–197.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010) Joint structure estimation for categorical markov networks. *Technical Report*. University of Michigan, Ann Arbor.

Hinkley, D. V. (1970) Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17.

Hoefling, H. (2010) BMN: the pseudo-likelihood method for pairwise binary markov networks. *R Package Version 1.02*. (Available from `http://CRAN.R-project.org/package=BMN`.)

Höfling, H. and Tibshirani, R. (2009) Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.*, **10**, 883–906.

Kolar, M., Song, L., Ahmed, A. and Xing, P. E. (2010) Estimating time varying networks. *Ann. Appl. Statist.*, **4**, 94–123.

Kolar, M. and Xing, P. E. (2012) Estimating networks with jumps. *Electron. J. Statist.*, **6**, 2069–2106.

Kosorok, R. M. (2008) *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Lan, Y., Banerjee, M. and Michailidis, G. (2009) Change-point estimation under adaptive sampling. *Ann. Statist.*, **37**, 1752–1791.

Leonardi, F. and Bühlmann, P. (2016) Computationally efficient change point detection for high-dimensional regression. *Preprint arXiv:1601.03704*.

Loader C. (1996) Change-point estimation using nonparametric regression. *Ann. Statist.*, **24**, 1667–1678.

Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Moody, J. and Mucha, P. (2013) Portrait of political party polarization *Netwrk Sci.*, **1**, 119–121.

Muller, H. (1992) Change-points in nonparametric regression analysis. *Ann. Statist.*, **20**, 737–761.

Nadaraya, E. A. (1965) On non-parametric estimation of density functions and regression curves. *Theor. Probab. Appl.*, **10**, 186–190.

Neghaban, S., Ravikumar, P., Wainwright, M. and Yu, B. (2010) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, **27**, 538–557.

Raimondo, M. (1998) Minimax estimation of sharp change-points. *Ann. Statist.*, **26**, 1379–1397.

Ravikumar, P., Wainwright, J. M. and Lafferty, D. J. (2010) High-dimensional Ising model selection using l_1-regularized logistic regression. *Ann. Statist.*, **38**, 1287–1319.

Rudelson, M. and Zhou, S. (2013) Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theor.*, **59**, 3434–3447.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Soh, Y. S. and Chandrasekaran, V. (2014) High-dimensional change-point estimation: combining filtering with convex optimization. *Preprint arXiv:1412.3731*.

Xue, L., Zou, H. and Cai, T. (2012) Non-concave penalized composite likelihood estimation of sparse ising models. *Ann. Statist.*, **40**, 1403–1429.

Zhou, S., Lafferty, J. and Wasserman, L. (2010) Time-varying undirected graphs. *Mach. Learn.*, **80**, 295–319.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to the paper "Change-point estimation in high-dimensional Markov random field models"'.