# Change-Point Estimation in High-Dimensional Markov Random Field Models

Sandipan Roy†, Yves Atchadé† and George Michailidis†

*University of Michigan, Ann Arbor, USA.*

**Summary**. This 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 pseudo-likelihood 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 proposed estimator 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 Pseudo-likelihood.

## 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, voting records of legislative bodies, etc.

In this work, we consider modeling the underlying network through a Markov random field (MRF) that exhibits a change in its structure at some point in time.

†*Address for correspondence:* Department of Statistics, 439 West Hall, University of Michigan, Ann Arbor, MI 48109-1107, USA

E-mail: sandipan@umich.edu, yvesa@umich.edu, gmichail@umich.edu

Specifically, suppose we have $T$ observations $\left\{X^{(t)}, 1 \leq t \leq T\right\}$ over $p$-variables with $X^{(t)} = \left(X_1^{(t)}, \ldots, X_p^{(t)}\right)$ and $X_j^{(t)} \in \mathsf{X}$, for some finite set $\mathsf{X}$. Further, we assume that there exists a time point $\tau_\star = \lceil \alpha_\star T \rceil \in \{1, \ldots, T-1\}$, with $\alpha_\star \in (0,1)$, such that $\left\{X^{(t)}, 1 \leq t \leq \tau_\star\right\}$ is an independent and identically distributed sequence from a distribution $g_{\theta_\star^{(1)}}(\cdot)$ parametrized by a real symmetric matrix $\theta_\star^{(1)}$, while the remaining observations $\left\{X^{(t)}, \tau_\star + 1 \leq t \leq T\right\}$ forms also an independent and identically distributed sequence from a distribution $g_{\theta_\star^{(2)}}(\cdot)$ parametrized by another real symmetric matrix $\theta_\star^{(2)}$. We assume that the two distributions $g_{\theta_\star^{(1)}}(\cdot)$, $g_{\theta_\star^{(2)}}(\cdot)$ belong to a parametric family of Markov random field distributions given by

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

for a non-zero function $B_0 : \mathsf{X} \to \mathbb{R}$, and a non-zero symmetric function $B : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ which encodes the interactions between the nodes. The term $Z(\theta)$ is the corresponding normalizing constant. Thus, the observations over time come from a MRF that exhibits a change in its structure at time $\tau_\star$ and the matrices $\theta_\star^{(1)}$ and $\theta_\star^{(2)}$ 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 $\tau_\star$, as well as the network structures $\theta_\star^{(1)}$ and $\theta_\star^{(2)}$. Although the problem of identifying a change point has a long history in statistics (see Bai (2010), Carlstein (1988), Hinkley (1970), Loader (1996), Lan, Banerjee and Michailidis (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 discussed in Section 6. Note that in a low-dimensional setting, the results obtained for the change-point depend on the regime 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)), while 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), Muller (1992)).

Note that the methodology 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 MRF model has been a very active research area (see e.g. Banerjee et al. (2008); Höfling and Tibshirani (2009); Ravikumar et al. (2010); Xue et al. (2012); Guo et al. (2010) and references therein). Sparsity (an often realistic assumption) 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)).

On the other hand, there is significant 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 considers 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, metabolites), nodewise change-point analysis would typically be preferred, whereas is many social network applications (such as the political network example 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 modeling framework 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, a strategy that has worked well in regression modeling where the number of parameters is the same as the number of observations. On the other hand, this paper assumes a high-dimensional framework where the number of nodes (and hence the number of parameters of interest, namely the edges) grow 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 pseudo-likelihood framework. As customary in the analysis

of change-point problems (Bai (2010); Lan, Banerjee and Michailidis (2009)), we employ a profile pseudo-likelihood function to obtain the estimate $\hat{\tau}$ of the true change-point $\tau_\star$. 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, $|(\hat{\tau}/T) - \alpha_\star| = O(\log(pT)/T)$, as $p, T \to \infty$. Note that in classical change-point problems with a fixed-magnitude change, it is well-known that the maximum likelihood estimator of the change-point satisfies $|(\hat{\tau}/T) - \alpha_\star| = O_p(1/T)$ (see e.g. Bhattacharya (1987), Bai (2010)). This suggests that our result is rate-optimal, up to the logarithm factor $\log(T)$. The derivation of the result requires a careful handling of model misspecification in Markov random fields as explained in Section 3, a novel aspect not present when estimating a single Markov random field from independent and identically distributed 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 developed methodology to analyze 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 dependencies between them. The analysis strongly points to the presence of a change-point around January, 1995, the beginning of the tenure of the 104th Congress. This change-point comes at the footsteps of the November 1994 election that witnessed the Republican Party 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. Modeling assumptions and the estimation framework are presented in Section 2, while Section 3 establishes the key technical results. Section 4 discusses computational issues and Section 5 evaluates the performance of the estimation procedure using synthetic data. Section 6 illustrates the procedure on the US Senate voting record. Finally, proofs are deferred to the Supplement.

## 2. Methodology

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

$$\mathbb{P}\left(X^{(t)} = x\right) = g_{\theta_\star^{(1)}}(x), \quad \text{for } t = 1, \ldots, \tau_\star,$$

and

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

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

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

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 $\mathsf{X}^p$ with probability mass function $g_\theta$ and $\mathbb{E}_\theta$ its expectation operator.

We are interested in estimating both the change point $\tau_\star$, as well as the parameters $\theta_\star^{(1)}, \theta_\star^{(2)}$. 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 $\langle \theta, \vartheta \rangle_{\mathsf{F}} \stackrel{\text{def}}{=} \sum_{k\le j} \theta_{jk}\vartheta_{jk}$, and the associated norm $\|\theta\|_{\mathsf{F}} \stackrel{\text{def}}{=} \sqrt{\langle \theta, \theta \rangle}$. 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 $\|\theta\|_1 \stackrel{\text{def}}{=} \sum_{k\le j} |\theta_{jk}|$, and $\|\theta\|_\infty \stackrel{\text{def}}{=} \sup_{k\le j} |\theta_{jk}|$. If $u \in \mathbb{R}^d$, for some $d \ge 1$, and $A$ is an ordered subset of $\{1, \ldots, d\}$, we define $u_A \stackrel{\text{def}}{=} (u_j,\ j \in A)$, and $u_{-j}$ is a shortcut for $u_{\{1,\ldots,d\}\setminus\{j\}}$.

To avoid some of the computational difficulties in dealing with the normalizing constant of $g_\theta$, we take a pseudo-likelihood approach. For $\theta \in \mathcal{M}_p$ and $j \in \{1, 2, \ldots, p\}$, define $f_\theta^{(j)}(u|x) \stackrel{\text{def}}{=} \mathbb{P}_\theta(X_j = u | X_{-j} = x_{-j})$, for $u \in \mathsf{X}$, and $x \in \mathsf{X}^p$. From the expression of the joint distribution $g_\theta$ in (1), we have

$$f_\theta^{(j)}(u|x) = \frac{1}{Z_\theta^{(j)}(x)} \exp\left(\theta_{jj} B_0(u) + \sum_{k\ne j} \theta_{jk} B(u, x_k)\right),\ u \in \mathsf{X},\ x \in \mathsf{X}^p, \tag{3}$$

where

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

The normalizing constant $Z_\theta^{(j)}(x)$ defined in (4) is actually a summation over $\mathsf{X}$, but for notational convenience we write it as an integral against the counting measure on $\mathsf{X}$. Next, we introduce

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

The negative log-pseudo-likelihood of the model (divided by $T$) is given by

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

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

$$\widehat{\theta}_{1,\tau}^{(\lambda)} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{\theta \in \mathcal{M}_p} \frac{1}{T} \sum_{t=1}^{\tau} \phi(\theta, X^{(t)}) + \lambda \|\theta\|_1,$$

and

$$\widehat{\theta}_{2,\tau}^{(\lambda)} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{\theta \in \mathcal{M}_p} \frac{1}{T} \sum_{t=\tau+1}^{T} \phi(\theta, X^{(t)}) + \lambda \|\theta\|_1.$$

We propose to estimate the change point $\tau_\star$ using a profile pseudo-likelihood approach. More precisely our estimator $\hat{\tau}$ is defined as

$$\widehat{\tau} = \operatorname*{Argmin}_{\tau \in \mathcal{T}} \ell_T(\tau; \widehat{\theta}_{1,\tau}, \widehat{\theta}_{2,\tau}), \tag{7}$$

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 $\tau \in \mathcal{T}$, $\widehat{\theta}_{1,\tau} = \hat{\theta}_{1,\tau}^{(\lambda_{1,\tau})}$ and $\widehat{\theta}_{2,\tau} = \hat{\theta}_{1,\tau}^{(\lambda_{1,\tau})}$, for some positive penalty parameters $\lambda_{1,\tau}$, $\lambda_{2,\tau}$. 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 behavior of the estimators, and we provide some guidelines below.

## 3. Theoretical Results

The recovery of $\tau_\star$ rests upon the ability of the estimators $\hat{\theta}_{j,\tau}$ to correctly estimate $\theta_\star^{(j)}$, $j \in \{1,2\}$. Estimators for the static version of the problem where one has i.i.d. 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 $\|\hat{\theta}_{j,\tau} - \theta_\star^{(j)}\|_2$, we borrow from the approach in Atchadé (2014). For penalty terms $\lambda_{j,\tau}$ as in (8) and under some regularity assumptions, we derive a bound on the estimator errors $\|\hat{\theta}_{j,\tau} - \theta_\star^{(j)}\|_2$, 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 $\tau \mapsto \ell_T(\tau; \theta_\star^{(1)}, \theta_\star^{(2)})$ and this allows us to establish a bound on the distance between $\hat{\tau}$ and the true change point $\tau_\star$.

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

$$\lambda_{1,\tau} = \frac{32c_0\sqrt{\tau \log(dT)}}{T} \text{ and } \lambda_{2,\tau} = \frac{32c_0\sqrt{(T-\tau)\log(dT)}}{T}, \tag{8}$$

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

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

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

For $j = 1, 2$, define $\mathcal{A}_j \stackrel{\text{def}}{=} \left\{1 \le k \le i \le p : \theta_{\star ik}^{(j)} \neq 0\right\}$, and define $s_j \stackrel{\text{def}}{=} |\mathcal{A}_j|$ the cardinality (and hence the sparsity) of the true model parameters. We also define

$$\mathbb{C}_j \stackrel{\text{def}}{=} \left\{ \theta \in \mathcal{M}_p : \sum_{(k,i) \in \mathcal{A}_j^c} |\theta_{ik}^{(j)}| \le 3 \sum_{(k,i) \in \mathcal{A}_j} |\theta_{ik}^{(j)}| \right\}, j \in \{1,2\}, \tag{10}$$

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

**H**1. [Restricted Strong Convexity] *For $j \in \{1,2\}$, and $X \sim g_{\theta_\star^{(j)}}$, there exists*

$\rho_j > 0$ *such that for all* $\Delta \in \mathbb{C}_j$,

$$\sum_{i=1}^{p} \mathbb{E}_{\theta_\star^{(j)}} \left[ \mathsf{Var}_{\theta_\star^{(j)}} \left( \sum_{k=1}^{p} \Delta_{ik} B_{ik}(X_i, X_k) | X_{-i} \right) \right] \geq 2\rho_j \, \|\Delta\|_2^2, \tag{11}$$

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

REMARK 1. *Assumption H1 is a (averaged) restricted strong convexity (RSC) assumption on the negative log-pseudo-likelihood function* $\phi(\theta, x)$. *This can be seen by noting that (11) can also be written as*

$$\Delta' \mathbb{E} \left[ \nabla^{(2)} \phi(\theta_\star^{(j)}, X^{(j)}) \right] \Delta \geq 2\rho_j \|\Delta\|_2^2, \quad X^{(j)} \sim g_{\theta_\star^{(j)}}, \;\; \Delta \in \mathbb{C}_j, \;\; j \in \{1, 2\}.$$

*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 e.g. Neghaban et al. (2010) and references therein). Note that the RSC assumption is expressed here in expectation, unlike Neghaban et al. (2010) which uses an almost sure version. Imposing this assumption in expectation (that is, 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 Supplement).*

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

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

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

*where* $\rho_1$, *and* $\rho_2$ *are as in H1.*

REMARK 2. *Note that the constants* $2^{11}$ *and* $48^2 \times 32^2$ *required in H2 will typically yield a very conservative bound on the sample size* $T$. *We believe these large constants are mostly artifacts of our techniques, and can be improved. The key point of H2 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 $\theta_\star^{(1)}$ to $\theta_\star^{(2)}$ be identifiable.

**H**3. [Identifiability Condition] *Assume that* $\theta_\star^{(1)} \neq \theta_\star^{(2)}$, *and*

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

REMARK 3. *Assumption H3 is needed for the identifiability of the change-point* $\tau_\star$. *Since the distributions* $g_\theta$ *are discrete data analogs of Gaussian graphical distributions, it is informative to look at H3 for Gaussian graphical distributions. Indeed, if* $g_\theta$ *is the density of the p-dimensional normal distribution* $\boldsymbol{N}(0, \theta^{-1})$ *with precision matrix* $\theta$, *and if we take* $\phi(\theta, x) = -\log g_\theta(x)$, *then it can be easily shown that*

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

*where* $L$ *is an upper bound on the largest eigenvalue of* $\theta_\star^{(1)}$ *and* $\theta_\star^{(2)}$. *Hence in this case H3 holds. Such a general result is more difficult to establish for discrete Markov random fields. However, it can be easily shown that H3 holds if*

$$\left(\theta_\star^{(1)} - \theta_\star^{(2)}\right)' \mathbb{E}_{\theta_\star^{(2)}}\left[\nabla^{(2)}\phi(\theta_\star^{(2)}, X)\right]\left(\theta_\star^{(1)} - \theta_\star^{(2)}\right)' > 0,$$
$$\text{and} \quad \left(\theta_\star^{(2)} - \theta_\star^{(1)}\right)' \mathbb{E}_{\theta_\star^{(1)}}\left[\nabla^{(2)}\phi(\theta_\star^{(1)}, X)\right]\left(\theta_\star^{(2)} - \theta_\star^{(1)}\right)' > 0. \tag{13}$$

*And in the particular setting where* $\theta_\star^{(1)}$ *and* $\theta_\star^{(2)}$ *have similar sparsity patterns (in the sense that* $\theta_\star^{(2)} - \theta_\star^{(1)} \in \mathbb{C}_1 \cap \mathbb{C}_2$*), then (13) follows from H1, and the discussion in Remark 1.*

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 $\tau \in \{\tau_\star + 1, \ldots, T\}$ such that

$$c_0 b(\tau - \tau_\star) \leq 2\sqrt{\tau \log(dT)}, \quad \text{and} \quad 64c_0^3 b s_1(\tau - \tau_\star) \leq \rho_1 \tau, \tag{15}$$

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

$$c_0 b(\tau_\star - \tau) \leq 2\sqrt{(T - \tau)\log(dT)}, \quad \text{and} \quad 64c_0^3 b s_2(\tau_\star - \tau) \leq \rho_2(T - \tau), \tag{16}$$

where

$$b \stackrel{\text{def}}{=} \sup_{1 \leq j \leq p} \sum_{k=1}^{p} \left|\theta_{\star jk}^{(2)} - \theta_{\star jk}^{(1)}\right|. \tag{17}$$

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

$$\tau \geq \max\left(2^{11}, (48 \times 32)^2\right) c_0^2 \left(\frac{s_1}{\rho_1}\right)^2 \log(dT),$$

$$\text{and} \quad T - \tau \geq \max\left(2^{11}, (48 \times 32)^2\right) c_0^2 \left(\frac{s_2}{\rho_2}\right)^2 \log(dT). \quad (18)$$

REMARK 4. *Notice that $\mathcal{T}$ is of the form $\{k_l, k_l + 1, \ldots, \tau_\star, \tau_\star + 1, \ldots, T - k_u\}$, since for $\tau$ close to $\tau_\star$ both (15), (16), and (18) hold provided that $T$ is large enough.*

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

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

THEOREM 1. *Consider the model posited in (2), and assume H1-H3. Let $\hat{\tau}$ be the estimator defined in (7), with $\lambda_{1,\tau}, \lambda_{2,\tau}$ as in (8), and with a search domain $\mathcal{T}$ that satisfies (15), (16), and (18). Then there exists a universal finite constant $a > 0$, such that with $\delta = aMc_0^2 \log(dT)$, we have*

$$\mathbb{P}\left(\left|\frac{\widehat{\tau}}{T} - \alpha_\star\right| > \frac{4\delta}{\kappa T}\right) \leq \frac{16}{d} + \frac{4\exp\left(-\frac{\delta}{32c_0^2 s}\left(\frac{\kappa}{\|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2}\right)^2\right)}{1 - \exp\left(-\frac{\kappa^2}{2^7 c_0^2 s \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2}\right)}, \quad (19)$$

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

Theorem 1 gives a theoretical guarantee that for large $p$ and for large enough sample size $T$ such that $(T/\log(T)) = O(\max(s_1^2, s_2^2)\log(p))$, $|\hat{\tau}/T - \alpha_\star| = O(\log(pT)/T)$ with high-probability. For fixed-parameter change-point problems, the maximum likelihood estimator of the change-point is known to satisfy $|\hat{\tau}/T - \alpha_\star| = O_P(1/T)$ (see e. g. Bai (2010)). This shows that our result is rate-optimal, up to the logarithm factor $\log(T)$. Whether one 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 Supplement. Whether such bound holds is currently an open problem, to the best of our knowledge. However, note that the $\log(p)$ term that appears in the theorem cannot be improve in general in the large $p$ regime.

If the signal $\kappa$ introduced in H3 satisfies

$$\kappa \geq \kappa_0 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2, \quad (20)$$

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

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

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

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

Another nice feature of Theorem 1 is the fact that the constant $M$ describes the behavior of the change-point estimator as a function of the key parameters of the problem. In particular, the bound in (19) shows that the change-point estimator improves as $s_1, s_2$ (the number of non-zero entries of the matrices $\theta_\star^{(1)}, \theta_\star^{(2)}$ resp.), 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 $\{x^{(t)}, 1 \le t \le T\}$, we propose the following algorithm to compute the change point $\hat{\tau}$, as well as the estimate the estimates $\left(\hat{\theta}_{1,\hat{\tau}}, \hat{\theta}_{2,\hat{\tau}}\right)$.

ALGORITHM 1 (**Basic Algorithm**). *Input: a sequence of observed $p$-dimensional vectors $\{x^{(t)}, 1 \le t \le T\}$, and $\mathcal{T} \subseteq \{1, \ldots, T\}$ the search domain.*

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

*(b) For each $\tau \in \mathcal{T}$, plug-in the estimates $\hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}$ in (6) and obtain the profile (negative) pseudo-log-likelihood function $\mathcal{P}\ell(\tau) \stackrel{\text{def}}{=} \ell_T(\tau; \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau})$.*

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

Note that 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 $\ell_T(\tau; \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}), \tau \in \mathcal{T}_1$. Subsequently, the profile likelihood function $\ell_T$ is smoothed using a Nadaraya-Watson kernel (Nadaraya (1965)). Based on 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 summarized next.

ALGORITHM 2 (**Fast Implementation Algorithm**). *Input: a sequence of observed p-dimensional vectors* $\{x^{(t)}, 1 \le t \le T\}$, *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* $\mathcal{P}\ell_T(\tau), \quad \tau \in \mathcal{T}_1$.

*(c) Compute the profile negative pseudo-log-likelihood over the interval* $[1, T]$ *by Nadaraya-Watson kernel smoothing:*

$$\widetilde{\mathcal{P}\ell_{1s}}(\tau) \stackrel{\text{def}}{=} \frac{\sum_{\tau_i \in \mathcal{T}_1} K_{h_\nu}(\tau, \tau_i)\, \ell(\tau_i; \widehat{\theta}_{1,\tau_i}, \widehat{\theta}_{2,\tau_i})}{\sum_{\tau_i \in \mathcal{T}_1} \ell\left(\tau_i; \widehat{\theta}_{1,\tau_i}, \widehat{\theta}_{2,\tau_i}\right)}, \; 1 \le \tau \le T.$$

*The first stage change-point estimate is then obtained as*

$$\widehat{\tau} = \underset{1<\tau<T}{\mathsf{Argmin}} \widetilde{\mathcal{P}\ell_{1s}}(\tau).$$

*(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* $\widehat{\widehat{\theta}}_{1,\tau}$ *and* $\widehat{\widehat{\theta}}_{2,\tau}$ *using steps (a) and (b) of the Basic Algorithm.*

*(e) Construct the second stage smoothed profile pseudo-likelihood*

$$\widetilde{\mathcal{P}\ell_{2s}}(\tau) \stackrel{\text{def}}{=} \frac{\sum_{\tau_i \in \mathcal{T}_2} K_{h_\nu}(\tau, \tau_i)\, \ell\left(\tau_i; \widehat{\widehat{\theta}}_{1,\tau_i}, \widehat{\widehat{\theta}}_{2,\tau_i}\right)}{\sum_{\tau_i \in \mathcal{T}_2} \ell\left(\tau_i; \widehat{\widehat{\theta}}_{1,\tau_i}, \widehat{\widehat{\theta}}_{2,\tau_i}\right)}, \ \min(\mathcal{T}_2) \leq \tau \leq \max(\mathcal{T}_2).$$

*The final change-point estimate is then given by*

$$\widehat{\widehat{\tau}} = \underset{\min(\mathcal{T}_2) \leq \tau \leq \max(\mathcal{T}_2)}{\textit{Argmin}} \ \widetilde{\mathcal{P}\ell_{2s}}(\tau).$$

## 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 (1) has $B_0(x_j) = x_j$, $B(x_j, x_k) = x_j x_k$ and $\mathsf{X} \equiv \{0, 1\}$. In all simulation setting the sample size is set to $T = 700$, and the true change-point is at $\tau_\star = 350$, while the network size $p$ varies from 40-100. All the simulation results 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 ∼10% of the entries are different than zero. Each off-diagonal element of $\theta^{(i)}_{\star jk}$, $(i = 1, 2)$ is drawn uniformly from $[-1, -0.5] \cup [0.5, 1]$ if there is an edge between nodes $j$ and $k$, otherwise $\theta^{(i)}_{\star jk} = 0$. All the diagonal entries are set to zero. Given the two matrices $\theta^{(1)}_\star$ and $\theta^{(2)}_\star$, we generate the data $\left\{X^{(t)}\right\}_{t=1}^{\tau_\star} \overset{\text{iid}}{\sim} g_{\theta^{(1)}_*}$ and $\left\{X^{(t)}\right\}_{t=\tau_\star+1}^{T} \overset{\text{iid}}{\sim} g_{\theta^{(2)}_*}$ by Gibbs sampling.

Different "signal strenghts" are considered, by setting the degree of similarity between $\theta^{(1)}_\star$ and $\theta^{(2)}_\star$ to 0%, 20% and 40%. The degree of similarity is the proportion of equal off-diagonal elements between $\theta^{(1)}_\star$ and $\theta^{(2)}_\star$. Thus, the difference $\|\theta^{(2)}_\star - \theta^{(1)}_\star\|_1$ becomes smaller for higher degree of similarity and as can be seen from Assumption H3, the estimation problem becomes harder in such cases.

The choice of the tuning parameters $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ were made based on Bayesian Information Criterion (BIC) where we search $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ over a grid $\Lambda$ and for each penalty parameter the $\lambda$ value that minimizes the BIC score (defined below) over $\Lambda$ is selected. If we define $\lambda_1^{BIC}$ and $\lambda_2^{BIC}$ as the selected $\lambda$ values for $\lambda_1$ and

$\lambda_2$ by BIC we have

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

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

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

For the fast algorithm (Algorithm 2), the first stage grid employed had a step size of 10 and ranged from 60 to 640, while 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), while there is a small degradation 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 negatives and can also be interpretated as (1-Type 1 error). On the other hand sensitivity is the proportion of true positives and can be interpreted as the power of the method. The results for Algorithm 2 for $p = 40, 60$ and $p = 100$, for the change-point estimates are given in Table 4, while the specificity, sensitivity and relative error of the estimated network structures are given in Table 5. These results show that Algorithm 2 has about 20% higher mean-squared error (MSE) compared to 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 5 times faster in a standard computing environment with 4 CPU cores. See also the results in Table 3 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 $\widetilde{\mathcal{P}\ell_{1s}}(\tau)$ and $\widetilde{\mathcal{P}\ell_{2s}}(\tau)$ from the first and second stage of Algorithm 2 are given in Figure 1.

Table 1: Change-point estimation results using the Basic Algorithm, for different percentages of similarity.

| $p$ | % of Similarity | $\widehat{\tau}$ | RMSE | CV |
|---|---|---|---|---|
| | 0 | 355 | 14.77 | 0.03 |
| 40 | 20 | 362 | 24.65 | 0.06 |
| | 40 | 375 | 38.49 | 0.08 |

Table 2: Specificity, sensitivity and relative error in estimating $\theta_\star^{(1)}$ and $\theta_\star^{(2)}$ from the Basic Algorithm, with different percentages of similarity.

| $p$ | % 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 |
| 40 | 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 |

Table 3: 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 |

Table 4: Change-point Estimation Results for different values of $p$ and different percentages of similarity for the Fast Implementation Algorithm.($T = 700$, $s_1 = s_2 = \frac{10p(p+1)}{2}\%$, $\tau^* = 354$)

| p | % of Similarity | $\widehat{\tau}$ | $\widehat{\widehat{\tau}}$ | RMSE | CV |
|---|---|---|---|---|---|
| | 0 | 360 | 360 | 17.89 | 0.04 |
| 40 | 20 | 363 | 361 | 30.07 | 0.08 |
| | 40 | 375 | 373 | 47.97 | 0.10 |
| | 0 | 357 | 356 | 23.05 | 0.06 |
| 60 | 20 | 388 | 386 | 43.20 | 0.08 |
| | 40 | 410 | 408 | 61.45 | 0.09 |
| | 0 | 356 | 355 | 35.93 | 0.10 |
| 100 | 20 | 408 | 401 | 62.89 | 0.10 |
| | 40 | 424 | 421 | 85.04 | 0.12 |

Table 5: Specificity, sensitivity and relative error of the two parameters for different values of $p$ and different percentages of similarity for the Fast Implementation Algorithm.

| p | % of Similarity | Specificity | | Sensitivity | | Relative error | |
|---|---|---|---|---|---|---|---|
| | | $\theta_*^{(1)}$ | $\theta_*^{(2)}$ | $\theta_*^{(1)}$ | $\theta_*^{(2)}$ | $\theta_*^{(1)}$ | $\theta_*^{(2)}$ |
| | 0 | 0.74 | 0.86 | 0.78 | 0.86 | 0.74 | 0.67 |
| 40 | 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 |
| | 0 | 0.81 | 0.83 | 0.77 | 0.82 | 0.75 | 0.66 |
| 60 | 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 |
| | 0 | 0.82 | 0.88 | 0.75 | 0.84 | 0.78 | 0.66 |
| 100 | 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 |

### *5.2. A community based network structure*

Next, we examine a setting similar to the one that emerges from the US Senate analysis 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

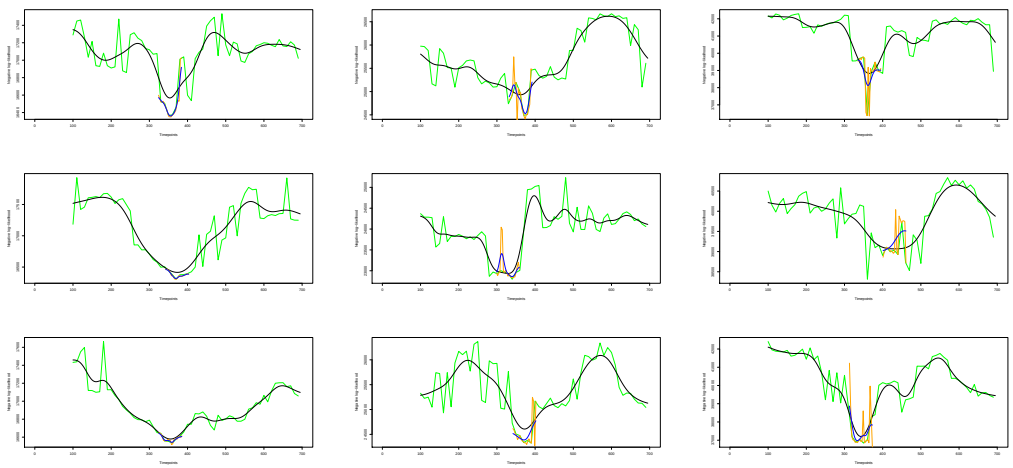

Fig. 1: Smoothed profile pseudo-log-likelihood functions from one run of Algorithm 2. Different values of similarity (0%, 20% and 40%) in rows. Different values of $p$ ($p = 40, 60$ & $100$) in column. The green curve is the non-smoothed profile pseudo-log-likelihood from Stage 1 of Algorithm 2, and the black curve is its smoothed version. The orange and the blue curve are respectively the non-smoothed and the smoothed profile pseudo-log-likelihood functions from Stage 2 of Algorithm 2.

one of them and decreased for the other after the occurrence of 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 (see Table 6), and 50% to community 2 (see Table 6), while 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 zeros. Given the two matrices $\theta^{(1)}_*$ and $\theta^{(2)}_*$, we generate data using the "BMN" package (Hoefling (2010)) as described earlier. The total sample size employed is $T = 1500$

and the true change-point is at $\tau^* = 750$. We choose the first stage grid comprising of 50 points with a step size of 27 and the second stage grid is chosen in a neighborhood of the first stage estimate with a step size of 3 with 20 points. We replicate the study 5 times and find that the estimated change-point averaged over the 5 replications as $\hat{\tau} = 768$. The relevant figure (see Figure 2) for this two community model is given below. The analysis indicates that our proposed methodology is able to estimate the true change-point sufficiently well in the presence of varying degrees of connections between two communities over two different time periods, a reassuring feature for the US Senate application presented next.

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

| Edges | Before | | | After | | |
|---|---|---|---|---|---|---|
| | comm 1 | comm 2 | between | comm 1 | comm 2 | between |
| positive | 50 | 63 | 0 | 52 | 21 | 0 |
| negative | 0 | 0 | 10 | 0 | 0 | 50 |
| Total | 50 | 63 | 10 | 52 | 21 | 50 |

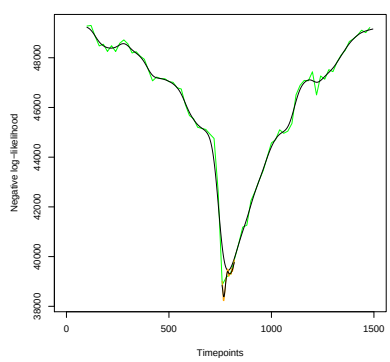


Fig. 2: Change-point estimate for the two community model with $p = 50$, $T = 1500$ and $\tau^*$=754

## 6. Application to Roll Call Data of the US Senate

The data examined correspond to voting records of the US Senate covering the period 1979 (96th Congress) to 2012 (112th Congress) and were obtained from the

website `www.voteview.com`. Specifically, for each of the 12129 votes cast during this period, the following information is recorded: the date that the vote occurred and the response to the bill/resolution under consideration -yes/no, or abstain- of the 100 Senate members. Due to 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 cabinet or other administrative positions, or physical demise. In order to hold the number of nodes fixed to 100 (the membership size 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.

Note that a significant number of the 12129 votes deal with fairly mundane procedural matters, thus resulting in nearly unanimous outcomes. Hence, only votes exhibiting conformity less than 75% (yes/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/no) of that member's party majority position on that particular vote. Note that other imputation methods of missing values were employed: (i) replacing all missing values by the value (yes/no) representing the winning majority on that bill and (ii) 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 Supplement.

Finally, the yes/no votes were encoded as 1/0, respectively. Under the posited model, votes are considered as i.i.d. 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 two year period that a Congress serves and probably even beyond that due to the slow turnover of Senate members. Nevertheless, the proposed model serves as a *working model* that captures essential features of the evolving voting dependency structure between Senate seats over time.

The likelihood function together with an estimate of a change-point are depicted in Figure 5 based on the Fast Implementation Algorithm presented in Section 4. We choose our first stage grid with a step-size of 50 that yields 157 points excluding time points close to both boundaries. In the second stage, we choose a finer-resolution grid with a step size of 20 in a neighborhood of the first stage change-point estimate. The vote corresponding to the change point occurred on January 17, 1995

at the beginning of the tenure of the 104th Congress. This change-point comes at the footsteps of the November 1994 election that witnessed the Republican Party 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", 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 heatmaps of the adjacency matrices in Figure 6. To obtain stable estimates of the respective 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 (blue), mixed (yellow) and Republican (red). 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 one. This means that if a seat was held for more than 5 out of the 8 Congresses in the pre change-point period and similarly 6 out of 9 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.

In the depicted heatmaps, the ordering 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 green dots represent positive edge weights, mostly corresponding to within categories interactions, while black dots represent negative edge weights, mostly between category interactions. It can be clearly seen an emergence of a significant number of black dots in the post change-point regimes, indicative of sharper disagreements between political parties and thus increased polarization. Further, it 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 further explore the reasons behind the presence of a change-point, we provide some network statistics in Figure 3 and Figure 4. Specifically, the two figures

present the proportion of positive and negative edges, before and after the estimated change-point using two different methods for selecting the penalty tuning parameters; an analogue of the Bayesian Information Criterion and threshold 0.8 for the stability selection method respectively. The patterns shown across the figures for the two different methods are very similar- high proportion of positive edges within groups and very low or almost negligible proportion 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 be accounted 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 same from pre to post change-point regime under BIC and stability selection both whereas the proportion of positive edges between the two groups decrease and the proportion of negative edges between them tend to increase from pre to post change-point regime for both the 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 heatmap.

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" one, in both the pre and post change-point network structures, whereas in terms of clustering coefficient values the "republican" group is ahead of the "democrat" one 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.

Table 7: Different network statistic values for stability selection with threshold=0.9 and 0.8 respectively

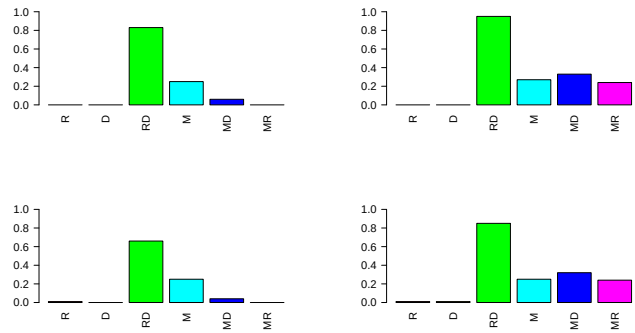


Fig. 3: Proportion of negative edges for network structures before (left figure) and after (right figure) the estimated change-point for BIC and stability selection with threshold=0.8

| Methods | Network Statistic | Before | | | After | | |
|---|---|---|---|---|---|---|---|
| | | Rep | Dem | Mixed | Rep | Dem | Mixed |
| Stable (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 |
| | | | | | | | |
| Stable (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 |

## References

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

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, **4**, 384-414.

Bai, J.(2010). Estimation of a change-point in multiple regression models. *The Review of Economics and Statistics*, **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.

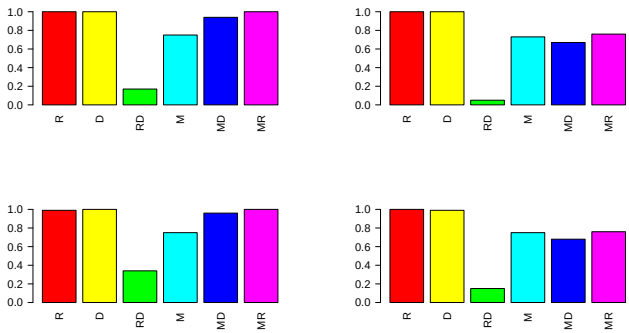


Fig. 4: Proportion of positive edges for network structures before (left figure) and after (right figure) the estimated change-point for BIC and stability selection with threshold=0.8

Basu, S. and Michailidis, G. (2015). Estimation in high dimensional vector autoregressive models. *Ann. Statist.* To Appear.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B.*, **36**, 192-236.

Bhattcharya, K. P. (1974). Maximum likelihood estimation of a change-point in the distribution of the independent random variables: General Multiparameter Case*J. Mult. Anls.*, **23**, 183-208.

Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199-227.

Bickel, P. J., Ritov, Y. and Tsybakov, A.B. (2009) Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705-1732.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.*, **17**, 453-510.

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

Drton, M. and Perlman, M.D. (2004). Model selection for Gaussian concentration graphs. *Biometrika.*, **91**, 591-602.

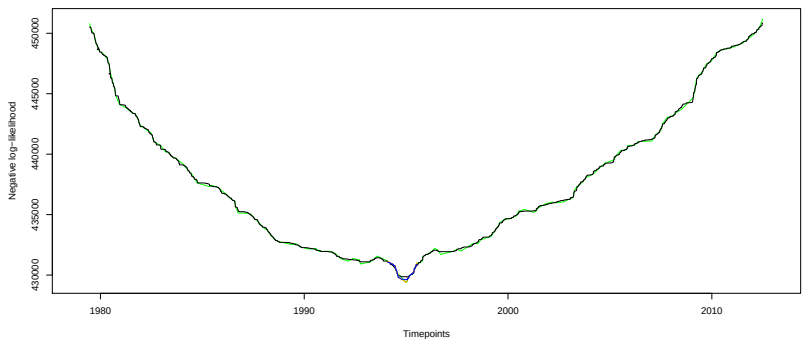


Fig. 5: Estimate of the change-point for the combined US senate data from 1979-2012

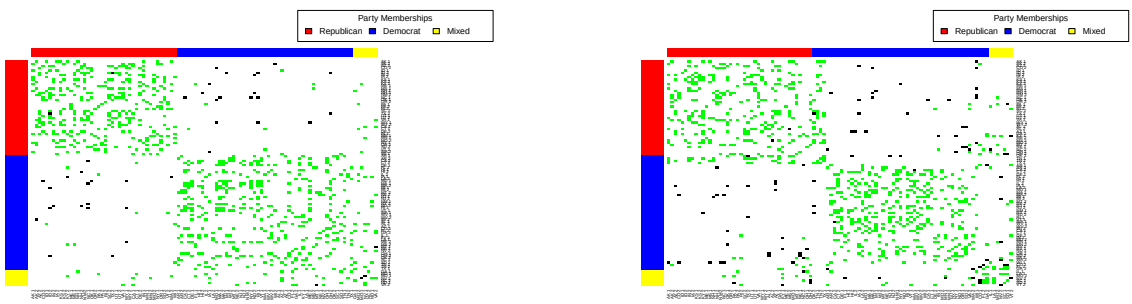


Fig. 6: Heatmap of the stable network structures before and after the estimated change-point

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1-22

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Joint structure estimation for categorical markov networks. *Tech. rep.*, Univ. of Michigan.

Han, F. and Liu, H. (2006). A direct estimation of high dimensional stationary vector autoregressions*arXiv:1307.0293v2 [stat.ML]* .

Hanneke, S. and Xing, P. E. (2006). Discrete temporal models of social networks. *Lecture Notes in Computer Science.*, **4503**, 115-125.

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

Hinkley, V. D. (1972). Time-ordered classification. *Biometrika.*, **59**, 509-523.

Hoefling, H.(2010). BMN: The pseudo-likelihood method for pairwise binary markov networks. *R package version 1.02, 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.

Hurvich, M. C., Simonoff, S.J. and Tsai, C. (1998). Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *J. R. Stat. Soc. Ser. B.*, **60**, 271-293.

Kolar, M., Song, L., Ahmed, A. and Xing, P. E. (2010). Estimating Time varying Networks. *Ann. App. Statist.*, **4**, 94-123.

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

Kosorok, R. M. (2012). Introduction to empirical processes and semiparametric inference. *Springer Series in Statistics.*

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix. *Ann. Statist.*, **37**, 4254-4278.

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

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

Massart, P. (2007). Concentration inequalities and model selection. *Springer Verlag.*

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. *J. R. Statist. Soc. B*, **72**, 417-473.

Moody, J. and Mucha, P.(2013). Portrait of political party polarization *Network Science*, **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. *Theory Prob. Applic.*, **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.

Pinelis, I. (2006). On the normal domination of (super) martingales. *Electronic Journal of Probability*, **39**, 1049-1070.

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.

Rothman, A. J., Bickel P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, **2**, 494-515.

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

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

Van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, **3**, 1360-1392.

Van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes. *Springer Series in Statistics.*

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

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

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika.*, **94**, 19-35.

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

# Supplementary Information

Although our main motivation is in discrete graphical models, the proposed methodology can be applied more broadly for model-based change-point estimation. With this in mind, we shall prove a more general result that can be useful with other high-dimensional change-point estimation problems. Theorem 1 follows as a special case.

## S7. High-dimensional model-based change-point detection

Let $\{X^{(t)},\ 1 \le t \le T\}$ be a sequence of $\mathbb{R}^p$-valued independent random variables. Let $\Theta \subseteq \mathbb{R}^d$ be an open, non-empty convex parameter space equipped with the Euclidean inner product $\langle \cdot, \cdot \rangle$, and norm$\| \cdot \|_2$. We will also use the $\ell^1$-norm $\|\theta\|_1 \stackrel{\text{def}}{=} \sum_{j=1}^d |\theta_j|$, and the $\ell^\infty$-norm $\|\theta\|_\infty \stackrel{\text{def}}{=} \max_{1 \le j \le d} |\theta_j|$. We assume that there exists a change point $\tau_\star \in \{1, \ldots, T-1\}$, parameters $\theta_\star^{(1)}, \theta_\star^{(2)} \in \Theta$, such that for $t = 1, \ldots, \tau_\star$, $X^{(t)} \sim g^{(t)}_{\theta_\star^{(1)}}$, and for $t = \tau_\star + 1, \ldots, T$, $X^{(t)} \sim g^{(t)}_{\theta_\star^{(2)}}$, where $g^{(t)}_{\theta_\star^{(1)}}$ and $g^{(t)}_{\theta_\star^{(2)}}$ are probability densities on $\mathbb{R}^p$. The goal is to estimate $\tau_\star, \theta_\star^{(1)}, \theta_\star^{(2)}$. This setting includes the Markov random field setting (our main motivation), where $g^{(t)}_{\theta_\star^{(1)}}$ and $g^{(t)}_{\theta_\star^{(2)}}$ does not depend $t$. It also includes regression models where the index $t$ in the distributions $g^{(t)}_{\theta_\star^{(1)}}$ and $g^{(t)}_{\theta_\star^{(2)}}$ accounts for the covariates of subject $t$.

For $t = 1 \ldots, T$, let $(\theta, x) \mapsto \phi_t(\theta, x)$ be jointly measurable functions on $\Theta \times \mathbb{R}^p$, such that $\theta \mapsto \phi_t(\theta, x)$ is convex and continuously differentiable for all $x \in \mathbb{R}^p$. We define

$$\ell_T(\tau; \theta_1, \theta_2) \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{\tau} \phi_t(\theta_1, X^{(t)}) + \frac{1}{T} \sum_{t=\tau+1}^{T} \phi_t(\theta_2, X^{(t)}),$$

and we consider the change-point estimator $\tau_\star$ given by

$$\widehat{\tau} = \operatorname*{Argmin}_{\tau \in \mathcal{T}} \ell_T(\tau; \widehat{\theta}_{1,\tau}, \widehat{\theta}_{2,\tau}), \tag{S22}$$

for a non-empty search domain $\mathcal{T} \subset \{1, \ldots, T\}$, where for each $\tau \in \mathcal{T}$, $\widehat{\theta}_{1,\tau}$ and $\widehat{\theta}_{2,\tau}$ are defined as

$$\widehat{\theta}_{1,\tau} \stackrel{\text{def}}{=} \underset{\theta \in \Theta}{\mathsf{Argmin}} \left[ \frac{1}{T} \sum_{t=1}^{\tau} \phi_t(\theta, X^{(t)}) + \lambda_{1,\tau} \|\theta\|_1 \right],$$

and

$$\widehat{\theta}_{2,\tau} \stackrel{\text{def}}{=} \underset{\theta \in \Theta}{\mathsf{Argmin}} \left[ \frac{1}{T} \sum_{t=\tau+1}^{T} \phi_t(\theta, X^{(t)}) + \lambda_{2,\tau} \|\theta\|_1 \right],$$

for some positive penalty parameters $\lambda_{1,\tau}, \lambda_{2,\tau}$. Note that by allowing the use of user-defined learning functions $\phi_t$, our framework can be used to analyze maximum likelihood and maximum pseudo-likelihood change-point estimators.

For $\tau \in \{1, \ldots, T-1\}$, we set

$$\mathcal{G}_\tau^1 \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{\tau} \nabla \phi_t(\theta_\star^{(1)}, X^{(t)}), \quad \text{and} \quad \mathcal{G}_\tau^2 \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla \phi_t(\theta_\star^{(2)}, X^{(t)}),$$

where $\nabla \phi_t(\theta, x)$ denotes the partial derivative of $u \mapsto \phi_t(u, x)$ at $\theta$. Also for $\tau \in \{1, \ldots, T-1\}$, and for $\theta \in \Theta$, we define,

$$\mathcal{L}_1(\tau, \theta) \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{\tau} \left[ \phi_t(\theta, X^{(t)}) - \phi_t(\theta_\star^{(1)}, X^{(t)}) - \left\langle \nabla \phi_t(\theta_\star^{(1)}, X^{(t)}), \theta - \theta_\star^{(1)} \right\rangle \right],$$

$$\text{and} \quad \mathcal{L}_2(\tau, \theta) \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=\tau+1}^{T} \left[ \phi_t(\theta, X^{(t)}) - \phi_t(\theta_\star^{(2)}, X^{(t)}) - \left\langle \nabla \phi_t(\theta_\star^{(2)}, X^{(t)}), \theta - \theta_\star^{(2)} \right\rangle \right].$$

For $j = 1, 2$, define $\mathcal{A}_j \stackrel{\text{def}}{=} \left\{ 1 \le k \le d : \theta_{\star k}^{(j)} \neq 0 \right\}$, $s_j = |\mathcal{A}_j|$ , and

$$\mathbb{C}_j \stackrel{\text{def}}{=} \left\{ \theta \in \Theta : \sum_{k \in \mathcal{A}_j^c} |\theta_k^{(j)}| \le 3 \sum_{k \in \mathcal{A}_j} |\theta_k^{(j)}| \right\}. \tag{S23}$$

The curvature of the function $\mathcal{L}_j(\tau, \cdot)$ is not always best described with the usual quadratic function $\theta \mapsto \|\theta - \theta_\star^{(j)}\|_2^2$. We will need a more flexible framework, in order to handle $\mathcal{L}_j(\tau, \cdot)$ in the case of discrete Markov random fields. Let $\mathsf{r} : [0, \infty) \to [0, \infty)$ be continuous function such that $x \mapsto \mathsf{r}(x)/x$ is strictly increasing and $\lim_{x \downarrow 0} \mathsf{r}(x)/x = 0$. We call $\mathsf{r}$ a rate function, and for $a > 0$, we define $\Psi_{\mathsf{r}}(a) \stackrel{\text{def}}{=} \inf\{x > 0 : \ \mathsf{r}(x)/x \ge a\}$ ($\inf \emptyset = +\infty$). For $\tau \in \{1, \ldots, T-1\}$, $\lambda > 0$, a rate

function $\mathsf{r}$, $c > 0$, and for $j = 1, 2$ we work with the event

$$\mathcal{E}_\tau^j(\lambda, \mathsf{r}, c) \stackrel{\text{def}}{=} \left\{ \|G_\tau^j\|_\infty \le \frac{\lambda}{2}, \quad \inf_{\theta \neq \theta_\star^{(j)},\, \theta - \theta_\star^{(j)} \in \mathbb{C}_j} \frac{\mathcal{L}_j(\tau, \theta)}{\mathsf{r}\left(\|\theta - \theta_\star^{(j)}\|_2\right)} \ge \frac{\tau}{T}, \right.$$
$$\left. \sup_{\theta \neq \theta_\star^{(j)},\, \theta - \theta_\star^{(j)} \in \mathbb{C}_j} \frac{\mathcal{L}_j(\tau, \theta)}{\|\theta - \theta_\star^{(j)}\|_2^2} \le \frac{\tau}{T}\frac{c}{2} \right\}.$$

Define

$$\kappa_0^{(t)} \stackrel{\text{def}}{=} \begin{cases} \mathbb{E}\left[\phi_t(\theta_\star^{(2)}, X^{(t)}) - \phi_t(\theta_\star^{(1)}, X^{(t)})\right] & \text{if } t \le \tau_\star \\ \mathbb{E}\left[\phi_t(\theta_\star^{(1)}, X^{(t)}) - \phi_t(\theta_\star^{(2)}, X^{(t)})\right] & \text{if } t > \tau_\star \end{cases},$$

and

$$U^{(t)} \stackrel{\text{def}}{=} \begin{cases} \phi_t(\theta_\star^{(2)}, X^{(t)}) - \phi_t(\theta_\star^{(1)}, X^{(t)}) - \kappa_0^{(t)} & \text{if } t \le \tau_\star \\ \phi_t(\theta_\star^{(1)}, X^{(t)}) - \phi_t(\theta_\star^{(2)}, X^{(t)}) - \kappa_0^{(t)} & \text{if } t > \tau_\star \end{cases}.$$

We make the following assumption.

**A** 1. *There exist finite constants $\sigma_{0t} > 0$ such that*

$$\mathbb{E}\left(e^{xU^{(t)}}\right) \le e^{x^2\sigma_{0t}^2\|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2/2}, \quad \textit{for all} \quad x > 0.$$

*Furthermore, there exist $B_0 > 0$, $\bar{\sigma}_0^2 > 0$, $\bar{\kappa}_0 > 0$ such that for all integer $k \ge B_0$,*

$$\min\left(\frac{1}{k}\sum_{t=\tau_\star - k + 1}^{\tau_\star} \kappa_0^{(t)},\ \frac{1}{k}\sum_{t=\tau_\star+1}^{\tau_\star + k} \kappa_0^{(t)}\right) \ge \bar{\kappa}_0 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2, \tag{S24}$$

*and*

$$\max\left(\frac{1}{k}\sum_{t=\tau_\star - k + 1}^{\tau_\star} \sigma_{0t}^2, \frac{1}{k}\sum_{t=\tau_\star+1}^{\tau_\star + k} \sigma_{0t}^2\right) \le \bar{\sigma}_0^2. \tag{S25}$$

THEOREM S1. *Assume A1, and $\theta_\star^{(1)} \neq \theta_\star^{(2)}$. Suppose that $\hat{\tau}$ is defined over a search domain $\mathcal{T} \ni \tau_\star$, and with penalty $\lambda_{j,\tau} > 0$ (for $j = 1, 2$). For $j = 1, 2$, take a rate function $\mathsf{r}_j$, constant $c_j > 0$, and define $\mathcal{E} \stackrel{\text{def}}{=} \cap_{\tau \in \mathcal{T}} \mathcal{E}_\tau^1(\lambda_{1,\tau}, \mathsf{r}_1, c_1) \cap \mathcal{E}_\tau^2(\lambda_{2,\tau}, \mathsf{r}_2, c_2)$. Set*

$$\delta(\tau) \stackrel{\text{def}}{=} \Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\left[2s_1^{1/2}T\lambda_{1,\tau} + \tau\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\right]$$
$$+ \Psi_{\mathsf{r}_2}\left(6\left(\frac{T}{T-\tau}\right)s_2^{1/2}\lambda_{2,\tau}\right)\left[2s_2^{1/2}T\lambda_{2,\tau} + (T-\tau)\Psi_{\mathsf{r}_2}\left(6\left(\frac{T}{T-\tau}\right)s_2^{1/2}\lambda_{2,\tau}\right)\right],$$

$\delta \stackrel{\text{def}}{=} \sup_{\tau\in\mathcal{T}} \delta(\tau)$, *and* $B \stackrel{\text{def}}{=} \max\left(B_0, \frac{4\delta}{\bar{\kappa}_0\|\theta_\star^{(2)}-\theta_\star^{(1)}\|_2^2}\right)$, *with* $B_0$ *as in A1. Then*

$$\mathbb{P}\left(|\hat{\tau}-\tau_\star|>B\right) \leq 2\mathbb{P}(\mathcal{E}^c) + \frac{4\exp\left(-\frac{\bar{\kappa}_0^2\delta}{2\bar{\sigma}_0^2}\right)}{1-\exp\left(-\frac{\bar{\kappa}_0^2\|\theta_\star^{(2)}-\theta_\star^{(1)}\|_2^2}{8\bar{\sigma}_0^2}\right)}. \tag{S26}$$

Proof. The starting point of the proof is the following variant of a result due to Neghaban et al. (2010).

Lemma 1. *Fix* $\tau\in\{1,2,\ldots,T-1\}$. *On* $\mathcal{E}_\tau^1(\lambda_{1,\tau},\mathsf{r}_1,c_1)\cap\mathcal{E}_\tau^2(\lambda_{2,\tau},\mathsf{r}_2,c_2)$, $\hat{\theta}_{j,\tau}-\theta_\star^{(j)}\in\mathbb{C}_j$, *(*$j=1,2$*), where* $\mathbb{C}_j$ *is defined in (S23), and*

$$\|\hat{\theta}_{1,\tau}-\theta_\star^{(1)}\|_2 \leq \Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right),$$
$$\text{and}\quad \|\hat{\theta}_{2,\tau}-\theta_\star^{(2)}\|_2 \leq \Psi_{\mathsf{r}_2}\left(6\left(\frac{T}{T-\tau}\right)s_2^{1/2}\lambda_{2,\tau}\right). \tag{S27}$$

Proof. We prove the first inequality. The second follows similarly. We set

$$\mathcal{U}(\theta) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{\tau}\phi_t(\theta,X^{(t)}) + \lambda_{1,\tau}\|\theta\|_1 - \left(\frac{1}{T}\sum_{t=1}^{\tau}\phi_t(\theta_\star^{(1)},X^{(t)}) + \lambda_{1,\tau}\|\theta_\star^{(1)}\|_1\right).$$

Since $\hat{\theta}_{1,\tau} = \mathsf{Argmin}_{\theta\in\Theta}\left[\frac{1}{T}\sum_{t=1}^{\tau}\phi_t(\theta,X^{(t)}) + \lambda_{1,\tau}\|\theta\|_1\right]$, and using the convexity of the functions $\phi_t$ we have

$$0 \geq \mathcal{U}(\hat{\theta}_{1,\tau}) \geq \left\langle G_\tau^1, \hat{\theta}_{1,\tau}-\theta_\star^{(1)}\right\rangle + \lambda_{1,\tau}\left(\|\hat{\theta}_{1,\tau}\|_1 - \|\theta_\star^{(1)}\|_1\right).$$

On $\mathcal{E}_\tau^1(\lambda_{1,\tau},\mathsf{r}_1,c_1)$, $\|G_\tau^1\|_\infty \leq \lambda_{1,\tau}/2$. Using this and some easy algebra as in Neghaban et al. (2010), shows that $\hat{\theta}_{1,\tau}-\theta_\star^{(1)}\in\mathbb{C}_1$. Set $b = \Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)$. We will show that for all $\theta\in\mathbb{R}^d$ such that $\theta-\theta_\star^{(1)}\in\mathbb{C}_1$, and $\|\theta-\theta_\star^{(1)}\|_2 > b$, we have $\mathcal{U}(\theta)>0$. Since $\mathcal{U}(\hat{\theta}_{1,\tau})\leq 0$, and $\hat{\theta}_{1,\tau}-\theta_\star^{(1)}\in\mathbb{C}_1$, the claim that $\|\theta-\theta_\star^{(1)}\|_2\leq b$ follows. On the event $\mathcal{E}_\tau^1(\lambda_{1,\tau},\mathsf{r}_1,c_1)$, and for $\theta-\theta_\star^{(1)}\in\mathbb{C}_1$, we have

$$\begin{aligned}
\mathcal{U}(\theta) &= \left\langle G_\tau^1, \theta-\theta_\star^{(1)}\right\rangle + \mathcal{L}_1(\tau,\theta) + \lambda_{1,\tau}\left(\|\theta\|_1 - \|\theta_\star^{(1)}\|_1\right)\\
&\geq \frac{\tau}{T}\mathsf{r}_1(\|\theta-\theta_\star^{(1)}\|_2) - \frac{3\lambda_{1,\tau}}{2}\|\theta-\theta_\star^{(1)}\|_1\\
&\geq \frac{\tau}{T}\left[\mathsf{r}_1(\|\theta-\theta_\star^{(1)}\|_2) - 6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\|\theta-\theta_\star^{(1)}\|_2\right].
\end{aligned}$$

Using the definition of $\Psi_{\mathsf{r}_1}$, we then see that $\mathcal{U}(\theta)>0$ for $\|\theta-\theta_\star^{(1)}\|_2>b$. This ends the proof. □

The next result follows easily.

LEMMA 2. *Fix* $\tau \in \{1, 2, \ldots, T-1\}$. *On* $\mathcal{E}_\tau^1(\lambda_{1,\tau}, \mathsf{r}_1, c_1) \cap \mathcal{E}_\tau^2(\lambda_{2,\tau}, \mathsf{r}_2, c_2)$,

$$\left|\ell_T(\tau, \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}) - \ell_T(\tau, \theta_\star^{(1)}, \theta_\star^{(2)})\right| \leq \frac{\delta(\tau)}{T},$$

*where*

$$\delta(\tau) \stackrel{\text{def}}{=} \Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\left[2s_1^{1/2}T\lambda_{1,\tau} + \frac{\tau c_1}{2}\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\right]$$
$$+\Psi_{\mathsf{r}_2}\left(6\left(\frac{T}{T-\tau}\right)s_2^{1/2}\lambda_{2,\tau}\right)\left[2s_2^{1/2}T\lambda_{2,\tau} + \frac{(T-\tau)c_2}{2}\Psi_{\mathsf{r}_2}\left(6\left(\frac{T}{T-\tau}\right)s_2^{1/2}\lambda_{2,\tau}\right)\right].$$

PROOF.

$$\ell_T(\tau, \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}) - \ell_T(\tau, \theta_\star^{(1)}, \theta_\star^{(2)}) = \frac{1}{T}\sum_{t=1}^{\tau}\left[\phi_t(\hat{\theta}_{1,\tau}, X^{(t)}) - \phi_t(\theta_\star^{(1)}, X^{(t)})\right]$$
$$+ \frac{1}{T}\sum_{t=\tau+1}^{T}\left[\phi_t(\hat{\theta}_{2,\tau}, X^{(t)}) - \phi_t(\theta_\star^{(2)}, X^{(t)})\right].$$

From the definition

$$\frac{1}{T}\sum_{t=1}^{\tau}\left[\phi_t(\hat{\theta}_{1,\tau}, X^{(t)}) - \phi_t(\theta_\star^{(1)}, X^{(t)})\right] = \left\langle G_\tau^1, \hat{\theta}_{1,\tau} - \theta_\star^{(1)}\right\rangle + \mathcal{L}_1(\tau, \hat{\theta}_{1,\tau}).$$

On $\mathcal{E}_\tau^1(\lambda_{1,\tau}, \mathsf{r}_1, c_1)$, and using Lemma 1, we have

$$\left|\left\langle G_\tau^1, \hat{\theta}_{1,\tau} - \theta_\star^{(1)}\right\rangle\right| \leq \frac{\lambda_{1,\tau}}{2}\|\hat{\theta}_{1,\tau} - \theta_\star^{(1)}\|_1 \leq 2s_1^{1/2}\lambda_{1,\tau}\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right),$$

and

$$\mathcal{L}_1(\tau, \hat{\theta}_{1,\tau}) \leq \frac{\tau}{T}\frac{c_1}{2}\|\hat{\theta}_{1,\tau} - \theta_\star^{(1)}\|_2^2 \leq \frac{\tau c_1}{2T}\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)^2.$$

Hence

$$\left|\frac{1}{T}\sum_{t=1}^{\tau}\left[\phi_t(\hat{\theta}_{1,\tau}, X^{(t)}) - \phi_t(\theta_\star^{(1)}, X^{(t)})\right]\right|$$
$$\leq \frac{1}{T}\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\left[2s_1^{1/2}T\lambda_{1,\tau} + \frac{\tau c_1}{2}\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\right].$$

A similar bound holds for the second term, and the lemma follows easily. □

We are now in position to prove Theorem S1. We have

$$\mathbb{P}\left(|\hat{\tau} - \tau_\star| > B\right) = \mathbb{P}\left(\hat{\tau} > \tau_\star + B\right) + \mathbb{P}\left(\hat{\tau} < \tau_\star - B\right).$$

We bound the first term $\mathbb{P}\left(\hat{\tau} > \tau_\star + B\right)$. The second term follows similarly by working with the reversed sequence $X^{(T)}, \ldots, X^{(1)}$.

For $\tau > \tau_\star$, we shall use $\ell_T\left(\tau\right)$ instead of $\ell_T\left(\tau; \hat{\theta}_{1,\tau}, \hat{\theta}_{2,\tau}\right)$ for notational convenience, and we define $r_T\left(\tau\right) \stackrel{\text{def}}{=} \ell_T(\tau) - \ell_T\left(\tau, \theta_\star^{(1)}, \theta_\star^{(2)}\right)$. We have

$$\begin{aligned}
\ell_T\left(\tau\right) &= \ell_T\left(\tau, \theta_\star^{(1)}, \theta_\star^{(2)}\right) + r_T(\tau), \\
&= \left[\ell_T\left(\tau, \theta_\star^{(1)}, \theta_\star^{(2)}\right) - \ell_T\left(\tau_\star, \theta_\star^{(1)}, \theta_\star^{(2)}\right)\right] + \ell_T\left(\tau_\star, \theta_\star^{(1)}, \theta_\star^{(2)}\right) + r_T(\tau).
\end{aligned}$$

Hence

$$\ell_T(\tau) - \ell_T(\tau_\star) = \left[\ell_T\left(\tau, \theta_\star^{(1)}, \theta_\star^{(2)}\right) - \ell_T\left(\tau_\star, \theta_\star^{(1)}, \theta_\star^{(2)}\right)\right] + r_T(\tau) - r_T(\tau_\star). \quad \text{(S28)}$$

It is straightforward to check that for $\tau > \tau_\star$,

$$\ell_T\left(\tau, \theta_\star^{(1)}, \theta_\star^{(2)}\right) - \ell_T\left(\tau_\star, \theta_\star^{(1)}, \theta_\star^{(2)}\right) = \frac{1}{T} \sum_{t=\tau_\star+1}^{\tau} \left(\phi_t(\theta_\star^{(1)}, X^{(t)}) - \phi_t(\theta_\star^{(2)}, X^{(t)})\right).$$

Therefore, and using the definition of $U^{(t)}$ and $\kappa_0^{(t)}$, (S28) becomes

$$\ell_T(\tau) - \ell_T(\tau_\star) = \frac{1}{T} \sum_{t=\tau_\star+1}^{\tau} \kappa_0^{(t)} + \frac{1}{T} \sum_{t=\tau_\star+1}^{\tau} U^{(t)} + r_T(\tau) - r_T(\tau_\star). \quad \text{(S29)}$$

We conclude from Lemma 2 that on the event $\mathcal{E}$,

$$\ell_T(\tau) - \ell_T(\tau_\star) = \frac{1}{T} \sum_{t=\tau_\star+1}^{\tau} \kappa_0^{(t)} + \frac{1}{T} \sum_{t=\tau_\star+1}^{\tau} U^{(t)} + \epsilon_T(\tau),$$
$$\text{where } \; |\epsilon_T(\tau)| \leq \frac{2 \sup_{\tau \mathcal{T}} |\delta(\tau)|}{T} = \frac{2\delta}{T}. \quad \text{(S30)}$$

Therefore,

$$\mathbb{P}\left(\hat{\tau} > \tau + B\right) \leq \mathbb{P}(\mathcal{E}^c) + \sum_{j \geq 0,\, \tau_\star + \lceil B \rceil + j \in \mathcal{T}} \mathbb{P}\left(\mathcal{E},\, \hat{\tau} = \tau_\star + \lceil B \rceil + j\right).$$

Using (S30), we have

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E},\, \hat{\tau} = \tau_\star + \lceil B \rceil + j\right) &\leq \mathbb{P}\left(\mathcal{E},\, \ell_T(\tau_\star + \lceil B \rceil + j) \leq \ell_T(\tau_\star)\right) \\
&\leq \mathbb{P}\left(\left|\sum_{t=\tau_\star+1}^{\tau_\star + \lceil B \rceil + j} U^{(t)}\right| > \sum_{t=\tau_\star+1}^{\tau_\star + \lceil B \rceil + j} \kappa_0^{(t)} - 2\delta\right).
\end{aligned}$$

However, since $B > B_0$, by Assumption A1,

$$\sum_{t=\tau_\star+1}^{\tau_\star+\lceil B\rceil+j} \kappa_0^{(t)} - 2\delta \geq (\lceil B\rceil + j)\,\bar{\kappa}_0 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2 - 2\delta \geq \frac{1}{2}\,(\lceil B\rceil + j)\,\bar{\kappa}_0 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2.$$

The first part of A1 implies that the random variables $Z^{(t)}$ are sub-Gaussian, and by standard exponential bounds for sub-Gaussian random variables, we then have

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E},\ \ell_T(\tau_\star + \lceil B\rceil + j) \leq \ell_T(\tau_\star)\right] &\leq 2\exp\left(-\frac{(\lceil B\rceil + j)^2\,\bar{\kappa}_0^2 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^4}{8\|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2 \sum_{t=\tau_\star+1}^{\tau_\star+\lceil B\rceil+j} \sigma_{0t}^2}\right), \\
&\leq 2\exp\left(-\frac{(\lceil B\rceil + j)\,\bar{\kappa}_0^2 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2}{8\bar{\sigma}_0^2}\right),
\end{aligned}$$

where the last inequality uses (S25). We can conclude that

$$\begin{aligned}
\mathbb{P}\left[\hat{\tau} > \tau_\star + B\right] &\leq \mathbb{P}(\mathcal{E}^c) + 2\sum_{j\geq 0} \exp\left(-\frac{(\lceil B\rceil + j)\,\bar{\kappa}_0^2 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2}{8\bar{\sigma}_0^2}\right) \\
&\leq \mathbb{P}(\mathcal{E}^c) + 2\frac{\exp\left(-\frac{B\bar{\kappa}_0^2 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2}{8\bar{\sigma}_0^2}\right)}{1 - \exp\left(-\frac{\bar{\kappa}_0^2 \|\theta_\star^{(2)} - \theta_\star^{(1)}\|_2^2}{8\bar{\sigma}_0^2}\right)}, \qquad \text{(S31)}
\end{aligned}$$

as claimed. □

## S8. Proof of Theorem 1

We will deduce Theorem 1 from Theorem S1. We take $\Theta$ as $\mathcal{M}_p$, the set of all $p \times p$ real symmetric matrices, equipped with the (modified) Frobenius inner product $\langle \theta, \vartheta \rangle_{\mathsf{F}} \stackrel{\text{def}}{=} \sum_{k\leq j} \theta_{jk}\vartheta_{jk}$, and the associated norm $\|\theta\|_{\mathsf{F}} \stackrel{\text{def}}{=} \sqrt{\langle \theta, \theta\rangle}$. With this inner product, we identify $\mathcal{M}_p$ with the Euclidean space $\mathbb{R}^d$, with $d = p(p+1)/2$. This puts us in the setting of Theorem S1.

We will use the following notation. If $u \in \mathbb{R}^q$, for some integer $q \geq 1$, and $A$ is an ordered subset of $\{1, \ldots, q\}$, we define $u_A \stackrel{\text{def}}{=} (u_j,\, j \in A)$, and $u_{-j}$ is a shortcut for $u_{\{1,\ldots,q\}\setminus\{j\}}$. We define the function $B_{jk}(x, y) = B_0(x)$ if $j = k$, and $B_{jk}(x, y) = B(x, y)$ if $j \neq k$.

In the present case, the function $\phi_t$ is $\phi$ as given in (5), and does not depend on $t$. The following properties of the conditional distribution (3) will be used below. It is well known (and easy to prove using Fisher's identity) that the function $\theta \mapsto \phi(\theta, x)$

is Lispchitz and

$$|\phi(\theta, x) - \phi(\vartheta, x)| \leq 2c_0 \|\theta - \vartheta\|_1, \;\; \theta, \vartheta \in \mathcal{M}_p, \; x \in \mathsf{X}^p, \tag{S32}$$

where $c_0$ is as in (9). From the expression (3) of the conditional densities, using straightforward algebra, it is easy to show that the negative log-pseudo-likelihood function $\phi(\theta, x)$ satisfies the following. For all $\theta, \Delta \in \mathcal{M}_p$, and $x \in \mathsf{X}^p$,

$$\begin{aligned}\phi(\theta + \Delta, x) - \phi(\theta, x) - \langle \nabla_\theta \phi(\theta, x), \Delta \rangle_{\mathsf{F}} \\ = \sum_{j=1}^{p} \left[ \log Z^{(j)}_{\theta+\Delta}(x) - \log Z^{(j)}_{\theta}(x) - \sum_{k=1}^{p} \Delta_{jk} \frac{\partial}{\partial \theta_{jk}} \log Z^{(j)}_{\theta}(x) \right].\end{aligned} \tag{S33}$$

Furthermore by Taylor expansion, we have

$$\begin{aligned}&\log Z^{(j)}_{\theta+\Delta}(x) - \log Z^{(j)}_{\theta}(x) - \sum_{k=1}^{p} \Delta_{jk} \frac{\partial}{\partial \theta_{jk}} \log Z^{(j)}_{\theta}(x) \\ &= \int_0^1 (1-t) \mathsf{Var}_{\theta + t\Delta} \left( \sum_{k=1}^{p} \Delta_{jk} B_{jk}(X_j, X_k) | X_{-j} \right) \mathrm{d}t \leq \frac{c_0^2}{2} \left( \sum_{k=1}^{p} |\Delta_{jk}| \right)^2.\end{aligned} \tag{S34}$$

By the self-concordant bound derived in Atchadé (2014) Lemma A2, we have

$$\begin{aligned}&\log Z^{(j)}_{\theta+\Delta}(x) - \log Z^{(j)}_{\theta}(x) - \sum_{k=1}^{p} \Delta_{jk} \frac{\partial}{\partial \theta_{jk}} \log Z^{(j)}_{\theta}(x) \\ &\qquad \geq \frac{1}{2 + c_0 \sum_{k=1}^{p} |\Delta_{jk}|} \mathsf{Var}_{\theta} \left( \sum_{k=1}^{p} \Delta_{jk} B_{jk}(X_j, X_k) | X_{-j} \right).\end{aligned} \tag{S35}$$

Proof (Proof of Theorem 1). Let us first show that under assumption H3 of Theorem 1, A1 holds. Since in this case $\phi_t$ does not actually depend on $t$, we can take $B_0 = 1$ in A1, and (S24) follows automatically from H3 with $\bar{\kappa}_0 = \kappa / \|\theta^{(2)}_\star - \theta^{(1)}_\star\|_2^2$. Also, (S32) implies that $|U^{(t)}| \leq 4c_0 \|\theta^{(2)}_\star - \theta^{(1)}_\star\|_1 \leq 4c_0 s^{1/2} \|\theta^{(2)}_\star - \theta^{(1)}_\star\|_2$, where $s$ denotes the number of non-zero entries of $\theta^{(2)} - \theta^{(1)}_\star$. Hence for all $x > 0$,

$$\mathbb{E}\left( e^{xU^{(t)}} \right) \leq \exp\left( 8x^2 c_0^2 s \|\theta^{(2)}_\star - \theta^{(1)}_\star\|_2^2 \right).$$

This establishes the sub-Gaussian condition of A1, and (S25) holds with $\bar{\sigma}_0^2 = 16 c_0^2 s$.

For $j = 1, 2$, let $\lambda_{1,\tau}$, $\lambda_{2,\tau}$ as in (8). We will apply Theorem S1 with $c_j = 64 c_0 s_j$, the rate function $\mathsf{r}_j(x) = \frac{\rho_j x^2}{2 + 4c_0 s_j^{1/2} x}$, $x > 0$, and with the event $\mathcal{E} =$

$\bigcap_{\tau\in\mathcal{T}}\left[\mathcal{E}^1_\tau(\lambda_{1,\tau},\mathsf{r}_1,c_1)\cap\mathcal{E}^2_\tau(\lambda_{2,\tau},\mathsf{r}_2,c_2)\right]$, where the search domain $\mathcal{T}$ satisfies (15), (16), and (18). Notice that if $\mathsf{r}(x)=\rho x^2/(2+bx)$, $\rho,b>0$, is a rate function, then for $a>0$, $\Psi_{\mathsf{r}}(a)\stackrel{\text{def}}{=}\inf\{x>0:\ r(x)\geq ax\}\leq 4a/\rho$, provided that $2ba\leq\rho$. Hence

$$\Psi_{\mathsf{r}_1}\left(6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}\right)\leq\frac{4}{\rho_1}6\left(\frac{T}{\tau}\right)s_1^{1/2}\lambda_{1,\tau}=24\times 32c_2\frac{s_1^{1/2}}{\rho_1}\sqrt{\frac{\log(dT)}{\tau}},$$

provided that $\tau\geq(48\times 32)^2c_0^2\left(\frac{s_1}{\rho_1}\right)^2\log(dT)$. Therefore, given that all $\tau\in\mathcal{T}$ satisfies (18), with some simple algebra we see that there exists a universal constant $a$ that we can take as $a=(24\times 32\times 64)^2$, such that for all $\tau\in\mathcal{T}$,

$$\delta(\tau)\leq\delta=ac_0^2M\log(dT),$$

where

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

Therefore in Theorem S1, we can take $B=\frac{4ac_0^2M\log(dT)}{\kappa}$, and by the conclusion of Theorem S1,

$$\mathbb{P}\left[|\hat{\tau}-\tau_\star|>B\right]\leq 2\mathbb{P}(\mathcal{E}^c)+\frac{4\exp\left(-\frac{\delta}{32c_0^2s}\left(\frac{\kappa}{\|\theta_\star^{(2)}-\theta_\star^{(1)}\|_2^2}\right)^2\right)}{1-\exp\left(-\frac{\kappa^2}{2^7c_0^2s\|\theta_\star^{(2)}-\theta_\star^{(1)}\|_2^2}\right)}.$$

We show in Lemma 3 and Lemma 4 below that $\mathbb{P}(\mathcal{E}^c)\leq 8/d$, and this ends the proof.

□

LEMMA 3. *Let $\lambda_{1,\tau},\lambda_{2,\tau}$ be as in equation (8). Suppose that the search domain $\mathcal{T}$ is such that (15)-(16) hold. Then*

$$\mathbb{P}\left[\max_{\tau\in\mathcal{T}}\lambda_{1,\tau}^{-1}\left\|G_\tau^1\right\|_\infty>\frac{1}{2}\right]\leq\frac{2}{d},\quad\textit{and}\quad\mathbb{P}\left[\max_{\tau\in\mathcal{T}}\lambda_{2,\tau}^{-1}\left\|G_\tau^2\right\|_\infty>\frac{1}{2}\right]\leq\frac{2}{d},$$

*where $d=p(p+1)/2$.*

PROOF. We carry the details for the first bound. The second is done similarly by working with the reversed sequence $X^{(T)},\ldots,X^{(1)}$. Fix $1\leq j\leq i\leq p$, $t\in\mathcal{T}$, and define $V_{ij}^{(t)}\stackrel{\text{def}}{=}\frac{\partial}{\partial\theta_{ij}}\phi(\theta_\star^{(1)},X^{(t)})$. We calculate that

$$V_{ij}^{(t)}=\begin{cases}-B_0(X_i^{(t)})+\mathbb{E}_{\theta_\star^{(1)}}(B_0(X_i|X_{-i}^{(t)}) & \text{if } i=j\\ -2B(X_i^{(t)},X_j^{(t)})+\mathbb{E}_{\theta_\star^{(1)}}\left(B(X_i,X_j^{(t)})|X_{-i}^{(t)}\right)+\mathbb{E}_{\theta_\star^{(1)}}\left(B(X_i,X_j^{(t)})|X_{-j}^{(t)}\right) & \text{if } j<i.\end{cases}$$

In the above display the notation $\mathbb{E}_{\theta_\star^{(1)}}\left(B(X_i, X_j^{(t)})|X_{-i}^{(t)}\right)$ is defined as the function $z \mapsto \mathbb{E}_{\theta_\star^{(1)}}\left(B(X_i, z_j)|X_{-i} = z_{-i}\right)$ evaluated on $X^{(t)}$. Since $X^{(1:\tau_\star)} \overset{i.i.d}{\sim} g_{\theta_\star^{(1)}}$, it follows that $\mathbb{E}(V_{ij}^{(t)}) = 0$ for $t = 1, \ldots, \tau_\star$. We set $\mu_{ij} \overset{\text{def}}{=} \mathbb{E}(V_{ij}^{(\tau_\star+1)}) = \mathbb{E}(V_{ij}^{(t)})$ for $t = \tau_\star + 1, \ldots, T$. We also set $\bar{V}_{ij}^{(t)} \overset{\text{def}}{=} V_{ij}^{(t)} - \mathbb{E}\left(V_{ij}^{(t)}\right)$. It is easy to see that $|\bar{V}_{ij}^{(t)}| \leq 4c_0$, where $c_0$ is defined in (9) . With these notations, for $\tau \in \mathcal{T}$, we can write

$$(G_\tau^1)_{ij} = \frac{1}{T}\sum_{t=1}^{\tau} \bar{V}_{ij}^{(t)} + \frac{(\tau - \tau_\star)_+ \mu_{ij}}{T},$$

where $a_+ \overset{\text{def}}{=} \max(a, 0)$. For $t > \tau_\star$, Lemma 5 can be used to write

$$\begin{aligned}
&\left|\mathbb{E}\left[B(X_i^{(t)}, X_j^{(t)}) - \mathbb{E}_{\theta_\star^{(1)}}\left(B(X_i, X_j^{(t)})|X_{-i}^{(t)}\right)\right]\right| \\
&\quad = \left|\mathbb{E}\left[\int_{\mathsf{X}} B(u, X_j^{(t)}) f_{\theta_\star^{(2)}}(u|X_{-i}^{(t)})\mathrm{d}u - \int_{\mathsf{X}} B(u, X_j^{(t)}) f_{\theta_\star^{(1)}}(u|X_{-i}^{(t)})\mathrm{d}u\right]\right| \\
&\quad \leq c_0^2 \sum_{j=1}^{p} |\theta_{\star,ij}^{(2)} - \theta_{\star,ij}^{(1)}| \leq b c_0^2,
\end{aligned}$$

where $b$ is as in (17). Hence

$$|\mu_{ij}| \leq 2 \max_{j \leq i} \left|\mathbb{E}_{\theta_\star^{(2)}}\left[B(X_i^{(t)}, X_j^{(t)}) - \mathbb{E}_{\theta_\star^{(1)}}\left(B(X_i^{(t)}, X_j^{(t)})|X_{-j}^{(t)}\right)\right]\right| \leq 2bc_0^2.$$

Set $\lambda_\tau \overset{\text{def}}{=} (A\sqrt{\tau}/T)$, where

$$A \overset{\text{def}}{=} 32 c_0 \sqrt{\log(dT)}.$$

By a union-bound argument,

$$\begin{aligned}
&\mathbb{P}\left[\max_{\tau \in \mathcal{T}} 2\lambda_\tau^{-1} \|G_\tau^1\|_\infty > 1\right] \\
&\qquad \leq \sum_{\tau \in \mathcal{T}} \sum_{i,j} \mathbb{P}\left[\frac{1}{A\sqrt{\tau}}\left|\sum_{t=1}^{\tau} \bar{V}_{ij}^{(t)}\right| + \frac{2bc_0^2(\tau - \tau_\star)_+}{A\sqrt{\tau}} > \frac{1}{2}\right]. \quad \text{(S36)}
\end{aligned}$$

Since $A = 32c_0\sqrt{\log(dT)}$, for $\tau \in \mathcal{T}$, and using (15) we see that $\max_{\tau \in \mathcal{T}} \frac{2bc_0^2(\tau - \tau_\star)_+}{A\sqrt{\tau}} \leq 1/4$. Hence

$$\begin{aligned}
\mathbb{P}\left[\max_{\tau \in \mathcal{T}} 2\lambda_\tau^{-1} \|G_\tau^1\|_\infty > 1\right] &\leq \sum_{\tau \in \mathcal{T}} \sum_{i,j} \mathbb{P}\left[\left|\sum_{t=1}^{\tau} \bar{V}_{ij}^{(t)}\right| > \frac{A\sqrt{\tau}}{4}\right], \quad \text{(S37)} \\
&\leq 2 \sum_{\tau \in \mathcal{T}} \sum_{i,j} \exp\left(-\frac{A^2}{8^3 c_0^2}\right) \leq \frac{2}{d}.
\end{aligned}$$

where the second inequality uses Hoeffding's inequality. □

REMARK 5. *The* $\log(dT)$ *term that appears in the convergence rate of Theorem 1 follows from the union bound and the exponential bound used in (S36), and (S37) respectively. Alternatively, it is easy to see that one could also write*

$$\mathbb{P}\left[\max_{\tau\in\mathcal{T}} 2\lambda_\tau^{-1}\|G_\tau^1\|_\infty > 1\right] \leq \sum_{i,j}\mathbb{P}\left[\max_{\tau\in\mathcal{T}}\left|\frac{1}{\sqrt{\tau}}\sum_{t=1}^{\tau}\bar{V}_{ij}^{(t)}\right| > \frac{A}{4}\right].$$

*Hence whether one can remote the* $\log(T)$ *term hinges on the existence of an exponential bound for the term* $\max_{\tau\in\mathcal{T}}\left|\tau^{-1/2}\sum_{t=1}^{\tau}\bar{V}_{ij}^{(t)}\right|$. *Unfortunately we are not aware of any such result in the literature. The closest results available deal with the unweighted sums:* $\max_{\tau\in\mathcal{T}}\left|\sum_{t=1}^{\tau}\bar{V}_{ij}^{(t)}\right|$ *(see for instance pinelis (2006) for some of the best bounds available).*

LEMMA 4. *Assume H1 and H2. Let* $\lambda_{1,\tau}$ *and* $\lambda_{2,\tau}$ *as in Equation* (8), *and let the search domain* $\mathcal{T}$ *be such that Equations* (15)-(16) *hold. Take* $c_1 = 64c_0s_1$, $c_2 = 64c_0s_2$ *and*

$$\mathsf{r}_1(x) = \frac{\rho_1 x^2}{2 + 4c_0 s_1^{1/2} x}, \quad \textit{and} \quad \mathsf{r}_2(x) = \frac{\rho_2 x^2}{2 + 4c_0 s_2^{1/2} x}, \ \ x \geq 0.$$

*Then the event* $\bigcap_{\tau\in\mathcal{T}}\left[\mathcal{E}_\tau^1(\lambda_{1,\tau},\mathsf{r}_1,c_1)\cap\mathcal{E}_\tau^2(\lambda_{2,\tau},\mathsf{r}_2,c_2)\right]$ *holds with probability at least* $1-\frac{8}{d}$.

PROOF. We have seen in Lemma 3 that with $\lambda_{1,\tau}$ and $\lambda_{2,\tau}$ as in equation (8), the event $\cap_{\tau\in\mathcal{T}}\left[\{\|G_\tau^1\|_\infty \leq \lambda_{1,\tau}/2\}\cap\{\|G_\tau^1\|_\infty \leq \lambda_{2,\tau}/2\}\right]$ holds with probability at least $1-2/d$. We have

$$\mathcal{L}_1(\tau,\theta) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{\tau}\left[\phi(\theta,X^{(t)}) - \phi(\theta_\star^{(1)},X^{(t)}) - \left\langle\nabla\phi(\theta_\star^{(1)},X^{(t)}),\theta-\theta_\star^{(1)}\right\rangle\right].$$

(S34) then implies that for all $\tau\in\mathcal{T}$, and $\theta-\theta_\star^{(1)}\in\mathbb{C}_1$,

$$\mathcal{L}_1(\tau,\theta) \leq \frac{\tau}{T}\frac{4c_0^2}{2}\|\theta-\theta_\star^{(1)}\|_1^2 \leq \frac{\tau}{T}\frac{64c_0^2s_1}{2}\|\theta-\theta_\star^{(1)}\|_2^2.$$

A similar bound holds for $j=2$. Hence $\cap_{\tau\in\mathcal{T}}\cap_{j=1}^2\left\{\sup_{\theta\neq\theta_\star^{(j)},\,\theta-\theta_\star^{(j)}\in\mathbb{C}_j}\frac{\mathcal{L}_j(\tau,\theta)}{\|\theta-\theta_\star^{(j)}\|_2^2} \leq \frac{\tau}{T}\frac{c_j}{2}\right\}$ holds with probability one.

Using (S35), we have

$$\mathcal{L}_1(\tau,\theta) \geq \frac{\tau}{T}\frac{1}{2+4c_0s_1^{1/2}\|\theta-\theta_\star^{(1)}\|_2} \times \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p}\mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p}B_{kj}(X_j^{(t)},X_k^{(t)})\left(\theta_{kj}-\theta_{\star,kj}^{(1)}\right)|X_{-j}^{(t)}\right). \quad \text{(S38)}$$

We will now show that for all $\tau \in \mathcal{T}$, and all $\theta - \theta_\star^{(1)} \in \mathbb{C}_1$, with probability at least $1 - 2/d$, we have

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p} \mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p} B_{kj}(X_j^{(t)}, X_k^{(t)})\left(\theta_{kj} - \theta_{\star,kj}^{(1)}\right)|X_{-j}^{(t)}\right) \geq \rho_1 \|\theta - \theta_\star^{(1)}\|_2^2.$$

Given (S38), this assertion will implies that $\mathcal{L}_1(\tau,\theta) \geq \frac{\tau}{T}\mathsf{r}_1(\|\theta - \theta_\star^{(1)}\|_2)$ for all $\theta - \theta_\star^{(1)} \in \mathbb{C}_1$ with probability at least $1 - 2/d$, where $\mathsf{r}_1(x) = \rho_1 x^2/(2 + 4c_0 s_1^{1/2} x)$. The lemma will then follow easily.

For $\Delta \in \mathcal{M}_p$, we define

$$\mathcal{V}^1(\tau,\Delta) \stackrel{\text{def}}{=} \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p} \mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p} B_{kj}(X_j^{(t)}, X_k^{(t)})\Delta_{kj}|X_{-j}^{(t)}\right),$$

and

$$W_{jkk'}^{(t)} \stackrel{\text{def}}{=} \mathsf{Cov}_{\theta_\star^{(1)}}\left(B(X_j^{(t)}, X_k^{(t)}), B(X_j^{(t)}, X_{k'}^{(t)})|X_{-j}^{(t)}\right) - \mathbb{E}\left[\mathsf{Cov}_{\theta_\star^{(1)}}\left(B(X_j^{(t)}, X_k^{(t)}), B(X_j^{(t)}, X_{k'}^{(t)})|X_{-j}^{(t)}\right)\right].$$

Then for $\Delta \in \mathbb{C}_1 \setminus \{0\}$,

$$\begin{aligned}\mathcal{V}^1(\tau,\Delta) &= \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}\mathbb{E}\left[\mathsf{Cov}_{\theta_\star^{(1)}}\left(B(X_j^{(t)}, X_k^{(t)}), B(X_j^{(t)}, X_{k'}^{(t)})|X_{-j}^{(t)}\right)\right] \\ &\quad + \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}W_{jkk'}^{(t)} \qquad \text{(S39)}\end{aligned}$$

Using H1, we deduce that

$$\begin{aligned}\mathcal{V}^1(\tau,\Delta) \geq 2\rho_1\|\Delta\|_2^2 &+ \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}W_{jkk'}^{(t)} \\ &+ \frac{(\tau-\tau_\star)_+}{\tau}\sum_{j=1}^{p}\mathbb{E}_{\theta_\star^{(2)}}\left[\mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p}\Delta_{jk}B_{ik}(X_j,X_k)|X_{-j}\right)\right] \\ &- \frac{(\tau-\tau_\star)_+}{\tau}\sum_{j=1}^{p}\mathbb{E}_{\theta_\star^{(1)}}\left[\mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p}\Delta_{jk}B_{ik}(X_j,X_k)|X_{-j}\right)\right]. \qquad \text{(S40)}\end{aligned}$$

By the comparison Lemma 5

$$\left|\mathbb{E}_{\theta_\star^{(2)}}\left[\mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p}\Delta_{jk}B_{ik}(X_j,X_k)|X_{-j}\right)\right]-\mathbb{E}_{\theta_\star^{(1)}}\left[\mathsf{Var}_{\theta_\star^{(1)}}\left(\sum_{k=1}^{p}\Delta_{jk}B_{ik}(X_j,X_k)|X_{-j}\right)\right]\right|$$
$$\leq c_0^3\left(\sum_{k=1}^{p}|\Delta_{jk}|\right)^2\sum_{k=1}^{p}|\theta^{(1)}_{\star jk}-\theta^{(2)}_{\star jk}|\leq c_0^3 b\left(\sum_{k=1}^{p}|\Delta_{jk}|\right)^2,$$

which implies that

$$\mathcal{V}^1(\tau,\Delta)\geq\left(2\rho_1-\frac{64}{\tau}(\tau-\tau_\star)_+s_1c_0^3b\right)\|\Delta\|_2^2+\frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}W^{(t)}_{jkk'}.$$

Given that on $\mathcal{T}_+$, $128(\tau-\tau_\star)s_1c_0^3b\leq\rho_1\tau$, it follows that for all $\tau\in\mathcal{T}$,

$$\mathcal{V}^1(\tau,\Delta)\geq\frac{3}{2}\rho_1\|\Delta\|_2^2+\frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}W^{(t)}_{jkk'}\tag{S41}$$

Set $Z^\tau_{jkk'}\stackrel{\text{def}}{=}\frac{1}{\tau}\sum_{t=1}^{\tau}W^{(t)}_{jkk'}$. We conclude from equation (S41) that if for some $\Delta\in\mathbb{C}_1\setminus\{0\}$, and for some $\tau\in\mathcal{T}$,

$$\mathcal{V}^1(\tau,\Delta)\leq\rho_1\|\Delta\|_2^2\tag{S42}$$

then

$$\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}Z^{(\tau)}_{jkk'}\leq-\frac{\rho_1}{2}\|\Delta\|_2^2.$$

But on the other hand, using the fact that $\Delta\in\mathbb{C}_1$,

$$\begin{aligned}\sum_{j=1}^{p}\sum_{k,k'=1}^{p}\Delta_{jk}\Delta_{jk'}Z^{(\tau)}_{jkk'}&\geq-\left(\sup_{j,k,k'}|Z^{(\tau)}_{jkk'}|\right)\left(\sum_{i=1}^{p}\sum_{k=1}^{p}|\Delta_{ik}|\right)^2\\&\geq-\left(\sup_{j,k,k'}|Z^{(\tau)}_{jkk'}|\right)4\|\Delta\|_1^2\\&\geq-64s_1\left(\sup_{j,k,k'}|Z^{(\tau)}_{jkk'}|\right)\|\Delta\|_2^2.\end{aligned}$$

Therefore if there exists a non-zero $\Delta\in\mathbb{C}_1$ and $\tau\in\mathcal{T}$ such that equation (S42) holds then $\left(\sup_{j,k,k'}|Z^{(\tau)}_{jkk'}|\right)\geq(\rho_1/s_1)(1/128)$. But by Hoeffding's inequality and a

union-sum bound,

$$\mathbb{P}\left[\sup_{j,k,k'} |Z^{(\tau)}_{jkk'}| \geq \frac{\rho_1}{128 s_1}\right] \leq 2\exp\left(3\log p - \frac{\tau\rho_1^2}{2^9 c_0^2 s_1^2}\right) \leq \frac{2}{p},$$

since for $\tau \in \mathcal{T}$, $\tau \geq 2^{11} c_0^2 s_1^2 \rho_1^{-2} \log p$. □

LEMMA 5. *Let $(\mathsf{Y}, \mathcal{A}, \nu)$ be a measure space where $\nu$ is a finite measure. Let $g_1, g_2, f_1, f_2 : \mathsf{Y} \to \mathbb{R}$ be bounded measurable functions. Set $Z_{g_i} \stackrel{\text{def}}{=} \int_{\mathsf{Y}} e^{g_i(y)}\nu(dy)$, $i \in \{1, 2\}$. Then*

$$\left|\frac{1}{Z_{g_1}}\int f_1(y)e^{g_1(y)}\nu(dy) - \frac{1}{Z_{g_2}}\int f_2(y)e^{g_2(y)}\nu(dy)\right| \leq \|f_2 - f_1\|_\infty + \frac{1}{2}\mathsf{osc}(g_2 - g_1)\left(\mathsf{osc}(f_1) + \mathsf{osc}(f_2)\right),$$

*where $\|f\|_\infty = \sup_{x\in\mathsf{Y}} |f(x)|$, and $\mathsf{osc}(f) \stackrel{\text{def}}{=} \sup_{x,y\in\mathsf{Y}} |f(x) - f(y)|$ is the oscillation of $f$.*

PROOF. The proof follows from Atchadé (2014) Lemma 3.4.

## S9. Different Methods of Missing Data Imputation for the Real Data Application

In the main paper we replaced the missing votes by the value (yes/no) of that member's party majority position on that particular vote. Here we employed two other missing data imputation techniques viz. (i) replacing all missing values by the value (yes/no) representing the winning majority on that bill and (ii) replacing the missing value of a Senator by the value that the majority of the opposite party voted on that particular bill. The estimated change-point obtained following these two imputation methods are not much different . The imputation technique (i) results in a estimated change-point at January 19, 1995 and the technique (ii) yields estimated change-point at January 17, 1995 respectively. The change-point estimate we obtained in the main paper was January 17, 1995. Clearly there is not much difference between the different imputation techniques and Fig. S7 also conveys the same message.


Fig. S7: Estimated Change-points via imputation technique (i) and (ii) respectively