Consistent Multiple Changepoint Estimation with Fused Gaussian Graphical Models

A. Gibberd S. Roy

Received: date / Revised: date

Abstract

We consider the consistency properties of a regularised estimator for the simultaneous identification of both changepoints and graphical dependency structure in multivariate time-series. Traditionally, estimation of Gaussian Graphical Models (GGM) is performed in an i.i.d setting. More recently, such models have been extended to allow for changes in the distribution, but primarily where changepoints are known a-priori. In this work, we study the Group-Fused Graphical Lasso (GFGL) which penalises partial-correlations with an L1 penalty while simultaneously inducing block-wise smoothness over time to detect multiple changepoints. We present a proof of consistency for the estimator, both in terms of changepoints, and the structure of the graphical models in each segment. We contrast our results, which are based on a global, i.e. graph wide likelihood, with those previously obtained for performing dynamic graph estimation at a node-wise (or neighbourhood) level.

Keywords

changepoint; regularisation; graphical model; asymptotics

1 Introduction

Many modern day datasets exhibit multivariate dependance structure that can be modelled using networks or graphs. For example, in social sciences, biomedical studies, financial applications etc. the association of datasets with latent network structures are ubiquitous. Many of these datasets are time-varying in nature and that motivates the modelling of dynamic networks. A network is usually characterised by a graph G with vertex set V (the collection of nodes) and edge set E (the collection of edges). We denote $G = (V, E)$. For example, in a biological application nodes may denote a set of genes and the edges may be the interactions among the genes. Alternatively, in neuroscience, the nodes may represent observed processes in different regions of the brain, and the edges represent functional connectivity or connectome. In both situations,

we may observe activity at nodes over a period of time— the challenge is to infer the dependency network and how this changes over time.

Although the topic of change-point estimation is well represented in statistics through methods such as binary segmentation [Fryzlewicz, 2014], dynamic programming [Killick et al., 2012], or more classical methods such as Bai [1997], Hinkley [1970], Raimondo [1998] and references therein; its application in the context of graphical models, i.e. how the conditional dependency structure between a set of variables changes, is relatively unexplored. We here consider a particular type of dynamic network model where the underlying conditional dependency structure, encoded as a graph, evolves in a piecewise fashion. The task is to estimate multiple change-points where the network structure changes, as well as the structures themselves. To this end, we formulate the problem as a joint optimization task whereby change-point estimation and structure recovery can be performed simultaneously. The particular class of networks we aim to estimate are encoded via a multivariate Gaussian that has a piecewise constant precision matrix and thus graph structure over certain blocks of time[1]. Specifically, we assume observations $X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ are drawn from the following model:

$$X^{(t)} \sim \mathcal{N}(0, \Sigma_0^{(k)}) \quad , \; t \in [\tau_{k-1}, \tau_k) \; , \tag{1}$$

where $t = 1, \ldots, T$ indexes the time of the observed data-point, and $k = 1, \ldots, B := K + 1$ indexes blocks $[\tau_{k-1}, \tau_k) := \{\tau_{k-1}, \ldots, \tau_k - 1\}$, separated by changepoints $\{\tau_k\}_{k=1}^K$, at which points the covariance matrix $\Sigma_0^{(k)}$ changes. The challenge is to assess, how well, or indeed if, we can recover both the changepoint positions τ_k and the correct precision matrices $\Theta_0^{(k)} := (\Sigma_0^{(k)})^{-1}$. In this paper we focus on the task of recovering $\{\Theta_0^{(k)}, \tau_k\}_{k=1}^B$ in the case where K is known in advance, that is, rather than focus on estimating the number of changes, we consider where and in what form these changes appear.

For dependency graph identification in the static i.i.d. setting there are two principal estimation approaches. Firstly, as suggested by Meinshausen and Bühlmann [2006], one may adopt a neighbourhood or *local* selection approach where edges are estimated at a node-wise level, an estimate for the network is then constructed by iterating across nodes. Alternatively, one may consider joint estimation of the edge structure across all nodes in a *global* fashion. In the i.i.d. setting a popular method to achieved this is via the Graphical lasso [Banerjee and Ghaoui, 2008], or explicitly constrained precision matrix estimation schemes such as CLIME [Cai et al., 2011].

In the non-stationary setting, one could consider extending regression methods, for instance utilising the methods of Lee et al. [2016], Leonardi and Bühlmann [2016] to estimate a graph where each node may exhibit multiple change points. The work of Roy et al. [2016] considers a joint estimation approach for networks in the presence of a single changepoint, while Kolar and Xing [2012] consider using the fused lasso [Harchaoui and Lévy-Leduc, 2010] to estimate multiple changepoints at a node-wise level. One may consider *fused* smoothing that re-

[1]For a reference on static graphical models the reader is directed to Lauritzen [1996]

stricts changes in the graph structure, either at an individual edge level via an ℓ_1 penalty [Gibberd and Nelson, 2014, Monti et al., 2014], or across multiple edges via a group-fused $\ell_{2,1}$ penalty [Gibberd and Nelson, 2017]. The work of Angelosante and Giannakis [2011] proposed to combine the graphical lasso with dynamic programming to estimate changepoints and graph structures.

In this paper we operate in the group-fused setting, and we provide theoretical analysis for the Group-Fused Graphical lasso (GFGL) estimator first proposed in Gibberd and Nelson [2017] and similarly in Hallac et al. [2017]. In these works, it was demonstrated empirically that GFGL can detect both changepoints and graphical structure in relatively high-dimensional settings. However, until now, the theoretical consistency properties of the estimator have remained unexplored. In this paper, we derive rates for the consistent recovery of both changepoints and model structure via upper bounds on the errors: $\max_k |\hat{\tau}^{(k)} - \tau_0^{(k)}|$ and $\|\hat{\Theta}^{(k)} - \Theta_0^{(k)}\|_\infty$ under sampling from the model in Eq. 1.

Definition 1 (Group-Fused Graphical Lasso)**.** *Let $\hat{S}^{(t)} := X^{(t)}(X^{(t)})^\top$ be the local empirical covariance estimator. The GFGL estimator is defined as the M-estimator*

$$\{\hat{\Theta}^{(t)}\}_{t=1}^T = \underset{\{U^{(t)} \succeq 0\}_{t=1}^T}{\arg\min} \Bigg[\underbrace{\sum_{t=1}^T \Big\{ -\log\det(U^{(t)}) + \mathrm{trace}(\hat{S}^{(t)} U^{(t)}) \Big\}}_{l_T(U,S)} + \underbrace{\lambda_1 \sum_{t=1}^T \|U_{\setminus ii}^{(t)}\|_1 + \lambda_2 \sum_{t=2}^T \|U^{(t)} - U^{(t-1)}\|_F}_{r_T(U)} \Bigg], \quad (2)$$

where $U_{\setminus ii}^{(t)}$ deotes the matrix $U^{(t)}$ with diagonal entries set to zero. Once the precision matrices have been estimated, changepoints are defined as the time-points $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\} := \{t \mid \hat{\Theta}^{(t)} - \hat{\Theta}^{(t-1)} \neq 0\}$. While changepoints in the traditional sense are defined above, it is convenient later on to consider the block separators $\hat{\mathcal{T}} := \{1\} \cup \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\} \cup \{T+1\}$, the added entries are denoted τ_0 and $\tau_{\hat{K}+1}$.

Remark 1. *Time and Block Notation*

Throughout the paper we will deal with parameters indexed by both individual time-steps, and as part of constant blocks, respectively, these are denoted using t and k. For instance, $\Theta_0^{(t)}$ denotes the tth time-step of the true model structure (in terms of precision matrix), whereas, $\Theta_0^{(k)}$ denotes the precision matrix for the kth block, i.e. inline with Eq. 1. Note: t is always reserved for time indexing, whereas we use k or l to reference blocks. Estimators are denoted with a hat notation, i.e. $\hat{\Theta}^{(t)}$ or $\hat{S}^{(t)}$, while ground-truth objects are denoted with $_0$, i.e. $\Theta_0^{(k)}$ refers to the true precision matrix in the kth block.

2 Theoretical Analysis of the GFGL Estimator

To assess the statistical properties of the GFGL estimator, we first need to derive a set of conditions which all minimisers of the cost function (2) obey. We connect this condition to the sampling of the model under (1) by the quantity $W^{(t)} := \hat{S}^{(t)} - \Sigma_0^{(t)}$ which represents the difference between the ground-truth covariance and the empirical covariance matrix $X^{(t)}(X^{(t)})^\top$. Since we are dealing with a fused regulariser, it is convenient to introduce a matrix $\Gamma^{(t)}$ corresponding to the differences in precision matrices. For the first step let $\Lambda^{(1)} = \Theta^{(1)}$, then for $t = 2, \ldots, T$, let $\Lambda^{(t)} = \Theta^{(t)} - \Theta^{(t-1)}$. The sub-gradients for the non-smooth portion of the cost function are denoted respectively as $\hat{R}_1^{(t)}, \hat{R}_2^{(t)} \in \mathbb{R}^{p \times p}$, for the ℓ_1 and the group-smoothing penalty. In full, these can be expressed as

$$\hat{R}_{1;ij}^{(t)} = \begin{cases} \text{sign}(\sum_{s \leq t} \hat{\Lambda}_{ij}^{(s)}) & \text{if } \sum_{s \leq t} \hat{\Lambda}_{ij}^{(s)} \neq 0 \\ [-1, 1] & \text{otherwise} \end{cases}$$

$$\hat{R}_2^{(t)} = \begin{cases} \frac{\hat{\Lambda}^{(t)}}{\|\hat{\Lambda}^{(t)}\|_F} & \text{if } \hat{\Lambda}^{(t)} \neq 0 \\ \mathbb{B}_F(0, 1) & \text{otherwise} \end{cases} ,$$

where $\mathbb{B}_F(0, 1)$ is the Frobenius unit ball.

Lemma 1 (GFGL Optimality Conditions)**.** *The minimiser $\{\hat{\Theta}^{(t)}\}_{t=1}^T$ of the GFGL objective satisfies the following*

$$\sum_{t=l}^T \left\{ (\Theta_0^{(t)})^{-1} - (\hat{\Theta}^{(t)})^{-1} \right\} + \sum_{t=l}^T W^{(t)} + \lambda_1 \sum_{t=l}^T \hat{R}_1^{(t)} + \lambda_2 \hat{R}_2^{(l)} = 0 ,$$

for all $l \in [T]$ and $\hat{R}_2^{(1)} = \hat{R}_2^{(T)} = 0$.

2.1 Consistent Changepoint Estimation

We here present a result for changepoint consistency with the GFGL estimator where $T \to \infty$, and the dimensionality p is fixed. Let $\{\delta_T\}_{T \geq 1}$ be a non-increasing positive sequence that converges to zero as $T \to \infty$. This quantity should converge at a rate which ensures an increasing absolute quantity $T\delta_T \to \infty$ as $T \to \infty$. The target of our results is to bound the maximum error to an ever decreasing proportion of the data, i.e. $\max_k |\hat{\tau}_k - \tau_k|/T \leq \delta_T$. To establish a bound, we consider the setting where the minimum true distance between changepoints $d_{\min} := \min_{k \in [B]} |\tau_k - \tau_{k-1}|$ increases with T, for simplicity let us assume this is bounded as a proportion $\gamma_{\min} < d_{\min}/T$. Furthermore, let us consider that the minimum jump size is denoted by:

$$\eta_{\min} := \min_{k \in [B]} \|\Sigma_0^{(k)} - \Sigma_0^{(k-1)}\|_F ,$$

and that the maximum jump size is finite in the Frobenius norm. The term $\eta_{\min}$ is important as in some sense it defines the strength of the jump signal

in our true sequence of covariance matrices. We can either consider $\eta_{\min}$ fixed as a function of T and demonstrate that we recover the jump positions with increasing precision, and/or we consider $\eta_{\min}$ to be a decreasing sequence in T and that we can recover (asymptotically in T) ever decreasing jump sizes.

The first result we present below requires that the regularisers λ_1, λ_2 are set in a very specific way, to achieve the correct number of changepoints. After discussing the result based on this somewhat restrictive assumption we will then present an alternative result which is more flexible in the specification of the regularisers.

Assumption 1. *Appropriate Regularisation*

For sequences of regularisers $\{\lambda_{1:T}\}$, $\{\lambda_{2:T}\}$ (and $\{\eta_{\min:T}\}$ if we consider this as a decreasing sequence) $\beta_1 := (\eta_{\min}\gamma_{\min}T)\lambda_2^{-1} > 2^5$, $\beta_2 := \eta_{\min}\lambda_1^{-1}(p(p-1))^{-1/2} > 2^3$, and $\beta_3 := (\eta_{\min}T\delta_T)\lambda_2^{-1} > 3$.

Theorem 1 (Changepoint Consistency, $\hat{K} = K$)**.** *Assume for all $T \geq T_0$, we have sequences which meet Assumption 1 such that GFGL problem (2) results in $|\hat{K}| = K$ changepoints, then we have*

$$P(\max_{k\in[K]} |\tau_k - \hat{\tau}_k| \geq T\delta_T) \leq f_\tau(T) := C_{K,p}\exp\{-T\delta_T\eta_{\min}^2/p^2c_3\}\,, \qquad (3)$$

where $C_{K,p} = 4p^2K(K^2 2^{K+1} + 4)$. The probabilistic bound holds for $c_3 = 5^4 2^7 \max_k \|\Sigma_0^{(k)}\|_\infty$, and $\eta_{\min} \leq 2^3 5^2 p \max_k \|\Sigma_0^{(k)}\|_\infty$.

The proof of the above follows a similar line of argument as used in Harchaoui and Lévy-Leduc [2010], Kolar and Xing [2012]. However, several modifications are required in order to allow analysis with the Gaussian likelihood and group-fused regulariser, we also utilise a different concentration bound. Full details can be found in Appendix A.

The results demonstrate that asymptotically changepoint error can be constrained to a decreasing fraction $\delta_T \to 0$ of the time-series as $T \to \infty$. Indeed, this occurs with increasing probability when estimation is performed with an increasing number of data-points. Unlike in high-dimensional settings [e.g. Negahban et al., 2012, Ravikumar et al., 2011], the regularisation parameters have less strict requirements on their form and the bound in Theorem 1 holds for any $T \geq T_0$ where T_0 is dictated by the choice of tuning parameter sequences in Assumption 1.

Remark 2. *Choosing the regularisers to scale as $\lambda_1 \asymp \lambda_2 = \mathcal{O}[\{\log(T)/T\}^{1/2}]$ enables convergence in probability with*

$$\delta_T = \log(T)^\alpha/T \quad ; \quad \eta_{\min} = \Omega\{(\log T)^{(1-\alpha)/2}\}\,, \qquad (4)$$

where $\alpha \geq 1$. Under such regularisation, the conditions in the theorem (specifically Assumption 1) are met (under the assumption $\hat{K} = K$) and the exponential bound of (3) decreases in order $f_\tau(T) \propto T^{-1}$. Alternatively, one may consider the polynomial quantities $\eta_{\min} = \Omega(T^{-b})$ and $\delta_T = T^{-a}$, where $a, b > 0$ and

$a + 2b = c < 1$. *In this case* $T\delta_T > 0$ *still increases with* T*, although the exponential bound is dependent on* c *of the form* $f_\tau(T) \propto \exp\{-T^{1-c}\}$*. Unlike in (4), when using the polynomial scaling there is a clear trade-off between the minimum jump size* $\eta_{\min}$*, and the amount of data* $T\delta_T$ *required to gain a certain level of changepoint consistency. For example, considering the case where* $b = 0$*, and thus* $\eta_{\min}$ *is a constant as* $T \to \infty$*, enables for a fixed value of* c *a larger value of* $a = c$*. In such cases changepoints may be recovered with greater accuracy.*

In Assumption 1 we made a strong assumption on the number of changepoints, in that we had $\hat{K} = K$ and that we somehow had access to a sequence of regularisers λ_1, λ_2 which attained such a condition. In practice, we will not often be able to meet the condition $\hat{K} = K$ as we don't know a-priori what the true number of changepoints K is. We note that this is a general challenge in changepoint estimation methods, as these are typically used in an exploratory data-analysis setting Harchaoui and Lévy-Leduc [2010], Killick et al. [2012]. What is perhaps more reasonable to assume is that we recover some $\hat{K}$ which upper bounds the truth. In the fused-lasso inspired changepoint estimators, such as GFGL studied here and that of Harchaoui and Lévy-Leduc [2010] it is typically seen experimentally that the regularisation imposes a bias on the estimated parameters. Particularly, unlike when using ℓ_0 penalised schemes such as AIC/BIC, for instance those which are typically used with dynamic programming schemes [Angelosante and Giannakis, 2011, Killick et al., 2012], the ℓ_1 or group smoothing can shrink the size of parameter changes resulting in less defined "jumps" and over-estimating the number of changepoints. For examples of this, the reader is referred to the synthetic experiments in Harchaoui and Lévy-Leduc [2010], Gibberd and Nelson [2017]. In the former paper, the authors suggest a second stage of estimation which further prunes the estimates from the fused-lasso smoother using a dynamic programming scheme. Potentially, this hybrid estimation scheme can also be adapted to GFGL, however, we leave this as further work. In the following result, we consider relaxing Assumption 1 and bounding the distance between any point in the estimated changepoint set $\hat{\mathcal{T}}_{\hat{K}}$, even in the case where $\hat{K} \geq K$.

Proposition 1 (Changepoint Error, $\hat{K} \geq K$)**.** *Consider the maximum distance of any estimated changepoint from its closest true changepoint as measured via*

$$h(\hat{\mathcal{T}}_{\hat{K}} \| \mathcal{T}_K) = \sup_{\tau \in \mathcal{T}_K} \inf_{\hat{\tau} \in \hat{\mathcal{T}}_{\hat{K}}} |\hat{\tau} - \tau| \, .$$

then under Assumption 1 and that $K \leq \hat{K} \leq K_{\max}$ *we have*

$$P[h(\hat{\mathcal{T}}_{\hat{K}} \| \mathcal{T}_K) \leq T\delta_T] \to 1 \, , \text{as } T \to \infty \, .$$

The proof of the above can be demonstrated along similar lines to that of Theorem 1 Harchaoui and Lévy-Leduc [2010]. The result is a direct extension of bounds used in proving Thm. 1.

2.2 Consistent Graph Recovery

One of the key properties of GFGL and similar fused estimation procedures is that they simultaneously estimate both the changepoint and model, i.e. conditional dependency structure. In this section we will turn our eye to the estimation of model structure in the form of the precision matrices between changepoints. In particular, we consider that a set of $\hat{K} = K$ changepoints have been identified as per Assumption 1 and Theorem 1. We here assume that such a bound $\max_{k\in[K]} |\tau_k - \hat{\tau}_k| \leq T\delta_T$ holds, and develop theory relating to the recovery of model structure and parameters in the relevant segments.

A key advantage of splitting the changepoint and model-estimation consistency arguments as we do here, is that we can consider a simplified model structure such that the GFGL estimator may be parameterised in terms of a $B = K + 1$ block-diagonal matrix $\Theta_{0;B} \in \tilde{\mathbb{R}}^{Bp\times Bp}$. Conditional on segmentation, we do not need to deal with the fact that the model may be arbitrarily mis-specified, as this is bounded by Theorem 1. As such, in this section the dimensionality of the model space is fixed with respect to an increasing number of time-points. The following results demonstrate, that as expected, gathering increasing amounts of data relating to a fixed number of blocks allows us to identify the model with increasing precision.

Let us define a set of pairs $\mathcal{M}_k$ which indicate the support set of the true model in block k and its complement $\mathcal{M}_k^\perp$ as $\mathcal{M}_k = \{(i,j) \mid \Theta^{(k)}_{0;ij} \neq 0\}$ and $\mathcal{M}_k^\perp = \{(i,j) \mid \Theta^{(k)}_{0;ij} = 0\}$ respectively. The recovery of the precision matrix sparsity pattern in true block l from estimated block k can be monitored by the sign-consistency event defined:

$$E_{\mathcal{M}}(\hat{\Theta}^{(k)}; \Theta_0^{(l)}) := \left\{ \text{sign}(\hat{\Theta}^{(k)}_{ij}) = \text{sign}(\Theta^{(l)}_{0;ij}) \; \forall i,j \in \mathcal{M}_l \right\} .$$

In order to derive bounds on model recovery, one must make assumptions on the true structure of Θ_0. Whilst the GFGL loss function is strictly convex over the space of positive definite matrices (Appendix D), one also needs to take into account more specific constraints referred to as *incoherence* or *irrepresentability* conditions. In the setting of graphical structure learning, these conditions act to limit correlation between edges and restrict the second order curvature of the loss function. In the case where we analyse GFGL under Gaussian sampling the Hessian $\Gamma_0 \equiv \nabla^2_\Theta L(\Theta)|_{\Theta_0}$ relates to the Fisher information matrix such that $\Gamma_{0;(j,k)(l,m)} = \text{Cov}(X_j X_k, X_l X_m)$. Written in this form we can understand the Fisher matrix as relating to the covariance between *edge variables* defined as $Z^{(t)}_{ij} = X^{(t)}_i X^{(t)}_j - \mathbb{E}[X^{(t)}_i X^{(t)}_j]$ for $i,j \in \{1,\ldots,p\}$.

Assumption 2. *Incoherence Condition*

Let $\mathcal{M}$ denote the set of components relating to true edges in the graph and $\mathcal{M}^\perp$ (for block k) its complement. For example, $\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}}$ refers to the sub matrix of the Fisher matrix relating to edges in the true graph. Assume that for each $k = 1,\ldots,B$ there exists some $\alpha_k \in (0,1]$ such that

$$\max_{e\in\mathcal{M}^\perp} \|\Gamma^{(k)}_{0;e\mathcal{M}}(\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}})^{-1}\|_1 \leq (1-\alpha_k) .$$

Notationally, we will use $U_{\mathcal{M}}^{(k)}$ to denote components of a matrix that are in the true support. In the multivariate Gaussian case we have

$$\max_{e \in \mathcal{M}^{\perp}} \|\mathbb{E}[Z_e^{(t)} Z_{\mathcal{M}}^{(t)\top}]\mathbb{E}[Z_{\mathcal{M}}^{(t)} Z_{\mathcal{M}}^{(t)\top}]^{-1}\|_1 \leq (1-\alpha_k) \ ,$$

for every $t \in \{\tau_k, \ldots, \tau_k + 1\}$ for each k where we denote and track $\alpha = \min_k\{\alpha_k\}$. One can interpret the incoherence condition as a statement on the correlation between edge variables which are outside the model subspace Z_{ij} such that $(i,j) \notin \mathcal{M}$, and those contained in the true model $(i,j) \in \mathcal{M}$. In practice, this sets bounds on the types of graph and associated covariance structures which estimators such as graphical lasso can recover (see the discussion in Sec. 3.1.1 of Ravikumar et al. 2011, and Meinshausen 2008). The model selection proof presented here can be seen as an extension of Ravikumar et al. [2011] to non-stationary settings. Similarly to their original analysis we will track the maximal row-wise sum via the upper bound $K_{\Sigma_0} := \max_k |||\Sigma_0^{(k)}|||_\infty$ and $K_{\Gamma_0} := \max_k |||\Gamma_0^{(k)}|||_\infty$.

When using GFGL there will generally be an error associated with the identification of changepoints and as such the estimated and ground-truth blocks do not directly align. With this in mind, the model consistency proof we present does not necessarily compare the kth estimated block, to the kth ground-truth block. Instead, the result we present is constructed such that the structure in an estimated block $k \in [\hat{B}]$ is compared to the ground-truth structure in block l such that the blocks k and l maximally overlap with respect to time. Notationally, let $\hat{n}_k = \hat{\tau}_k - \hat{\tau}_{k-1}$ and $\hat{n}_{lk} = |\{\hat{\tau}_{k-1}, \ldots, \hat{\tau}_k\} \cap \{\tau_{l-1}, \ldots, \tau_l\}|$, the maximally overlapping block is then defined as $k_{\max} = \arg\max_l\{\hat{n}_{lk}\}$.

Theorem 2 (Bounds on Estimation Error)**.** *Consider the GFGL estimator with Assumption 2, and in the case where changepoint error is bounded according to the event* $E_\tau := \{\max_{k \in [K]} |\tau_k - \hat{\tau}_k| < T\delta_T\}$*, this event occurs in probability* $1 - f_\tau(T)$ *under Theorem 1. Assume* $\lambda_1 = 16\alpha^{-1}\epsilon$*,* $\lambda_2 = \rho\lambda_1$ *for some finite* $\rho > 0$ *and any*

$$\epsilon \in \left(\sqrt{\frac{c_4 \log(4p^2)}{(\gamma_{\min} - 2\delta_T)T}}, \min\left\{ \frac{1}{2^3 5 \max_k \|\Sigma_0^{(k)}\|_\infty}, \frac{1}{6dz_T^2 \max\{K_{\Sigma_0}K_{\Gamma_0}, K_{\Sigma_0}^3 K_{\Gamma_0}^2\}} \right\} \right) \ , \quad (5)$$

where $c_4 = 2^7 5^2 \max_k \|\Sigma_0^{(k)}\|_\infty$*, and* $z_T := 1 + 2^4\alpha^{-1}(1 + 2/(\gamma_{\min} - 2\delta_T)T)$*, then we have*

$$\|\hat{\Theta}^{(k)} - \Theta_0^{(k_{\max})}\|_\infty \leq 2K_{\Gamma_0}\epsilon z_T \ , \quad (6)$$

in probability greater than $1 - f_\tau(T) - f_V(T) \to 1$ *as* $T \to \infty$*, where*

$$f_V(T) := 4p^2 \exp\{-c_4^{-1}\epsilon^2(\gamma_{\min} - 2\delta_T)T\} \ .$$

Corollary 1 (Model-selection consistency)**.** *In addition to the assumptions in Theorem 2. Let* $\theta^{(k)}_{\min} := \min_{ij} |\Theta^{(k)}_{0;ij}|$ *for all* $(i,j) \in \mathcal{M}_k$ *and for each* $k = 1, \ldots, B$*. Let* $v_\theta = 2K_{\Sigma_0} z_{\hat{n}_k} \theta^{-1}_{\min}$*, then if* ϵ *satisfies Eq. 5, and* $\epsilon \leq 1/2K_{\Sigma_0} z_{\hat{n}_k} \theta^{-1}_{\min}$ *, then GFGL attains sign-consistency* $P\{E_{\mathcal{M}}(\hat{\Theta}^{(k)}; \Theta^{(k_{\max})}_0)\} \geq 1 - f_\tau(T) - f_V(T)$ *with probability tending to one as* $T \to \infty$*.*

Theorem 2 is obtained utilising a primal-dual-witness approach conditional on the event E_τ. This follows a similar argument to that used in Ravikumar et al. [2011], but requires modifications due to the smoothing regulariser in GFGL and posssible (but limited) mis-specification related to changepoints. Corollary 1 follows from Theorem 2 subject to the condition that the true entries in the precision matrix are sufficiently large, i.e. bounded away from zero. Full details of the proofs are found in Appendix B.

The above bounds suggest that indeed, if regularisation is appropriately set one can not only recover changepoint structure, but also obtain a consistent estimate of the precision matrices using GFGL. However, there are several important insights we can take from the results. Firstly, we clearly see the effect of the smoothing regulariser in (6) where a larger ρ will result in a larger upper bound for the error, even though asymptotically this bias diminishes. In the analagous results from the i.i.d. graphical lasso case [Ravikumar et al., 2011], the bound on Eq. 6 is of a form $2K_{\Gamma_0}(1 + 8\alpha^{-1})\epsilon$. In fact, our results suggest that the additional error in the precision matrix is a function of the ratio $\lambda_2/\lambda_1\hat{n}_k$, if we now let $\rho_T = \lambda_2/\lambda_1$ vary as a function of T, but so it doesn't grow faster than $\hat{n}_k \simeq \mathcal{O}(T)$, then estimation consistency can still be achieved. For example, if we set $\lambda_1 = \mathcal{O}(\{\log(p)/T\}^{1/2})$ then noting $\hat{n}_k = \mathcal{O}(T)$ (under changepoint consistency), gives λ_2 the flexibility to grow with increasing T, for instance at a rate $\lambda_2 = \mathcal{O}(\{T \log p\}^{1/2})$. We note that under such scaling it is possible to satisfy the conditions of Assumption 1, thus both changepoint and structure estimation consistency can be achieved.

3 Discussion

The grouped nature of GFGL assumes that the graph structure underlying a process changes in some sense systematically, where changepoints are considered at the full precision matrix scale as opposed to seperately for each node $i = 1, \ldots, p$. As a consequence of this changepoint definition, we may expect that the minimum jump size $\eta_{\min}$ required in our results (Theorem 1) are greater than that utilised in neighbourhood selection case [Meinshausen and Bühlmann, 2006, Kolar and Xing, 2012]. To compare, lets consider defining the analogous quantity $\eta^{(i)}_{\min} := \min_k \|\Sigma^{(k+1)}_{0;i,\cdot} - \Sigma^{(k)}_{0;i,\cdot}\|_2$, at the neighbourhood of each node $i = 1, \ldots, p$. The jumps as measured through the group-norm can now be related to those measured in a Frobenius sense, such that $\eta_{\min} \leq \sum_i \eta^{(\mathrm{i})}_{\min} \leq p^{1/2}\eta_{\min}$. Thus, even though the minimum jump size in the GFGL case is greater, i.e. $\eta_{\min} > \eta^{(i)}_{\min}$, it is not proportionally greater when one considers summing over nodes. In our analysis it should be noted that consistent recovery

of changepoints requires a tradeoff between the minimum jump-size $\eta_{\min}$ and the amount of data T. For example, a smaller minimum jump-size will generally require more data; as expected it is harder to detect small jumps. The relation $\eta_{\min} \leq \sum_i \eta^{(i)}_{\min}$ suggests that the minimum jump-size at a graph-wide (precision matrix wide) level is proportionally smaller when measured in the Frobenius norm, than at a node-wise level. As a result, for equivalent scaling of $\eta_{\min}$ and $\eta^{(\mathrm{i})}_{\min}$ the graph-wide GFGL method will be able to detect smaller (graph-wide) jumps with an equivalent level of data. Conversely, if the jumps one is interested in occur at the neighbourhood level the neighbourhood based method would be more appropriate.

Neighbourhood selection is the result of a set of p group-fused optimisation problems, and changepoints are selected at the node level. However, in many situations it is not trivial how we combine edge estimates when using a neighbourhood selection approach. For example, we may consider an edge to exist if it is estimated according to one node "OR" another, or alternatively, when it is estimated by one node "AND" another. Because edges are estimated locally they may be inconsistent across nodes which can make it difficult to interpret global changes. Clearly, the OR rule will result in a graph which is less sparse, but more prone to false positives. In the context of changepoint detection for the block-constant GGM (1) the results of Kolar and Xing [2012] require one to use a union bound over the p separate nodes. The bulk of the theory in that paper operates at the node level, with the argument being that when one has successfully recovered the local structure then global structure can be recovered by combining results over nodes. In the case of GFGL, the estimator is obtained by optimising over the whole set of precision matrices jointly, and thus we have no need to combine estimates.

In practice, whether a global or local approach is more appropriate will depend on the application, and thus it is important to study both cases. For example, in certain biological applications (proteins, biomolecules etc.) node-wise changes would be more appropriate, whereas for genetic interaction networks, social interaction or brain networks global structural changes may be of more interest. To this end, the results presented here complement previous results derived in the literature, and extend our theoretical understanding of regularised fused estimators for graphical model recovery. One may note that the results here are presented in a Gaussian setting, i.e. we analyse the Gaussian generating process (1), however, the proof of the result is actually more general in that results will also hold for sub-Gaussian sampling. This suggests some level of robustness to these estimators, even when the likelihood is mis-specified. We conclude with a remark about the path towards a true high-dimensional changepoint theory and note that while in our case, the model-selection arguments can operate in high-dimensions, we have yet to prove consistency of changepoint estimation in high-dimensions. We leave this as a direction for future work.

A Proof of Changepoint Consistency

We relate the proof bounding the maximum deviation between estimated and true changepoints to the probability of an individual changepoint breaking the bound. Following Harchaoui and Lévy-Leduc [2010], we utilise the union bound

$$P\left[\max_{k\in[K]}|\tau_k-\hat{\tau}_k|\geq T\delta_T\right]\leq\sum_{k\in[K]}P\left[|\tau_k-\hat{\tau}_k|\geq T\delta_T\right]\ .$$

The complement of the event on the LHS is equivalent to the target of proof; we wish to demonstrate $P[\max_k|\tau_k-\hat{\tau}_k|\leq T\delta_T]\to 1$. In order to show this, we need to show the LHS above goes to zero as $T\to\infty$. It is sufficient, via the union bound, to demonstrate that the probability of the *bad* events:

$$A_{T,k}:=\{|\tau_k-\hat{\tau}_k|>T\delta_T\}\ , \tag{7}$$

go to zero for all $k\in[K]$. The strategy presented here separates the probability of $A_{T,k}$ occurring across complementary events. In particular, let us construct what can be thought of as a good event, where the estimated changepoints are within a region of the true ones:

$$C_T:=\left\{\max_{k\in[K]}|\hat{\tau}_k-\tau_k|<d_{\min}/2\right\}\ . \tag{8}$$

The task is then to show that $P[A_{T,k}]\to 0$ by showing $P[A_{T,k}\cap C_T]\to 0$ and $P[A_{T,k}\cap C_T^c]\to 0$ as $T\to 0$.

A.1 Stationarity induced bounds

As a first step, let us introduce some bounds based on the optimality conditions which occur in probability one. We base our results on a set of events which occur in relation to these conditions. By intersecting these events with $A_{T,k}\cap C_T$ and $A_{T,k}\cap C_T^c$, we can construct an upper bound on the probability for changepoint error exceeding a level $T\delta_T$.

Without loss of generality, consider the optimality equations (Lemma. 1) with changepoints $l=\tau_k$ and $l=\hat{\tau}_k$ such that $\hat{\tau}_k<\tau_k$. We note, that an argument for the reverse situation $\tau_k>\hat{\tau}_k$ follows through symmetry. Taking the differences between the equations we find

$$\|\textstyle\sum_{t=\hat{\tau}_k}^{\tau_k-1}(\hat{\Sigma}^{(t)}-\Sigma_0^{(t)})-\sum_{t=\hat{\tau}_k}^{\tau_k-1}W^{(t)}+\lambda_1\sum_{t=\hat{\tau}}^{\tau_k-1}\hat{R}_1^{(t)}\|_F\leq 2\lambda_2\ . \tag{9}$$

The gradient from the ℓ_1 term $\sum_{t=\hat{\tau}_k}^{\tau_k-1}\lambda\hat{R}_1^{(t)}$ can obtain a maximum value of $\pm\lambda_1(\tau_k-\hat{\tau})$ for each entry in the precision matrix. Transferring this to the RHS and splitting the LHS in terms of the stochastic and estimated terms we obtain

$$\|\textstyle\sum_{t=\hat{\tau}_k}^{\tau_k-1}(\hat{\Sigma}^{(t)}-\Sigma^{(t)})\|_F-\|\sum_{t=\hat{\tau}_k}^{\tau_k-1}W^{(t)}\|_F\leq 2\lambda_2+\lambda_1\sqrt{p(p-1)}(\tau_k-\hat{\tau}_k)\ . \tag{10}$$

The next step is to replace the time indexed inverse precision matrices $\Theta^{(t)}$ with the block-covariance matrices indexed $\Sigma_0^{(k)}$ and $\Sigma_0^{(k+1)}$. We can re-express the

difference in precision matrices as the sum of a difference between true values before τ_k , i.e. $\Sigma_0^{(k+1)} - \Sigma_0^{(k)}$, and the difference between the next $(k+1)$st true block and estimated block, i.e. $\hat{\Sigma}^{(k+1)} - \Sigma_0^{(k+1)}$ to obtain:

$$\begin{aligned}\lambda_2 + \lambda_1\sqrt{p(p-1)}(\tau_k - \hat{\tau}_k) \quad \geq \quad & \underbrace{\|\textstyle\sum_{t=\hat{\tau}_k}^{\tau_k-1}(\Sigma_0^{(k)} - \Sigma_0^{(k+1)})\|_F}_{\|R_1\|_F} - \underbrace{\|\textstyle\sum_{t=\hat{\tau}_k}^{\tau_k-1}(\hat{\Sigma}^{(k+1)} - \Sigma_0^{(k+1)})\|_F}_{\|R_2\|_F} \\ & - \underbrace{\|\textstyle\sum_{t=\hat{\tau}_k}^{\tau_k-1} W^{(t)}\|_F}_{\|R_3\|_F} , \end{aligned} \tag{11}$$

which holds with probability one. Define the events:

$$\begin{aligned} E_1 :=& \{\lambda_2 + \lambda_1\sqrt{p(p-1)}(\tau_k - \hat{\tau}_k) \geq \frac{1}{3}\|R_1\|_F\} \\ E_2 :=& \{\|R_2\|_F \geq \frac{1}{3}\|R_1\|_F\} \\ E_3 :=& \{\|R_3\|_F \geq \frac{1}{3}\|R_1\|_F\} \end{aligned}$$

Since we know that the bound (11) occurs with probability one, then the union of these three events must also occur with probability one, i.e. $P[E_1 \cup E_2 \cup E_3] = 1$.

A.2 Bounding the Good Cases

One of the three events above are required to happen, either together, or separately. We can thus use this to bound the probability of both the good C_T and bad $A_{T,k}$ events. Similar to Harchaoui and Lévy-Leduc [2010], Kolar and Xing [2012] we obtain

$$P[A_{T,k} \cap C_T] \quad \leq \quad P[\overbrace{A_{T,k} \cap C_T \cap E_1}^{A_{T,k,1}}] + P[\overbrace{A_{T,k} \cap C_T \cap E_2}^{A_{T,k,2}}] + P[\overbrace{A_{T,k} \cap C_T \cap E_3}^{A_{T,k,3}}] .$$

The following sub-sections describe how to separately bound these sub-events.

Unlike in the work of Kolar and Xing [2012], there is no stochastic element (related to the data X_t) within the first event $A_{T,k,1}$. We can bound the probability of $P[A_{T,k,1}]$ by considering the event $\{\frac{1}{3}\|R_1\|_F \leq \lambda_2 + \lambda_1\sqrt{p(p-1)}(\tau_k - \hat{\tau}_k)\}$. Given $\|R_1\|_F = \|\sum_{t=\hat{\tau}_k}^{\tau_k-1}\Sigma_0^{(k)} - \Sigma_0^{(k+1)}\|_F \geq (\tau_k - \hat{\tau}_k)\eta_{\min}$ we therefore obtain the bound

$$P[A_{T,k,1}] \leq P[(\tau_k - \hat{\tau}_k)\eta_{\min}/3 \leq \lambda_2 + \lambda_1\sqrt{p(p-1)}(\tau_k - \hat{\tau}_k)] .$$

When the events $C_T, A_{T,k}$ occur we have $T\delta_T < \tau_k - \hat{\tau}_k \leq d_{\min}/2$ to ensure the event $A_{T,k,1}$ does not occur, we need:

$$\eta_{\min} T\delta_T > 3\lambda_2 \quad ; \quad \eta_{\min} > 3\lambda_1\sqrt{p(p-1)} . \tag{12}$$

These conditions are satisfied by Assumption 1. Thus, for a large enough T, we can show that the probability $P[A_{T,k,1}] = 0$, the size of this T depends on the quantities in Eq. (12).

Now let us consider the event $A_{T,k,2}$. Consider the quantity $\bar{\tau}_k := \lfloor(\tau_k + \tau_{k+1})/2\rfloor$. On the event C_n, we have $\hat{\tau}_{k+1} > \bar{\tau}_k$ so $\hat{\Sigma}^{(t)} = \hat{\Sigma}^{(k+1)}$ for all $t \in [\tau_k, \bar{\tau}_k]$. Using the optimality conditions (Prop 1) with changepoints at $l = \bar{\tau}_k$ and $l = \tau_k$ we obtain

$$2\lambda_2 + \lambda_1\sqrt{p(p-1)}(\bar{\tau}_k - \tau_k) \geq \|\textstyle\sum_{t=\tau_k}^{\bar{\tau}_k - 1}(\hat{\Sigma}^{(k+1)} - \Sigma_0^{(k+1)})\|_F - \|\sum_{t=\tau_k}^{\bar{\tau}_k - 1} W^{(t)}\|_F \ ,$$

and thus

$$\|\hat{\Sigma}^{(k+1)} - \Sigma_0^{(k+1)}\|_F \leq \frac{4\lambda_2 + 2\lambda_1\sqrt{p(p-1)}(\bar{\tau}_k - \tau_k) + 2\|\sum_{t=\tau_k}^{\bar{\tau}_k - 1} W^{(t)}\|_F}{\tau_{k+1} - \tau_k} \ . \quad (13)$$

We now combine the bounds for events E_1 and E_2, via $E_2 := \{\|R_2\|_F \geq \frac{1}{3}\|R_1\|_F\}$ and the bounds $\|R_1\|_F \geq (\tau_k - \hat{\tau}_k)\eta_{\min}$ and $\|R_2\|_F \leq (\tau_k - \hat{\tau}_k)\|\hat{\Sigma}^{k+1} - \Sigma_0^{k+1}\|_F$. Substituting in (13) we have

$$P[A_{T,k,2}] \leq P[E_2] = P\left[\eta_{\min} \leq \frac{12\lambda_2 + 6\lambda_1\sqrt{p(p-1)}(\bar{\tau}_k - \tau_k) + 6\|\sum_{t=\tau_k}^{\bar{\tau}_k - 1} W^{(t)}\|_F}{\tau_{k+1} - \tau_k}\right] . \quad (14)$$

Splitting the probability into three components, we obtain

$$P[A_{T,k,2}] \leq P[\eta_{\min} d_{\min} \leq 12\lambda_2] + P[\eta_{\min} \leq 3\lambda_1\sqrt{p(p-1)}] + P\left[\eta_{\min} \leq \frac{6\|\sum_{t=\tau_k}^{\bar{\tau}_k - 1} W^{(t)}\|_F}{\tau_{k+1} - \tau_k}\right] . \quad (15)$$

Convergence of the first two terms follows as in $A_{T,k,1}$, the second is exactly covered in $A_{T,k,1}$; however, the third term $\eta_{\min} \leq 3\|\sum_{t=\tau_k}^{\bar{\tau}_k - 1} W^{(t)}\|_F/(\bar{\tau}_k - \tau_k)$ requires some extra treatment. As $\bar{\tau}_k < \tau_{k+1}$, we can relate the covariance matrix of the ground-truth (time-indexed) and block (indexed by k) such that $\Sigma^{(t)} = \Sigma_0^{(k)}$ for all $t \in [\tau_k, \tau_{k+1}]$. One can now write the average sampling error across time according to:

$$V_{|s^{(k)}|}^{(k)} := \hat{S}_{|s^{(k)}|}^{(k)} - \Sigma_0^{(k)} \ ,$$

where

$$\hat{S}_{|s^{(k)}|}^{(k)} := \frac{1}{|s^{(k)}|}\sum_{t \in s^{(k)}} X^{(t)}(X^{(t)})^\top$$

and $s^{(k)} \subseteq [\tau_k, \tau_{k+1}]$ is a subset of the kth changepoint interval. To simplify the notation, we will refer to the above quantities with a general subset of data $n = |s^{(k)}|$, principally, this is because in a block the samples are i.i.d so only the length n distinguishes the quantity.

Lemma 2 (Error Bound on Empirical Covariance Matrix)**.** *Let* $X^{(t)} \sim \mathcal{N}(0, \Sigma_0^{(k)})$*, or more generally be sub-Gaussian with parameter* σ_t *for* $t = 1, \ldots, n$*, then*

$$P\left(\|V_n^{(k)}\|_F \geq \epsilon\right) \leq 4p^2 \exp\left\{-n\epsilon^2/p^2 c_1\right\} \ , \tag{16}$$

where $c_1 = \max_{i,t}\{\Sigma_{0;ii}^{(t)}\} 2^7 (1 + 4\max_t\{\sigma_t^2\})^2$ *for all*

$$\epsilon \leq 2^3 p \max_{i,t}\{\Sigma_{0;ii}^{(t)}\}(1 + 4\min_t\{\sigma_t^2\})^2/(1 + 4\max_t\{\sigma_t^2\}) \ .$$

The result is a corollary of relating the empirical covariance under sub-Gaussian sampling to the sum of sub-exponential random variables, see C.1 for details.

Finally, let us turn to $A_{T,k,3}$. Recall $P(A_{T,k,3}) := P(A_{T,k} \cap C_T \cap E_3) := P(A_{T,k} \cap C_T \cap \{\| \sum_{t=\hat{\tau}_k}^{\tau_k - 1} W^{(t)}\|_F \geq \|R_1\|_F/3\})$. Given that $\|R_1\|_F \geq (\tau_k - \hat{\tau}_k)\eta_{\min}$ with probability 1, an upper bound on $P[A_{T,k,3}]$ can be found using the same concentration bounds (Lemma 2) as for $A_{T,k,2}$. The only difference is that we need to replace the integration interval n with $T\delta_T$. Noting that $T\delta_T < \tau_k - \hat{\tau}_k \leq d_{\min}/2$, the overall bound will be dominated by the concentration results requiring $n > T\delta_T$.

A.3 Bounding the Bad Cases

In order to complete the proof, we need to demonstrate that $P[A_{T,k} \cap C_T^c] \to 0$. Again, the argument below follows that of Harchaoui and Lévy-Leduc [2010], whereby the bad case is split into several events:

$$\begin{aligned} D_T^{(l)} : &= \{\exists k \in [K],\ \hat{\tau}_k \leq \tau_{k-1}\} \cap C_T^c, \\ D_T^{(m)} : &= \{\forall k \in [K],\ \tau_{k-1} < \hat{\tau}_k < \tau_{k+1}\} \cap C_T^c, \\ D_T^{(r)} : &= \{\exists k \in [K],\ \hat{\tau}_k \geq \tau_{k+1}\} \cap C_T^c, \end{aligned}$$

where $C_T^c = \{\max_{k \in [K]} |\hat{\tau}_k - \tau_k| \geq d_{\min}/2\}$ is the complement of the good event. The events above correspond to estimating a changepoint; a) before the previous true changepoint ($D_T^{(l)}$); b) between the previous and next true changepoint ($D_T^{(m)}$), and c) after the next true changepoint ($D_T^{(r)}$). The events $D_T^{(l)}$ and $D_T^{(r)}$ appear to be particularly bad as the estimated changepoint is very far from the truth, due to symmetry we can bound these events in a similar manner. Focussing on the middle term $P[A_{T,k} \cap D_T^{(m)}]$, let us again assume $\hat{\tau}_k < \tau_k$, the reverse arguments hold by symmetry.

The probability of the intersection of $A_{T,k}$ and $D_T^{(m)}$ can be bounded from above by considering the events

$$E_k^{'} := \{(\hat{\tau}_{k+1} - \tau_k) \geq d_{\min}/2\} \ , \tag{17}$$

$$E_k^{''} := \{(\tau_k - \hat{\tau}_k) \geq d_{\min}/2\} \ . \tag{18}$$

In particular, one can demonstrate that:

$$P[A_{T,k} \cap D_T^{(m)}] \leq P[A_{T,k} \cap E_k^{'} \cap D_T^{(m)}] + \sum_{j=k+1}^{K} P[E_j^{''} \cap E_j^{'} \cap D_T^{(m)}] \, . \quad (19)$$

Let us first assess $P[A_{T,k} \cap D_T^{(m)} \cap E_k^{'}]$, and consider the stationarity conditions (10) with start and end points set as $l = \hat{\tau}_k$, $l = \tau_k$ and $l = \hat{\tau}_k, l = \tau_{k+1}$. We respectively obtain:

$$|\tau_k - \hat{\tau}_k| \|\Sigma_0^{(k)} - \hat{\Sigma}^{(k+1)}\|_F \leq 2\lambda_2 + \lambda_1 \sqrt{p(p-1)}(\tau_k - \hat{\tau}_k) + \|\textstyle\sum_{t=\hat{\tau}_k}^{\tau_k - 1} W^{(t)}\|_F \quad (20)$$

and

$$|\tau_k - \hat{\tau}_{k+1}| \|\Sigma_0^{(k+1)} - \hat{\Sigma}^{(k+1)}\|_F \leq 2\lambda_2 + \lambda_1 \sqrt{p(p-1)}(\hat{\tau}_{k+1} - \tau_k) + \|\textstyle\sum_{t=\tau_k}^{\hat{\tau}_{k+1} - 1} W^{(t)}\|_F \, . \quad (21)$$

Using the triangle inequality, we bound $\|\Sigma_0^{(k+1)} - \Sigma_0^{(k)}\|_F$ conditional on $E_k^{'} := \{(\hat{\tau}_{k+1} - \tau_k) \geq d_{\min}/2\}$ and $A_{T,k} := \{|\tau_k - \hat{\tau}_k| > T\delta_T\}$. Specifically, we construct the event

$$\begin{aligned} H_T^{\Sigma} \quad := \quad & \{\|\Sigma_0^{(k+1)} - \Sigma_0^{(k)}\|_F \leq 2\lambda_1 \sqrt{p(p-1)} + 2\lambda_2((T\delta_T)^{-1} + 2/d_{\min}) \\ & + \|V_{\tau_k - \hat{\tau}_k}^{(k)}\|_F + \|V_{\hat{\tau}_{k+1} - \tau_k}^{(k+1)}\|_F\} \, , \end{aligned} \quad (22)$$

which bounds the first term of (19) such that $P[A_{T,k} \cap E_k^{'} \cap D_T^{(m)}] \leq P[H_T^{\Sigma} \cap \{\tau_k - \hat{\tau}_k \geq T\delta_T\} \cap E_k^{'}]$. Splitting the intersection of events we now have five terms to consider

$$\begin{aligned} & P(A_{T,k} \cap E_k^{'} \cap D_T^{(m)}) \\ \leq \quad & P(\lambda_1 \sqrt{p(p-1)} \geq \eta_{\min}/10) + P(\lambda_2/T\delta_T \geq \eta_{\min}/10) + P(\lambda_2/d_{\min} \geq \eta_{\min}/20) \\ & + P(\|V_{\tau_k - \hat{\tau}_k}^{(k)}\|_F \geq \eta_{\min}/5\} \cap \{\tau_k - \hat{\tau}_k \geq T\delta_T\}) \\ & + P(\{\|V_{\hat{\tau}_{k+1} - \tau_k}^{(k+1)}\|_F \geq \eta_{\min}/5\} \cap \{\hat{\tau}_{k+1} - \tau_k \geq d_{\min}/2\}) \, . \end{aligned}$$

The stochastic error terms (containing $V_{\tau_k - \hat{\tau}_k}^{(k)}$) can then be shown to converge similarly to $P(A_{T,k} \cap C_T)$. Again, it is worth noting that the term involving $T\delta_T$ will be slowest to converge, as $d_{\min} = \gamma_{\min} T > \delta_T T$ for large T. The first three terms are bounded through the assumptions on $d_{\min}, \lambda_1, \lambda_2$, and δ_T as required by the theorem (and enforce a similar requirement to those used to bound $P(A_{T,k,1})$ in Eq. 12). The other terms in (19), i.e. $\sum_{j=k+1}^{K} P[E_j^{''} \cap E_j^{'} \cap D_T^{(m)}]$ can be similarly bounded. Instead of using exactly the event H_T^{Σ} one simply replaces the term $1/T\delta_T$ in (22) with $2/d_{\min}$.

Now let us consider the events $D_T^{(l)} := \{\exists k \in [K], \, \hat{\tau}_k \leq \tau_{k-1}\} \cap C_T^c$. The final step of the proof is to show that the bound on $A_{T,k} \cap D_T^{(l)}$, and similarly $A_{T,k} \cap D_T^{(r)}$ tends to zero. To achieve this, we introduce an upper bound derived by the combinatorics of estimated changepoints:

Claim 1. *The probability of* $D_T^{(l)}$ *is bounded by*

$$P(D_T^{(l)}) \leq 2^K \sum_{k=1}^{K-1} \sum_{l \geq k}^{K-1} P(E_l^{''} \cap E_l^{'}) + 2^K P(E_K^{'}) \; .$$

We omit proof for brevity, however, remark that a similar argument is made in Harchaoui and Lévy-Leduc [2010, Eq. 31]. In order to bound the above probabilities we relate the events $E_l^{''}$ and $E_l^{'}$ to the optimality conditions as before, via Eq. 10 . Setting $k = l$ and invoking the triangle inequality gives us (similarly to Eq. 22), the event

$$J_T^{\Sigma_0} := \left\{ \|\Sigma_0^{(l+1)} - \Sigma_0^{(l)}\|_F \leq 2\lambda_1 \sqrt{p(p-1)} + M + \|V^{(l)}_{\tau_l - \hat{\tau}_l}\|_F + \|V^{(l+1)}_{\hat{\tau}_{l+1} - \tau_l}\|_F \right\} \; ,$$

where $M = 2\lambda_2(|\tau_l - \hat{\tau}_l|^{-1} + |\hat{\tau}_{l+1} - \tau_l|^{-1})$. Conditioning on the event $E_l^{''} \cap E_l^{'}$ implies that $M = 8\lambda_2/d_{\min}$. We can thus write

$$\begin{aligned} P(E_l^{''} \cap E_l^{'}) \quad \leq \quad & P(\eta_{\min} \leq 8\lambda_1 \sqrt{p(p-1)}) + P(\eta_{\min} \leq 32\lambda_2/d_{\min}) \\ & + P(\{\|V^{(l)}_{\tau_l - \hat{\tau}_l}\|_F \geq \eta_{\min}/4\} \cap \{\tau_l - \hat{\tau}_l \geq d_{\min}/2\}) \\ & + P(\{\|V^{(l+1)}_{\hat{\tau}_{l+1} - \tau_l}\|_F \geq \eta_{\min}/4\} \cap \{\hat{\tau}_{l+1} - \tau_l \geq d_{\min}/2\}) \; . \end{aligned}$$

Finally, the term corresponding to the last changepoint can be bounded by noting that when $k = K$ we have $M = 6\lambda_2/d_{\min}$, and

$$\begin{aligned} P(E_K^{''}) \leq \; & P(\eta_{\min} \leq 8\lambda_1 \sqrt{p(p-1)}) + P(\eta_{\min} \leq 24\lambda_2/d_{\min}) \\ & + P(\{\|V^{(K)}_{\tau_K - \hat{\tau}_K}\|_F \geq \eta_{\min}/4\} \cap \{\tau_K - \hat{\tau}_K \geq d_{\min}/2\}) \\ & P(\|V^{(K+1)}_{T+1-\tau_K}\|_F \geq \eta_{\min}/4) \; . \end{aligned} \tag{23}$$

A.4 Summary

The bounds derived in A1-A3 demonstrate that $P(A_{T,k}) \to 0$ since $P(A_{T,k} \cap C_T) \to 0$ and $P(A_{T,k} \cap C_T^c) \to 0$. However, to achieve these bounds, the regularisers must be set appropriately. The event $E_l^{''} \cap E_l^{'}$ establishes a minimal condition on T in conjunction with $\eta_{\min}$ and the regularisers, such that $\eta_{\min} d_{\min}/\lambda_2 > 32$ and $\eta_{\min}/\lambda_1 \sqrt{p(p-1)} > 8$. A final condition for $A_{T,k,1}$ requires $\eta_{\min} T\delta_T/\lambda_2 > 3$. Once T is large enough to satisfy these conditions, the probabilistic bound is determined either by the smallest block size $d_{\min} = \gamma_{\min} T$ or by the minimum error $T\delta_T$. Let $k_\infty = \arg\max_k \{\max_{ii} \Sigma^{(k)}_{0;ii}\}$ select the block which results in the largest expected covariance error. Summing the probabilities, one obtains the upper bound:

$$\begin{aligned} P[|\tau_k - \hat{\tau}_k| \geq T\delta_T] \leq & 2 \times 2^K \big((K-1)^2 + 1\big) P(\|V^{k_\infty; d_{\min}/2}\|_F \geq \eta_{\min}/4) \\ & + 2P(\|V^{(k_\infty)}_{T\delta_T}\|_F \geq \eta_{\min}/5) \\ & + 2P(\|V^{(k_\infty)}_{T\delta_T}\|_F \geq \eta_{\min}/3) \; , \end{aligned}$$

where the top row corresponds to $D_T^{(l)}$ and $D_T^{(r)}$; the middle $D_T^{(m)}$, and the bottom $A_{T,k,2}$ and $A_{T,k,3}$. Since $\delta_T T < \gamma_{\min} T$ for large T, the above bounds will be dominated by errors $V_{T\delta_T}^{(k_\infty)}$ integrated over the relatively small distance $T\delta_T$. A suitable overall bound on the probability is

$$\begin{aligned} P(\max_{k\in[K]} |\tau_k - \hat{\tau}_k| \geq T\delta_T) &\leq K^3 2^{K+1} P(\|V_{d_{\min}/2}^{(k_\infty)}\|_F \geq \eta_{\min}/4) \\ &\quad + 4KP(\|V_{T\delta_T}^{(k_\infty)}\|_F \geq \eta_{\min}/5) \\ &\leq C_K P(\|V_{T\delta_T}^{(k_\infty)}\|_F \geq \eta_{\min}/5) \ . \end{aligned}$$

In the Gaussian case where $\sigma_t = 1$ for all t, we have

$$P(\max_{k\in[K]} |\tau_k - \hat{\tau}_k| \geq T\delta_T) \leq C_K 4p^2 \exp\{-T\delta_T \eta_{\min}^2 / p^2 c_3\}$$

where $C_K = K(K^2 2^{K+1}+4)$, $c_3 = 5^4 2^7 \|\Sigma_0^{(k_\infty)}\|_\infty$ for all $\eta_{\min} \leq 2^3 5^2 p \|\Sigma_0^{(k_\infty)}\|_\infty$. We thus arrive at the result of Theorem 1.

B Proof of Model-Selection Consistency

Assumption 3. The event $E_\tau := \{\max_k |\hat{\tau}_k - \tau_k| \leq T\delta_T\}$ holds with some increasing probability $1 - f_\tau(T) \to 1$ as $T \to \infty$.

For the model selection consistency argument, we extend that presented in the stationary i.i.d setting as discussed in Ravikumar et al. [2011]. Specifically, this follows the *primal-dual witness* method as described in Wainwright [2009]. Since $\{\hat{\Theta}^{(k)}\}_{k=1}^B$ is an optimal solution for GFGL, for each estimated block $k, l = 1, \ldots, \hat{B} = K+1$ it needs to satisfy

$$\left\{ \sum_{l\neq k}^{\hat{B}} \hat{n}_{lk} (V_{\hat{n}_{lk}}^{(l)}) \right\} + \hat{n}_{kk} V_{\hat{n}_{kk}}^{(k)} - \hat{n}_k \hat{\Sigma}^{(k)} + \lambda_1 \hat{n}_k \hat{R}_1^{(\hat{\tau}_{k-1})} + \lambda_2 (\hat{R}_2^{(\hat{\tau}_{k-1})} - \hat{R}_2^{(\hat{\tau}_k)}) = 0 \ , \tag{24}$$

where $\hat{n}_{lk}$ describes the proportion of overlap between the lth true block and the kth estimated block. The term $\sum_{l\neq k\in[\hat{B}]} \hat{n}_{lk}(V_{\hat{n}_{lk}}^{(l)})$ can be though of as providing a sampling bias due to estimation error in the changepoints, wheras the term $\hat{n}_{kk} V^{k;\hat{n}_{kk}}$ compares samples and the ground-truth of the same underlying covariance matrix.

We will now proceed to construct an oracle estimator $\bar{\Theta} \in \mathbb{M}_{\hat{B}} := \{U^{(k)} \in \mathbb{R}^{p\times p} |\ U_{\mathcal{M}^\perp}^{(k)} = 0, U^{(k)} \succ 0\}_{k=1}^{\hat{B}}$. The oracle is constructed through solving the restricted problem

$$\begin{aligned} \bar{\Theta} := \underset{U\in\mathbb{M}_{\hat{B}}}{\arg\min} \Bigg[\sum_{k=1}^{\hat{B}} \Big\{ \sum_{l=1}^{\hat{B}} \hat{n}_{lk} \mathrm{tr}(\hat{S}_{\hat{n}_{lk}}^{(l)} U^{(k)}) - \hat{n}_k \log\det(U^{(k)}) \Big\} + \lambda_1 \sum_{k=1}^{\hat{B}} \hat{n}_k \|U^{(k)}\|_1 \\ + \lambda_2 \sum_{k=2}^{\hat{B}} \|U^{(k)} - U^{(k-1)}\|_F \Bigg] \ . \end{aligned}$$

The construction above does not utilise oracle knowledge to enforce changepoint positions, only the sparsity structure of the block-wise precision matrices. Again, for each estimate block, we obtain a set of optimality conditions like (24). Let us denote the sub-gradient of the restricted problem evaluated at the oracle solution as $\bar{R}_1^{(k)} \equiv \bar{R}_1^{(\hat{\tau}_{k-1})}$ for the ℓ_1 penalty, and $\bar{R}_2^{(\hat{\tau}_{k-1})}, \bar{R}_2^{(\hat{\tau}_k)}$ for the smoothing components. By definition the matrices $\bar{R}_2^{(\hat{\tau}_{k-1})}, \bar{R}_2^{(\hat{\tau}_k)}$ are members of the sub-differential and hence dual feasible. To show that $\bar{\Theta}$ is also a minimiser of the unrestricted GFGL problem (2), we will show that $\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty \leq 1$ and is hence dual-feasible.

Ravikumar et al. [2011, Lemma 4] demonstrates that for the standard graphical lasso problem strict dual-feasibiility can be obtained by bounding the maxima of both the sampling and estimation error. The estimation error (on the precision matrices) is tracked through the difference (remainder) between the gradient of the log-det loss function and its first-order Taylor expansion. In our case we will track the precision matrices at each block k via the *remainder function* defined as

$$\mathcal{E}(\Delta) = \bar{\Theta}^{-1} - \Theta_0^{-1} + \Theta_0^{-1}\Delta\Theta_0^{-1} \,, \tag{25}$$

where $\Delta = \bar{\Theta} - \Theta_0 \in \mathbb{R}^{p\times p}$.

Lemma 3 (Dual Feasibility)**.** *The out-of-subspace parameters are dual feasible such that* $\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty < 1$ *if*

$$\max\Big\{\|\tilde{V}^{(k)}\|_\infty, \|\mathcal{E}(\Delta)\|_\infty, \lambda_2\hat{n}_k^{-1}\|\bar{R}_2^{(\hat{\tau}_{k-1})}\|_\infty, \lambda_2\hat{n}_k^{-1}\|\bar{R}_2^{(\hat{\tau}_k)}\|_\infty\Big\} \leq \alpha\lambda_1/16 \,, \tag{26}$$

where

$$\tilde{V}^{(k)} := \hat{n}_k^{-1}(\textstyle\sum_{l\neq k}^{\hat{B}} \hat{n}_{lk} V^{(l)}_{\hat{n}_{lk}} + \hat{n}_{kk} V^{(k)}_{\hat{n}_{kk}}) \,. \tag{27}$$

We note at this point, that the condition (26) in the setting where $T \to \infty$ converges to that of the standard graphical lasso Ravikumar et al. [2011]. Specifically, if changepoint error is bounded according to the event $E_\tau := \{\max_k |\hat{\tau}_k - \tau_k| \leq T\delta_T\}$, the mis-specification error averaged across the block converges to the exact case $\tilde{V}^{(k)} \to V^{(k)}_{\text{exact}}$, where exact refers to the setting with zero changepoint estimation error.

Lemma 4. *The average sampling error over estimated block k is bounded for some $\epsilon < c_5 := 2^3 5\|\Sigma_0^{(k_\infty)}\|_\infty$ according to*

$$P[\|\tilde{V}^{(k)}\|_\infty > \epsilon] \leq 4p^2 e^{-\epsilon^2\hat{n}_k/c_4} \,,$$

for $c_4 = 2^7 5^2\|\Sigma_0^{(k_\infty)}\|_\infty$.

Applying this result to (26) and making the choice of regulariser $\lambda_1 = 16\alpha^{-1}\epsilon$, enables the condition $\|\tilde{V}^{(k)}\|_\infty \leq \alpha\lambda_1/16$ in (26) to be satisfied in high probability. Specifically, we have

$$P[\|\tilde{V}^{(k)}\|_\infty > \alpha\lambda_1 2^{-4}] \leq f_V(T) \,,$$

where $f_V(T) = 4p^2 \exp\{-c_4^{-1}\epsilon^2(\gamma_{\min} - 2\delta_T)T\}$, and note that $f_V(T) \to 0$ as $T \to \infty$.

We now turn our attention to the size of the remainder $\|\mathcal{E}(\Delta)\|_\infty$. In the first step, we directly invoke a result from Ravikumar et al. [2011]:

Lemma 5 (Ravikumar et al. [2011], Lemma 5)**.** *If the bound* $\|\Delta\|_\infty \leq (3K_{\Sigma_0}d)^{-1}$ *holds and* d *is the maxmimum node degree, then*

$$\|\mathcal{E}(\Delta)\|_\infty \leq \frac{3}{2}d\|\Delta\|_\infty^2 K_{\Sigma_0}^3 \ .$$

While we can use the same relation as Ravikumar et al. [2011] to map $\|\Delta\|_\infty$ to $\|\mathcal{E}(\Delta)\|_\infty$ we need to modify our argument for the actual control on $\|\Delta\|_\infty$.

Lemma 6. *The elementwise* ℓ_∞ *norm of the error is bounded such that* $\|\bar{\Delta}\|_\infty = \|\bar{\Theta} - \Theta_0\|_\infty \leq r$ *if*

$$r := 2K_{\Gamma_0}\{\|\tilde{V}^{(k)}\|_\infty + \lambda_1 + \lambda_2\hat{n}_k^{-1}(\|\bar{R}_2^{(\hat{\tau}_{k-1})}\|_\infty + \|\bar{R}_2^{(\hat{\tau}_k)}\|_\infty)\} \ , \qquad (28)$$

and $r \leq \min\{(3K_{\Sigma_0}d)^{-1}, (3K_{\Sigma_0}^3 K_{\Gamma_0}d)^{-1}\}$.

Note that the contribution of the fused sub-gradient is bounded $\lambda_2\hat{n}_k^{-1}(\|\bar{R}_2^{(\hat{\tau}_{k-1})}\|_\infty + \|\bar{R}_2^{(\hat{\tau}_k)}\|_\infty) \leq 2\lambda_2\hat{n}_k^{-1}$. Let us further assume that $\lambda_2 = \lambda_1\rho$ for $\rho > 0$, we now upper bound (28) with the stated form of λ_1 such that

$$r \leq r_{E_V} := 2K_{\Gamma_0}\{\epsilon + \lambda_1(1 + 2\rho\hat{n}_k^{-1})\} = 2K_{\Gamma_0}z_{\hat{n}_k}\epsilon \ ,$$

where

$$z_{\hat{n}_k} := 1 + 16\alpha^{-1}(1 + 2\rho\hat{n}_k^{-1}) \ .$$

We have two constraints on ϵ, one from the concentration bound (Lemma 5) whereby $\epsilon \leq 2^3 5\|\Sigma_0^{(k_\infty)}\|_\infty$, then a second from Lemma 6 gives

$$r \leq r_{E_V} \leq \min\{(3K_{\Sigma_0}d)^{-1}, (3K_{\Sigma_0}^3 K_{\Gamma_0}d)^{-1}\}$$

which implies $\epsilon \leq 1/v_{E_V}$ where $v_{E_V} := 6dz_{\hat{n}_k}\max\{K_{\Sigma_0}K_{\Gamma_0}, K_{\Sigma_0}^3 K_{\Gamma_0}^2\}$. Now lets use Lemma 5 to obtain

$$\begin{aligned}
\|\mathcal{E}(\Delta)\|_\infty &\leq \frac{3}{2}d\|\Delta\|_\infty^2 K_{\Sigma_0}^3 \\
&\leq 6dK_{\Gamma_0}^2 K_{\Sigma_0}^3 z_{\hat{n}_k}^2\epsilon^2 \\
&= \left[6dK_{\Gamma_0}^2 K_{\Sigma_0}^3 z_{\hat{n}_k}^2\epsilon\right] 2^{-4}\lambda_1\alpha \ ,
\end{aligned}$$

where the last line comes from setting $\epsilon = \lambda_1\alpha/16$. To demonstrate dual feasibility we therefore have to satisfy the further constraint that $\epsilon \leq 1/v_{\mathcal{E}}$ where $v_{\mathcal{E}} := 6dK_{\Gamma_0}^2 K_{\Sigma_0}^3 z_{\hat{n}_k}^2$. Dual feasibility is obtained in the case where

$$\epsilon \in \left(0, \min\left\{\frac{1}{2^3 5\|\Sigma_0^{(k_\infty)}\|_\infty}, \frac{1}{6dz_{\hat{n}_k}^2\max\{K_{\Sigma_0}K_{\Gamma_0}, K_{\Sigma_0}^3 K_{\Gamma_0}^2\}}\right\}\right) \ .$$

Consider the lower bound on $\hat{n}_k > (\gamma_{\min} - 2\delta_T)T \to \gamma_{\min} T$, as $T \to \infty$ then the term $z_{\hat{n}_k} \to 1 + 16\alpha^{-1}$. For convergence of the tail bound (Lemma 5) we need

$$\epsilon^2 T > c_4 \log(4p^2) \big/ (\gamma_{\min} - 2\delta_T) \; .$$

Thus, we can choose a rate $\epsilon = \Omega(T^{-1/2})$ and still maintain dual feasibility.

The final step of the proof is to demonstrate that all possible solutions to GFGL maintain this relation. In the case of GFGL, the following lemma states that the objective function is strictly convex on the positive-definite cone. Hence, if we find a minima it is the global minima, and the dual-feasibility condition ensures that the suggested bounds are achieved.

Lemma 7 (Strict Convexity)**.** *For matrices $\Theta_T \in \mathcal{S}_{++}^T := \{\{U^{(t)}\}_{t=1}^T \,|\, U^{(t)} \succ 0 \,,\, U^{(t)} = U^{(t)\top}\}$ the GFGL cost function is strictly convex.*

C Proof of Lemmata

C.1 Bounds for Empirical Covariance Error

Proof of Lemma 2. We want to bound $P[\|V_n^{(k)}\|_\infty \geq \epsilon]$. Our approach is to extend the single block tail-bound of Ravikumar et al. [2011], derived in an i.i.d setting, to the case where samples are independent, but not necessarily identically sampled. Without loss of generality consider the event

$$\mathbb{A}_{ij}(v) := \left\{ |\sum_{t=1}^{n} X_i^{(t)} X_j^{(t)} - \Sigma_{0;ij}^{(t)}| > d \right\} \; . \tag{29}$$

Recall, the deviation of a sub-exponential random variable Z with $E[Z] = \mu_Z$ is given by the inequality

$$P(Z \geq \mu_Z + v) \leq \begin{cases} \exp(\frac{-v^2}{2\gamma^2}) & \text{if } 0 \leq v \leq \gamma^2/b \\ \exp(-\frac{v}{2b}) & \text{if } v > \gamma^2/b \end{cases} \; . \tag{30}$$

For independent $Z^{(t)}$, sub-exponentials with parameters (γ_t^2, b_t) the sum $Z_{ij} = Z_{ij}^{(1)} + \ldots + Z_{ij}^{(t)}$ is sub-exponential with parameters $(\sum_t \gamma_t^2, \max_t b_t)$. Now consider the event $\mathbb{A}_{ij}(v)$, and construct the auxilary variables $A_{ij}^{(t)} := \check{X}_i^{(t)} + \check{X}_j^{(t)}$ and $B_{ij}^{(t)} := \check{X}_i^{(t)} - \check{X}_j^{(t)}$ where $\check{X}_i^{(t)} = X_i^{(t)} / \sqrt{\Sigma_{0;ii}^{(t)}}$. We note that Lemma 9 of Ravikumar et al. [2011] holds at the individual t step level such that, if $\bar{X}_i^{(t)}$ is sub-Gaussian with parameter σ_t, then the random variables $A_{ij}^{(t)}$ and $B_{ij}^{(t)}$ are sub-Gaussian with parameter $2\sigma_t$, and for all $d > 0$

$$\begin{aligned} P[\mathbb{A}_{ij}(v)] \leq P\left[|\sum_{t=1}^{n} (A_{ij}^{(t)})^2 - 2(1 - \rho_{ij}^*)| > 2d \big/ \sqrt{\Sigma_{0;ii}^{(t)} \Sigma_{0;jj}^{(t)}} \right] \\ + P\left[|\sum_{t=1}^{n} (B_{ij}^{(t)})^2 - 2(1 - \rho_{ij}^*)| > 2d \big/ \sqrt{\Sigma_{0;ii}^{(t)} \Sigma_{0;jj}^{(t)}} | \right] \; . \end{aligned}$$

Through Lemma 10 (Ravikumar et al. [2011]) we have $Z_{ij}^{(t)} := (A_{ij}^{(t)})^2 - 2(1+\rho_{ij}^{(t)})$ is sub-exponential with parameter $\gamma = b = 16(1+4\sigma^2)$. Application of (30) with the tighter bound ($0 \leq v \leq \gamma^2/b$), gives us

$$P\left[|\sum_{t=1}^{n} (A_{ij}^{(t)})^2 - 2(1-\rho_{ij}^{*})| > 2d/\sqrt{\Sigma_{0;ii}^{(t)}\Sigma_{0;jj}^{(t)}} \right] \leq 2\exp\left\{-d^2/nc_1\right\} ,$$

where $c_1 = \max_{i,t}\{\Sigma_{0;ii}^{(t)}\} 2^7 (1 + 4\max_t\{\sigma_t^2\})^2$ for all $d \leq nc_2$ where $c_2 = 2^3 \max_{i,t}\{\Sigma_{0;ii}^{(t)}\}(1 + 4\min_t\{\sigma_t^2\})^2/(1 + 4\max_t\{\sigma_t^2\})$. A bound on the required quantity $P(\|B_n^{(k)}\|_\infty \geq \epsilon)$ follows from application of the union bound over both A and B, and then further over the individual elements in the $p \times p$ matrices. In a general setting we obtain

$$P\left[\|\sum_{t=1}^{n} (\check{X}^{(t)})^\top \check{X}^{(t)} - \Sigma_0^{(t)}\|_\infty > d \right] \leq 4p^2 \exp\left\{-d^2/nc_1\right\} , \tag{31}$$

where $c_1 = \max_{i,t}\{\Sigma_{0;ii}^{(t)}\} 2^7 (1 + 4\max_t\{\sigma_t^2\})^2$ for all

$$d \leq 2^3 n \max_{i,t}\{\Sigma_{0;ii}^{(t)}\}(1 + 4\min_t\{\sigma_t^2\})^2/(1 + 4\max_t\{\sigma_t^2\}) .$$

The required bound is given by the inequality $\|X\|_F \leq p\|X\|_\infty$ and setting $d = \epsilon n$ with $\epsilon \leq c_2$.

□

Proof of Lemma 4. For the quantity $\|\tilde{V}^{(k)}\|_\infty$ we consider adapting the proof of Lemma 2. A bound on this can be derived from (31) setting $n = \hat{n}_k$, $d = \epsilon\hat{n}$ letting $\max_{i,t}\{\Sigma_{0;ii}^{(t)}\} = \|\Sigma_0^{(k_\infty)}\|_\infty$ and $\max_t\{\sigma_t^2\} = \min_t\{\sigma_t^2\} = 1$, i.e. all variates are Gaussian. We then obtain

$$P[\|\tilde{V}^{(k)}\|_\infty > \epsilon] \leq 4p^2 \exp\{-\epsilon^2\hat{n}_k/c_4\} ,$$

for $c_4 = 2^7 5^2 \|\Sigma_0^{(k_\infty)}\|_\infty$ and $\epsilon < c_5 := 2^3 5 \|\Sigma_0^{(k_\infty)}\|_\infty$. □

C.2 Dual-feasibility with Mis-Specification (Proof of Lemma 3)

Proof. We can write the block-wise optimality conditions (24) for the restricted estimator as

$$(\Theta_0^{(k)})^{-1}\Delta^{(k)}(\Theta_0^{(k)})^{-1} - \mathcal{E}(\Delta^{(k)}) + \frac{1}{\hat{n}_k}\left(\sum_{l\neq k}^{\hat{B}} \hat{n}_{lk} V_{\hat{n}_{lk}}^{(l)} + \hat{n}_{kk} V_{\hat{n}_{kk}}^{(k)}\right) + \lambda_1 \bar{R}_1^{(k)} + \frac{\lambda_2}{\hat{n}_k}(\bar{R}_2^{(\hat{\tau}_{k-1})} - \bar{R}_2^{(\hat{\tau}_k)}) = 0 .$$

As pointed out in Ravikumar et al. [2011], this equation may be written as an ordinary linear equation by vectorising the matrices, for instance $\text{vec}\{(\Theta_0^{(k)})^{-1}\Delta^{(k)}(\Theta_0^{(k)})^{-1}\} = \{(\Theta_0^{(k)})^{-1} \otimes (\Theta_0^{(k)})^{-1}\}\text{vec}(\Delta^{(k)}) \equiv \Gamma_0\text{vec}(\Delta)$. Utilising the fact $\Delta_{\mathcal{M}^\perp} = 0$ we can split the optimality conditions into two blocks of linear equations

$$\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}}\text{vec}(\Delta^{(k)}_{\mathcal{M}}) + \text{vec}(G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}}) = 0 \tag{32}$$

$$\Gamma^{(k)}_{0;\mathcal{M}^\perp\mathcal{M}}\text{vec}(\Delta^{(k)}_{\mathcal{M}}) + \text{vec}(G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}^\perp}) = 0\ , \tag{33}$$

where

$$G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2) := \tilde{V}^{(k)} - \mathcal{E}(\Delta^{(k)}) + \lambda_1\bar{R}^{(k)}_1 + \hat{n}_k^{-1}\lambda_2(\bar{R}^{(\hat{\tau}_{k-1})}_2 - \bar{R}^{(\hat{\tau}_k)}_2)\ .$$

Solving (32) for $\text{vec}(\Delta^{(k)}_{\mathcal{M}})$ we find $\text{vec}(\Delta^{(k)}_{\mathcal{M}}) = -(\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}})^{-1}\text{vec}\{G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}}\}$. Substituting this into (33) and re-arranging for $\bar{R}_{1;\mathcal{M}^\perp}$ gives

$$\text{vec}(G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}^\perp}) = \Gamma^{(k)}_{0;\mathcal{M}^\perp\mathcal{M}}(\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}})^{-1}\text{vec}\{G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}}\}\ ,$$

and thus, letting $H^{(k)} := \Gamma^{(k)}_{0;\mathcal{M}^\perp\mathcal{M}}(\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}})^{-1}$ we obtain

$$\begin{aligned}\bar{R}^{(k)}_{1;\mathcal{M}^\perp} = \frac{1}{\lambda_1}H^{(k)}\text{vec}\{\tilde{V}^{(k)}_{\mathcal{M}} - \mathcal{E}_{\mathcal{M}}(\Delta^{(k)})\} + \frac{\lambda_2}{\hat{n}_k\lambda_1}H^{(k)}\text{vec}\{(\bar{R}^{(\hat{\tau}_{k-1})}_2 - \bar{R}^{(\hat{\tau}_k)}_2)_{\mathcal{M}}\} + H^{(k)}\text{vec}(\bar{R}^{(k)}_{1;\mathcal{M}}) \\ - \frac{1}{\lambda_1}\text{vec}\{\tilde{V}^{(k)}_{\mathcal{M}^\perp} - \mathcal{E}_{\mathcal{M}^\perp}(\Delta^{(k)})\} - \frac{\lambda_2}{\hat{n}_k\lambda_1}\text{vec}\{(\bar{R}^{(\hat{\tau}_{k-1})}_2 - \bar{R}^{(\hat{\tau}_k)}_2)_{\mathcal{M}^\perp}\}\ .\end{aligned}$$

Taking the ℓ_∞ norm of both sides gives

$$\begin{aligned}\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty \leq \frac{1}{\lambda_1}|||H^{(k)}|||_\infty(\|\tilde{V}^{(k)}_{\mathcal{M}}\|_\infty + \|\mathcal{E}_{\mathcal{M}}(\Delta^{(k)})\|_\infty) + \|H^{(k)}\text{vec}(\bar{R}^{(k)}_{1;\mathcal{M}})\|_\infty \\ + \frac{1}{\lambda_1}(\|\tilde{V}^{(k)}_{\mathcal{M}^\perp}\|_\infty + \|\mathcal{E}_{\mathcal{M}^\perp}(\Delta^{(k)})\|_\infty) \\ + \frac{\lambda_2}{\hat{n}_k\lambda_1}\left\{|||H^{(k)}|||_\infty(\|\bar{R}^{(\hat{\tau}_{k-1})}_{2;\mathcal{M}}\|_\infty + \|\bar{R}^{(\hat{\tau}_k)}_{2;\mathcal{M}}\|_\infty) + \|\bar{R}^{(\hat{\tau}_{k-1})}_{2;\mathcal{M}^\perp}\|_\infty + \|\bar{R}^{(\hat{\tau}_k)}_{2;\mathcal{M}^\perp}\|_\infty\right\}\ .\end{aligned}$$

Claim 2. *The error in the model-space dominates that outside such that*

$$\|\tilde{V}^{(k)}_{\mathcal{M}^\perp}\|_\infty \leq \|\tilde{V}^{(k)}_{\mathcal{M}}\|_\infty\ , \tag{34}$$

$$\|\mathcal{E}_{\mathcal{M}^\perp}(\Delta^{(k)})\|_\infty \leq \|\mathcal{E}_{\mathcal{M}}(\Delta^{(k)})\|_\infty\ . \tag{35}$$

Furthermore, the maximum size of the sub-gradient in the model subspace is bounded $\|\bar{R}^{(k)}_{1;\mathcal{M}}\|_\infty \leq 1$.

A similar claim in made in Ravikumar et al. [2011], and thus via the results above, we obtain $\|H^{(k)}\text{vec}(\bar{R}^{(k)}_{1;\mathcal{M}})\|_\infty \leq 1-\alpha$ and

$$\begin{aligned}\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty \leq \lambda_1^{-1}(2-\alpha)\{\|\tilde{V}^{(k)}_{\mathcal{M}}\|_\infty + \|\mathcal{E}_{\mathcal{M}}(\Delta^{(k)})\|_\infty \\ + \lambda_2\hat{n}_k^{-1}(\|\bar{R}^{(\hat{\tau}_{k-1})}_{2;\mathcal{M}}\|_\infty + \|\bar{R}^{(\hat{\tau}_k)}_{2;\mathcal{M}}\|_\infty)\} + \|H^{(k)}\text{vec}(\bar{R}^{(k)}_{1;\mathcal{M}})\|_\infty\ .\end{aligned}$$

The condition (26) stated in the lemma now ensures $\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty < 1$. □

C.3 Control of Estimation Error (Proof of Lemma 6)

Note that $\bar{\Theta}^{(k)}_{\mathcal{M}^\perp} = \Theta^{(k)}_{0;\mathcal{M}^\perp} = 0$ and thus $\|\Delta^{(k)}\|_\infty = \|\Delta^{(k)}_{\mathcal{M}}\|_\infty$. We follow Lemma 6 from Ravikumar et al. [2011] in the spirit of our proof. The first step is to characterise the solution $\bar{\Theta}_{\mathcal{M}}$ in terms of its zero-gradient condition (of the restricted oracle problem). Define a function to represent the block-wise optimality conditions (akin to Eq. 75 Ravikumar et al. [2011])

$$Q(\Theta^{(k)}_{\mathcal{M}}) = -(\Theta^{(k)}_{\mathcal{M}})^{-1} + \hat{n}_k^{-1}\left(\sum_{t=\hat{\tau}_{k-1}}^{\hat{\tau}_k - 1} \hat{S}^{(t)}_{\mathcal{M}}\right) + \lambda_1 \bar{R}^{(k)}_1 + \lambda_2 \hat{n}_k^{-1}(\bar{R}^{(\hat{\tau}_{k-1})}_2 - \bar{R}^{(\hat{\tau}_k)}_2) = 0 \ .$$

Now construct a map $F : \Delta^{(k)}_{\mathcal{M}} \mapsto F(\Delta^{(k)}_{\mathcal{M}})$ such that its fixed points are equivalent to the zeros of the gradient expression in terms of $\Delta^{(k)}_{\mathcal{M}}$. To simplify the analysis, let us work with the vectorised form and define the map

$$F(\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}})) := -(\Gamma_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\{Q(\Theta^{(k)}_{\mathcal{M}})\} + \mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}) \ ,$$

such that $F\{\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}})\} = \mathrm{vec}(\Delta^{(k)}_{\mathcal{M}})$ iff $Q(\Theta^{(k)}_{0;\mathcal{M}} + \Delta^{(k)}_{\mathcal{M}}) = Q(\Theta^{(k)}_{\mathcal{M}}) = 0$. Now, to ensure all solutions that satisfy the zero gradient expression may have their error bounded within the ball we demonstrate that F maps a ℓ_∞ ball $\mathbb{B}(r) := \{\Theta^{(k)}_{\mathcal{M}} \,|\, \|\Theta^{(k)}_{\mathcal{M}}\|_\infty \leq r\}$ onto itself. Expanding $F(\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}))$, we find

$$\begin{aligned} F(\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}})) &= -(\Gamma_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\{Q(\Theta^{(k)}_{0;\mathcal{M}} + \Delta^{(k)}_{\mathcal{M}})\} + \mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}) \\ &= T_1 - T_2 \ , \end{aligned}$$

where

$$\begin{aligned} T_1 &:= (\Gamma_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\big[\{(\Theta^{(k)}_0 + \Delta^{(k)})^{-1} - (\Theta^{(k)}_0)^{-1}\}_{\mathcal{M}} + \mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}) \\ T_2 &:= (\Gamma_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\big[\tilde{V}^{(k)}_{\mathcal{M}} + \lambda_1 \bar{R}^{(k)}_{1;\mathcal{M}} + \lambda_2 \hat{n}_k^{-1}(\bar{R}^{(\hat{\tau}_{k-1})}_{2;\mathcal{M}} - \bar{R}^{(\hat{\tau}_k)}_{2;\mathcal{M}})\big] \ . \end{aligned}$$

The rest of the proof follows from Ravikumar et al. [2011], where via Lemma 5, one can show

$$\|T_1\|_\infty \leq \frac{3}{2} d K^3_{\Sigma_0} K_{\Gamma_0} \|\Delta^{(k)}\|^2_\infty \leq \frac{3}{2} d K^3_{\Sigma_0} K_{\Gamma_0} r^2 \ ,$$

under the assumptions of the lemma we obtain $\|T_1\| \leq r/2$. Combined with the stated form of r, we also find $\|T_2\|_\infty \leq r/2$ and thus $\|F(\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}))\|_\infty \leq r$. Through the construction of F, we have $\|\Delta^{(k)}_{\mathcal{M}}\|_\infty \leq r$ iff $Q(\Theta^{(k)}_{0;\mathcal{M}} + \Delta^{(k)}_{\mathcal{M}}) = Q(\Theta^{(k)}_{\mathcal{M}}) = 0$ and since $Q(\bar{\Theta}^{(k)}_{\mathcal{M}}) = 0$ for any $\bar{\Theta}^{(k)}_{\mathcal{M}}$ we obtain $\|\bar{\Delta}^{(k)}_{\mathcal{M}}\|_\infty \leq r$ where $\bar{\Delta}^{(k)}_{\mathcal{M}} := \bar{\Theta}^{(k)} - \Theta^{(k)}_0$. Finally, the existence of a solution $\bar{\Theta}^{(k)}_{\mathcal{M}}$ corresponding to $\mathrm{vec}(\bar{\Delta}^{(k)}_{\mathcal{M}}) \in \mathbb{B}(r)$ is guaranteed by Brouwer's fixed point theorem.

D Further Results

D.1 Strict Convexity of GFGL (Proof of Lemma 7)

Proof. The negative log-det barrier $-\log\det(\Theta^{(t)})$ is strictly convex on $\Theta^{(t)} \in \mathcal{S}_{++}^1$. While the frobenius norm is strictly convex on a given matrix $\|A\|_F$ for $A \in \mathcal{S}_{++}^1$ it is not strictly convex when considering the mixed norm $\sum_{t=2}^T \|\Theta^{(t)} - \Theta^{(t-1)}\|_F$ for $\Theta^{(t)} \in \mathcal{S}_{++}^{(T)}$. However, due to Lagrangian duality, we can re-write the GFGL problem as an explicitly constrained problem

$$\min_{\Theta_T \in \mathcal{S}_{++}^{(T)}} \Big\{ \sum_{t=1}^T \big[\langle \Theta^{(t)}, \hat{S}^{(t)} \rangle - \log\det(\Theta^{(t)}) \big] \Big\} \text{ such that}$$

$$\sum_{t=1}^T \|\Theta_{-ii}^{(t)}\|_1 + \frac{\lambda_2}{\lambda_1} \sum_{t=2}^T \|\Theta^{(t)} - \Theta^{(t-1)}\|_F \leq C(\lambda_1) \; .$$

We can alternatively write

$$\min_{\Theta_T \in \mathcal{S}_{++}^{(T)}} \Big\{ \sum_{t=1}^T \big[\langle \Theta^{(t)}, \hat{S}^{(t)} \rangle - \log\det(\Theta^{(t)}) \big] \Big\}$$

$$\text{such that} \quad \sum_{t=1}^T \|\Theta_{-ii}^{(t)}\|_1 \leq C_{\text{sparse}}(\lambda_1, \lambda_2) \quad \text{and} \quad \sum_{t=2}^T \|\Theta^{(t)} - \Theta^{(t-1)}\|_F \leq C_{\text{smooth}}(\lambda_1, \lambda_2) \; .$$

A similar argument to that used in Ravikumar et al. [2011] now holds. Specifically, we note that even the rank one estimate $\hat{S}^{(t)}$ will have positive diagonal entries $\hat{S}_{ii}^{(t)} > 0$ for all $i = 1, \ldots, p$. The off-diagonal entries in the precision matrix are restricted through the ℓ_1 term. Unlike in the standard static case, the size of this norm is not related to just a single precision matrix, rather it counts the size of the off-diagonals over the whole set $\{\Theta^{(t)}\}_{t=1}^T$. Thus, to obtain strict convexity, one also needs to include appropriate smoothing. To borrow the same argument as used in Ravikumar et al. [2011], we need to demonstrate that for any time-point t we can construct a problem of the form $\min_{\Theta^{(t)} \in \mathcal{S}_{++}^{(1)}} \big\{ \langle \Theta^{(t)}, \hat{S}^{(t)} \rangle - \log\det(\Theta^{(t)}) \big\}$ such that $\|\Theta_{-ii}^{(t)}\|_1 \leq C_t(\lambda_1, \lambda_2)$. The constraint due to smoothing allows exactly this, for instance, one may obtain a bound $\|\Theta_{-ii}^{(t)}\|_1 \leq C_{\text{sparse}}(\lambda_1, \lambda_2) - \sum_{s \neq t} \|\Theta_{-ii}^{(s)}\|_1$. Writing $\Theta^{(s)} = \Theta^{(1)} + \sum_{q=2}^s (\Theta^{(q)} - \Theta^{(q-1)})$ for $s \geq 2$ we obtain

$$\begin{aligned} \sum_{s \neq t} \|\Theta^{(s)}\|_1 &\leq \|\Theta^{(1)}\|_1 + \sum_{s \neq t} \sum_{q=1}^s \|\Theta^{(q)} - \Theta^{(q-1)}\|_1 \\ &\leq p C_{\text{smooth}}(\lambda_1, \lambda_2) + \|\Theta^{(1)}\|_1 \; , \end{aligned}$$

where we note $\|\cdot\|_1 \leq p\|\cdot\|_F$. Converting to the bound for the ℓ_1 norm at time t we find

$$\|\Theta_{-ii}^{(t)}\|_1 \leq C_{\text{sparse}}(\lambda_1, \lambda_2) - p C_{\text{smooth}}(\lambda_1, \lambda_2) - \|\Theta^{(1)}\|_1 \equiv C_t(\lambda_1, \lambda_2) \; ,$$

and thus an effective bound on the time-specific ℓ_1 norm can be obtained. □

D.2 Consistency when $\hat{K} \geq K$ (Proof Sketch for Proposition 1)

Proof. The below constitutes a sketch of the proof for Proposition 1. Let us start by recalling the assumption $K \leq \hat{K} \leq K_{\max}$. We can split the bound into two, one corresponding to the correctly identified number of changes and one for those which exceed this number:

$$P\left[\left\{h(\hat{\mathcal{T}}_{\hat{K}} \| \mathcal{T}_K) \geq T\delta_T\right\} \cap \left\{K \leq \hat{K} \leq K_{\max}\right\}\right] \leq \underbrace{P[h(\hat{\mathcal{T}}_K \| \mathcal{T}_K) \geq T\delta_T]}_{\to 0 \text{ via Thm.1}} + \sum_{L>K}^{K_{\max}} P[h(\hat{\mathcal{T}}_L \| \mathcal{T}_K) \geq T\delta_T] \, .$$

Focusing on the second bound, Harchaoui and Lévy-Leduc [2010] demonstrates this can be broken down in terms of three events:

$$\sum_{L>K}^{K_{\max}} P[h(\hat{\mathcal{T}}_L \| \mathcal{T}_K) \geq T\delta_T] \leq \sum_{L>K}^{K_{\max}} \sum_{k=1}^{K} P[F_{T,k,1}] + P[F_{T,k,3}] + P[F_{T,k,3}]$$

where

$$F_{T,k,1} = \{|\hat{\tau}_l - \tau_k| \geq T\delta_T \text{ and } \hat{\tau}_l < \tau_k \, , \, \forall 1 \leq l \leq L\}$$

$$F_{T,k,2} = \{|\hat{\tau}_l - \tau_k| \geq T\delta_T \text{ and } \hat{\tau}_l > \tau_k \, , \, \forall 1 \leq l \leq L\}$$

$$F_{T,k,3} = \{\exists 1 \leq l \leq L \,|\, |\hat{\tau}_l - \tau_k| \geq T\delta_T, \, |\hat{\tau}_{l+1} - \tau_k| \geq T\delta_T, \text{ and } \hat{\tau}_l < \tau_k < \hat{\tau}_{l+1}\} \, .$$

The probability of these events can all be bounded in a similar way to the that of Eq. 14, i.e. by considering specific start and end-points in the optimality conditions given by Lemma 1. In the interests of space, we refer the reader to Harchaoui and Lévy-Leduc [2010] for technical details of this procedure.

Finally, in the statement of the above proposition, we asserted that there was some $K_{\max}$ which upper bounded the number of estimated changepoints. By considering the bias of the GFGL estimator as a function of specific λ_1, λ_2 which scale as discussed in the main paper one should be able to demonstrate that the estimated number of changepoints is indeed bounded from above. In the seminal paper of Harchaoui and Lévy-Leduc [2010], this is achieved via Lemma 2 of Meinshausen and Yu [2009]. In the case of GFGL, with the log-determinant constrained likelihood and the additional regulariser, a new result would be needed to upper bound $\hat{K}$. One possible pathway to achieving this would be via the analysis provided in Rothman et al. [2008], and extending this to the non-stationary setting. Proof of such a specific result is left as future work.

□

References

D. Angelosante and G. B. Giannakis. Sparse graphical modeling of piecewise-stationary time-series. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1960–1963, 2011.

J. Bai. Estimation of a change point in multiple regression models. *The Review of Economics and Statistics*, 79(4):551–563, 1997.

O. Banerjee and L. E. Ghaoui. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014.

A. J. Gibberd and J. D. B. Nelson. High dimensional changepoint detection with a dynamic graphical lasso. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2684–2688, 2014.

A. J. Gibberd and J. D. B. Nelson. Regularized estimation of piecewise constant gaussian graphical models : The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 2017.

D. Hallac, Y. Park, S. Boyd, and J. Leskovec. Network inference via the time-varying graphical lasso. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492): 1480–1493, 2010.

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

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

M. Kolar and E. P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6:2069–2106, 2012.

Steffen L. Lauritzen. *Graphical Models*. OUP Oxford, 1996.

S. Lee, M. H. Seo, and Y. Shin. The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(1):193–210, 2016.

F. Leonardi and P. Bühlmann. Computationally efficient change point detection for high-dimensional regression. pages 1–32, 2016. URL `http://arxiv.org/abs/1601.03704`.

N. Meinshausen. A note on the lasso for gaussian graphical model selection. *Statistics & Probability Letters*, 78:880–884, 2008.

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

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

R. P. Monti, P. Hellyer, D. Sharp, R. Leech, S. Anagnostopoulos, and G. Montana. Estimating time-varying brain connectivity networks from functional mri time series. *NeuroImage*, 103:427–443, 2014.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

M. Raimondo. Minimax estimation of sharp change points. *Annals of Statistics*, 26(4):1379–1397, 1998.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(January 2010):935–980, 2011.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2(0):494–515, 2008.

S. Roy, Y. Atchadé, and G. Michailidis. Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.