Multiple Changepoint Estimation in High-Dimensional Gaussian Graphical Models

A. Gibberd and S. Roy

December 18, 2017

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 only 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.

1 Introduction

1.1 Motivation

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 datsets 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.

In this paper we consider a particular type of dynamic network model where the underlying dependency structure evolves in a piecewise fashion. We desire to estimate multiple change-points where the network structure changes, as well as the structures themselves. To this end, the task is formulated as a joint optimization problem such that 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 over certain blocks of time. Specifically, we assume observations $X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ are drawn from the following model

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

where $t = 1, \ldots, T$ indexes the time of the observed data-point, and $k = 1, \ldots, K+1$ indexes blocks seperated by changepoints $\{\tau_k\}_{k=1}^K$, at which point the covariance matrix $\Sigma_0^{(k)}$ changes. The task is to assess, how well, or indeed if, we can recover both the changepoint positions τ_k and the correct precision matrices $\Theta_0^{(k)} := (\Sigma_0^{(k)})^{-1}$.

1.2 Literature Review

Although the topic of change-point estimation is well represented in statistics (see Bai [1997], Hinkley [1970], Loader [1996], Lan et al. [2009], Muller [1992], Raimondo [1998] and references therein), its application in high dimensional graphical models is relatively unexplored. We will first review the literature on change-point detection before moving onto graphical model estimation and the intersection of these methodologies.

One of the most popular methods of change-point detection is binary segmentation [Fryzlewicz, 2014, Cho and Fryzlewicz, 2015]. The method hinges on dividing the entire data into two segments based on a discrepancy measure and relating the procedure on subsequent segments until there are no change-point available. Usually a cumulative summation type test-statistic [Page, 1954] is used to compute the discrepancy based on possible two segments of the data. Further approaches for multiple change-point estimation in multivariate time series (not necessarily high dimensional) include: SLEX (Smooth Localized Complex Exponentials), a complexity penalized optimization for time series segmentation [Ombao et al., 2005]; approaches which utilise a penalised gaussian likelihood via dynamic programming [Lavielle and Teyssiere, 2006, Angelosante and Giannakis, 2011]; or penalized regression utilising the group lasso [Bleakley and Vert, 2011]. More recently, work on high dimensional time series segmentation includes: sparsified binary segmentation via thresholding CUSUM statistics [Cho and Fryzlewicz, 2015]; change-point estimation for high dimensional time series with missing data via subspace tracking [Xie et al., 2013], and projection based approaches [Wang and Samworth, 2017]. Some recent works on high dimensional regression with change-points are also worth mentioning here. For instance Lee et al. [2016] consider high dimensional regression with a possible change-point due to a covariate threshold and develop a lasso based estimator for the regression coefficients as well as the threshold parameter. Leonardi and Bühlmann [2016] extend this to multiple change-point detection in high dimensional regression, the resultant estimator can be used in conjunction with dynamic programming or in an approximate setting via binary segmentation. The theoretical guarantees for the two methods appear to be similar and estimate both the number and locations of change-points, as well as the parameters in the corresponding segments.

Moving on to graph estimation, in the static i.i.d. case there are two principle approaches to estimating conditional independence graphs. Firstly, as suggested by

Meinshausen and Bühlmann [2006], one may adopt a neighbourhood or *local* selection approach where edges are estimated at a nodewise 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 dynamic setting, one could consider extending static neighbourhood selection 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 neighbourhood selection 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 the node level. In the global estimation setting, Angelosante and Giannakis [2011] proposed to combine the graphical lasso with dynamic programming to estimate changepoints and graph structures. Alternatively, one may consider smoothing variation 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].

1.3 Paper Contribution

Unlike previous literature in high dimensional time series segmentation or high dimensional regression settings with possible change-points, our framework is in the context of a graphical model that changes its underlying structure with time in a piecewise manner. We therefore desire to detect jointly, both the change-points and the parameters specifying the underlying network structure. To achieve such estimation, we construct an M-estimator which jointly penalises sparsity in the graph structure while additionally smoothing the structure over time.

In this paper we focus on describing multiple change-point impacting the global structural change in the underlying Gaussian graphical model. Whether a global, or local approach is appropriate will depend on the application. 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. Additionally, in many situations it is not trivial how we combine edge estimates when performing a neighbourhood selection approach. Because edges are estimated locally they may be inconsistent across nodes which can make it difficult to interpret global changes.

To avoid these problems, we opt to perform global estimation and assume the presence of changepoints which impact many of the edges together. Specifically, this paper deals with analysing the Group-Fused Graphical lasso (GFGL) estimator first proposed in Gibberd and Nelson [2017]. In previous work it was demonstrated that empirically GFGL can detect both changepoints and graphical structure while operating in relatively high-dimensional settings. However, until now, the theoretical consistency properties of the estimator have remained unknown. In this paper, we derive rates for the consistent recovery of both changepoints and model structure via upper bounds on the

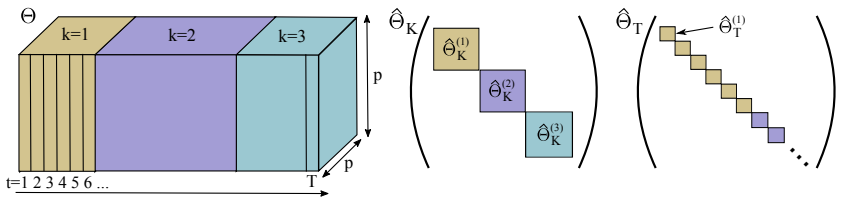

Figure 1: Diagramatic representation of matrix indexing and notation.

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. We discuss the obtained convergence rates with those already described in the literature, such as changepoint recovery rates with fused neighbourhood selection [Kolar and Xing, 2012], and node-wise binary-segmentation/dynamic programming routines [Leonardi and Bühlmann, 2016]. Our theoretical results demonstrate that we can asymptotically recover precision matrices, even in the presence of a-priori unknown changepoints. We demonstrate that error bounds of a similar form to those presented in the static i.i.d. graphical lasso case Ravikumar et al. [2011] can be obtained for sufficient T and appropriate smoothing regularisation. Using such bounds, we attempt to quantify the price one pays for allowing non-homogeneous sampling and the presence of changepoints. To this end, our results indicate the efficiency of precision matrix estimation with GFGL may be limited for shorter sequences due to the influence and bias imposed by our assumed smoothing regularisers.

1.4 Notation

Throughout the paper we will use a kind of notational overloading whereby $\Theta^{(k)} \in \mathbb{R}^{p\times p}$ refers to a *block* indexed precision matrix and $\Theta^{(t)}$ refers to a *time*-indexed one, for $t = 1, \ldots, T$. In general, each block $k = 1, \ldots, K+1$ will contain multiple timepoints. It is expected that many $\Theta^{(t)}$ will be the same, however, only a few $\Theta^{(k)}$ will be similar, a diagrammatic overview of the notation can be found in Figure 1.

Vector norms ℓ_a are denoted $\|x\|_a$, similarly the application of vector norms to a vectorised matrix is denoted $\|X\|_a$. For instance, the maximum element of a matrix is denoted $\|X\|_\infty = \max_{i,j} |X_{ij}|$. Where the Frobenius norm is used, this is denoted $\|X\|_F$. Additionally, we utilise the operator norm $|||X|||_2$ which denotes the largest singular values of X, and the $\ell_{\infty,\infty}$ norm $|||X|||_\infty := \max_i \sum_j |X_{ij}|$.

2 Estimator Formulation and Computation

2.1 The Group-Fused Graphical Lasso

To simultaneously estimate multivariate Gaussian model structure alongside changepoints, we propose to use the GFGL estimator first examined in Gibberd and Nelson [2017]. This takes multivariate observations $x^{(t)} \in \mathbb{R}^p$ for $t = 1, \ldots, T$ and estimates a set of T precision matrices represented in a block-diagonal model space

$\hat{\Theta}_T \in \tilde{\mathbb{R}}^{Tp \times Tp}$. In the following definition, we reference each diagonal matrix as $\hat{\Theta}^{(t)}$ with the full model being described by $\{\hat{\Theta}^{(t)}\}_{t=1}^T$.

Definition 2.1. *Group-Fused Graphical Lasso estimator*

Let $\hat{S}^{(t)} := x^{(t)}(x^{(t)})^\top$ be the local empirical covariance estimator. The GFGL estimator is defined as the M-estimator

$$\{\hat{\Theta}^{(t)}\}_{t=1}^T = \underset{\{U^{(t)} \succeq 0\}_{t=1}^T}{\textsf{Argmin}} \left\{ l_T((U^{(t)}, x^{(t)})_{t=1}^T) + r_T((U^{(t)})_{t=1}^T) \right\} \tag{2}$$

where the loss-function and regulariser are respectively defined as

$$l_T((U^{(t)}, \hat{S}^{(t)})_{t=1}^T) := \sum_{t=1}^T \left\{ -\log\det(U^{(t)}) + \text{trace}(\hat{S}^{(t)}U^{(t)}) \right\} , \tag{3}$$

$$r_T((U^{(t)})_{t=1}^T) := \lambda_1 \sum_{t=1}^T \sum_{i \neq j} |U_{i,j}^{(t)}| + \lambda_2 \sum_{t=2}^T \{ \sum_{i,j=1}^p (U_{i,j}^{(t)} - U_{i,j}^{(t-1)})^2 \}^{1/2} . \tag{4}$$

Once the precision matrices have been estimated, changepoints are defined as the time-points:

$$\left\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\right\} := \left\{ t \,|\hat{\Theta}^{(t)} - \hat{\Theta}^{(t-1)} \neq 0 \right\} .$$

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 $\tau_{\hat{K}+1}$.

2.2 Comparison with Alternative Estimators

The discussion in this paper studies the impact of implementing *graph-wise* smoothness constraints on precision matrix estimation through penalising edge variation with the group Frobenius norm. This contrasts with previous research that has predominently focussed on enforcing smoothness constraints at an edge-by-edge level [Kolar and Xing, 2011, 2012, Monti et al., 2014, Danaher et al., 2013, Saegusa and Shojaie, 2016]. For instance, one may replace the group-fused penalty in (4) with a linearly seperable norm such as the ℓ_1 norm, i.e. $\lambda_2 \sum_{t=1}^T \|\Theta^{(t)} - \Theta^{(t-1)}\|_1$. In the case of ℓ_1 fusing one may interpret the estimator as a convex relaxation of the total number of jumps across all edges, we may define the number of changepoints as $K_0 := (1/2)\sum_{t=2}^{T-1} \|\Theta_{\setminus ii}^{(t)} - \Theta_{\setminus ii}^{(t-1)}\|_0$, where $\|\cdot\|_0$ counts the number of non-zero differences.

Alternatively, the grouped nature of GFGL assumes that the graph structure underlying a process changes in some sense systematically. Rather than edges changing individually, we presume that many of the edges change dependency structure at the same time. In this case, we consider smoothness at the graph level by counting the number of non-zero differences in a non-seperable norm, such that $K_0 := |\{\|\Theta_{\setminus ii}^{(t)} - \Theta_{\setminus ii}^{(t-1)}\| \neq 0\}|$. In this case, the estimator assumes that many of the active edges may change at the same time and that changes in graphical structure are therefore syncronised.

2.3 Numerical Optimisation

Due to both the convexity and linear composition properties of the cost function (2), optimisation may be performed relatively efficiently. Although not the main focus of this work, we here suggest an *alternating directed method of moments (ADMM)* algorithm which is suitable for the task of minimising the GFGL cost. The algorithm presented here is a modified version of that found in the original GFGL paper [Gibberd and Nelson, 2017], a similar independently developed algorithm that is applicable to this problem can be found in the work of Hallac et al. [2017].

Specifically, we propose to minimise an augmented Lagrangian for (2) where the variables are grouped into sets corresponding to; primal U, auxiliary V, transformations of auxiliary variables W, and dual variables $\{\mathcal{V}_1, \mathcal{V}_2, \mathcal{W}\}$:

$$\mathcal{L}(U, V_1, V_2, W, \mathcal{V}_1, \mathcal{V}_2, \mathcal{W}) := \sum_{t=1}^{T} \Big(- \log\det(U^{(t)}) + \mathrm{tr}(U^{(t)} S^{(t)}) \Big) \dots$$
$$+ \lambda_1 \sum_{t=1}^{T} \| [V_1^{(t)}]_{\setminus ii} \|_1 + \lambda_2 \sum_{t=2}^{T} \| W^{(t)} \|_F + \frac{\gamma_{V_1}}{2} \Big(\sum_{t=1}^{T} \| U^{(t)} - V_1^{(t)} + \mathcal{V}_1^{(t)} \|_F^2 - \| \mathcal{V}_1^{(t)} \|_F^2 \Big) \dots$$
$$+ \frac{\gamma_{V_2}}{2} \Big(\sum_{t=1}^{T-1} \| U^{(t)} - V_2^{(t)} + \mathcal{V}_2^{(t)} \|_F^2 - \| \mathcal{V}_2^{(t)} \|_F^2 \Big) \dots$$
$$+ \frac{\gamma_W}{2} \Big(\sum_{t=2}^{T} \| (V_1^{(t)} - V_2^{(t-1)}) - W^{(t)} + \mathcal{W}^{(t)} \|_F^2 - \| \mathcal{W}^{(t)} \|_F^2 \Big) .$$

The quantities $\gamma_{V_1}, \gamma_{V_2}, \gamma_W$ are algorithmic tuning parameters which assign weight to deviation between the auxilary and primal parameters, for instance via terms like $(\gamma_{V_1}/2)\|U^{(t)} - V_1^{(t)}\|_F$. In practice, the algorithm appears numerically stable for equal weighting $\gamma_{V_1} = \gamma_{V_2} = \gamma_W = 1$. The ADMM algorithm proceeds to perform dual ascent iteratively (and sequentially) minimising $\mathcal{L}\{\cdot\}$ for U, V_1, V_2, W and then updating the dual variables. Pseudocode for these operations are given in Algorithm 1, a more detailed schema can be found in the Supplementary Material.

Algorithm 1. ADMM algorithm for the minimisation of the GFGL cost function.

Input: Data and regulariser parameters: X, λ_1, λ_2, convergence thresholds: $t_p,\ t_d$

Output: Precision matrices: $\{\hat{\Theta}^{(t)}\}_{t=1}^{T}$ and changepoints: $\{\hat{\tau}_1, \dots, \hat{\tau}_{\hat{K}}\}$

Init: $U^{(t)} = V^{(t)} = W^{(t)} = I \in \mathbb{R}^{p \times p}$

while $\epsilon_p > t_p$, $\epsilon_d > t_d$ **do**

- Update Primal Estimate: $U_{n+1}^{(t)} = \arg\min_U \mathcal{L}\{\}$ - Eigen-decomposition;
- Update Auxilary Estimates:
- $V_{1;n+1}^{(t)} = \arg\min_{V_1} \mathcal{L}\{\cdot\}$ - Soft-threshold;
- $V_{2;n+1}^{(t)} = \arg\min_{V_2} \mathcal{L}\{\cdot\}$ - Matrix multiplication;
- $W_{n+1}^{t} = \arg\min_{W} \mathcal{L}\{\cdot\}$ - Group soft-threshold differences;
- Update Dual variables: c.f. $\mathcal{V}_{1;n+1}^{(t)} = \mathcal{V}_{1;n}^{(t)} + U_{n+1}^{t} - V_{1;n}^{(t)}$;

return $\{\hat{\Theta}^{(t)}\}$

In particular, due to seperability of the ADMM updates, this algorithm can be trivially parallelised across time-points and obtains an iteration time complexity of order $\mathcal{O}(p^3T)$ in terms of the number of data-points and dimension p. This contrasts favourably with dynamic programming, which has a naiive cost $\mathcal{O}(T^2)$ [Angelosante and Giannakis, 2011], or $\mathcal{O}(T\log(T))$ in the pruned case [Killick et al., 2012], the binary segmentation scheme of Leonardi and Bühlmann [2016] also obtains this latter rate. The cubic cost with regards to dimension is due to the eigen-decomposition step. Potentially, this may be reduced if some form of covariance pre-screeneing can be achieved, for instance as suggested in Danaher et al. [2013] it may be possible to break the p dimensional problem into a set of independent $p_1, \ldots, p_M$ dimensional problems where $\sum_{m=1}^{M} p_m = p$. We leave such investigations as directions for further work.

2.4 Optimality conditions

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 obey. We connect this condition to the sampling of the model under (1) by the quantity $\Psi^{(t)} := \hat{S}^{(t)} - \Sigma^{(t)}$ 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 convinient to introduce a matrix $\Gamma^{(t)}$ corresponding to the differences in precision matrices. For the first step let $\Gamma^{(1)} = \Theta^{(1)}$, then for $t = 2, \ldots, T$, let $\Gamma^{(t)} = \Theta^{(t)} - \Theta^{(t-1)}$. 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

$$\hat{R}_{1;(i,j)}^{(t)} = \begin{cases} \text{sign}(\sum_{s\leq t}\Gamma_{i,j}^{(s)}) & \text{if } \sum_{s\leq t}\Gamma_{i,j}^{(s)} \neq 0 \\ [-1,1] & \text{otherwise} \end{cases} ; \; \hat{R}_2^{(t)} = \begin{cases} \frac{\hat{\Gamma}^{(t)}}{\|\hat{\Gamma}^{(t)}\|_F} & \text{if } \hat{\Gamma}^{(t)} \neq 0 \\ \mathcal{B}_F(0,1) & \text{otherwise} \end{cases},$$

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

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

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

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

3 Theoretical Properties of the GFGL Estimator

3.1 Consistent Changepoint Estimation

We here present two complimentary results for changepoint consistency with the GFGL estimator. The first represents a standard asymptotic setting where p is fixed and $T \to \infty$. The second result utilises a different concentration bound to constrain

changepoint estimation error even in high-dimensions where $p > T$, in the doubly asymptotic setting $(p, T) \to \infty$. Further to the Gaussian sampling model in (1), the high-dimensional result holds for any collection $(X^{(1)}, \ldots, X^{(T)})$ where each $X_i^{(t)}/(\Sigma_{ii}^{(t)})^{1/2}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, p$ is sub-Gaussian with parameter σ.

The changepoint consistency result presented here will take the form of an upper bound on the maximum error of estimating a changepoint. 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 $d_{\min} := \min_{k \in [K+1]} |\tau_k - \tau_{k-1}|$ increases with T, for simplicity let us assume this is bounded as a proportion $\gamma_{\min} < d_{\min}/T$.

In order to control the variance in the underlying model it is required to introduce several assumptions on the generating process:

Assumption 1. *Bounded Eigenvalues*

There exist a constant $\phi_{\max} < \infty$ that gives the maximum eigenvalues of the true covariance matrix (across all blocks) such that $\phi_{\max} = \max_{k \in [K+1]}\{\Lambda_{\max}(\Sigma^{(k)})\}$.

Assumption 2. *Bounded jump size*

There exists a constant $M > 0$ such that the difference between any two blocks is bounded by a constant $\max_{k,k' \in [K+1]} \|\Sigma^{(k')} - \Sigma^{(k)}\|_F \leq M$. Additionally, the jumps are lower bounded according to $\min_{k \in [K+1]} \|\Sigma^{k} - \Sigma^{k-1}\|_F \geq \eta_{\min}$.

In addition to controlling variation in the sampling, we also need to ensure the regularisers λ_1, λ_2 are set appropriately. Generally one can increase or decrease the smoothing regulariser λ_2 to respectively decrease or increase the number of estimated changepoints However, in practice, we do not know a-priori the true number of changepoints in the model. Pragmatically, we may adopt a cross-validation strategy to choose the regularisers from data. However, in the theoretical setting, we assume that the regularisers are such that the correct number of changepoints are estimated.

Assumption 3. *Appropriate Regularisation*

Let λ_1 and λ_2 be specified such the solution of the GFGL problem (2) results in $|\hat{K}| = K$ changepoints. Furthermore, assume that T is large enough such that $\beta_1 := (\eta_{\min}\gamma_{\min}T)\lambda_2^{-1} > 32$, $\beta_2 := \eta_{\min}\lambda_1^{-1}(p(p-1))^{-1/2} > 8$, and $\beta_3 := (\eta_{\min}T\delta_T)\lambda_2^{-1} > 3$.

Theorem 3.1. *Changepoint Consistency (Standard-dimensional asymptotics)*

Given Assumptions 1,2,3. For a finite, but large T such that $2 < p \leq T\delta_T$ and the jump size is lower bounded according to $\eta_{\min} > 20\phi_{\max}e^{-1/2}p^{1/2}\sqrt{\log(T\delta_T)/T\delta_T}$, then the maximum changepoint error of GFGL (under sampling Eq. 1) is bounded according to probability

$$P(\max_{k \in [K]} |\tau_k - \hat{\tau}_k| \leq T\delta_T) \geq 1 - C_K 2(T\delta_T)^{-p/2} \to 1 \quad \text{as} \quad T \to \infty \,,$$

where $C_K = K(K^2 2^{K+1} + 4)$.

Theorem 3.2. *Changepoint-consistency (High-dimensional asymptotics)*

Given Assumptions 1,2,3. If $\eta_{\min} > 5c_\sigma 2^5 p\sqrt{\log(p^{\beta/2})/T\delta_T}$*, where* $c_\sigma = (1 + 4\sigma^2)\max_{ii,k}\{\Sigma^{(k)}_{0;ii}\}$ *and* $\eta_{\min} \in (0, 2^4 c_\sigma p)$*, then*

$$P(\max_{k\in[K]} |\tau_k - \hat{\tau}_k| \leq T\delta_T) \geq 1 - C_K 1/p^{\beta-2} \to 1 \quad \text{as} \quad (T,p) \to \infty\,.$$

Proof. The proofs of the above follow 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. Additionally, for Theorem 3.2 we investigate the use of concentration bounds that operate in the setting where $p > T\delta_T$. Full details can be found in the Appendix (ref). □

The results demonstrate that asymptotically changepoint error can be constrained to a decreasing fraction $\delta_T \to 0$ of the time-series. Indeed, this occurs with increasing probability when estimation is performed with an increasing number of data-points. In both results, the minimal jump size must be sufficient in order to consistently detect changes. In the standard-dimensional setting we require $\eta_{\min} = \Omega(p^{1/2}\sqrt{\log(T\delta_T)/(T\delta_T)})$, and therefore asymptotically we can recover increasingly small changes. In the doubly asymptotic setting of Theorem 3.2 we have $\eta_{\min} = \Omega(p\sqrt{\log(p^{\beta/2})/(T\delta_T)})$, thus we still require $\eta_{\min}$ to grow with p. One should note that the high-dimensional bound is only guaranteed for $\eta_{\min} \in (0, 2^4 c_\sigma p)$ which stipulates a minimal sample size $T\delta_T = \Omega(\log(p^{\beta/2}))$.

3.2 Consistent Graph Recovery

One of the key properties of GFGL is that it simultaneously estimates both the changepoint and model 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 3 and Theorem 3.2. In the previous section we developed theory for the recovery of changepoints, for a given (large) number of data-points we demonstrated that the changepoint error can be bounded within a region $T\delta_T \to 0$. We here assume that such a bound holds, and develop the theory in the high-dimensional setting analogous to Theorem 3.2.

Assumption 4. *Changepoint error is small*

Consider that the number of changepoints is estimated correctly, and the changepoint consistency event

$$E_\tau := \left\{ \max_{k\in[B]} |\hat{\tau}_k - \tau_k| \leq T\delta_T \right\}$$

holds with some increasing probability $f_\tau(p,T) \to 1$ *as* $(p,T) \to \infty$*.*

A key advantage of splitting the changepoint and model-estiamtion 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 $K+1$ block-diagonal matrix $\Theta_K \in \tilde{\mathbb{R}}^{Bp \times Bp}$. Conditional on segmentation, we do not need to deal with the fact that the model may be arbitrarily mis-specified (this is bounded by Assumption 4). 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.

To demonstrate model recovery, let us define a set of pairs $\mathcal{M}_k$ which indicate the support set of the true model in block k and its compliment $\mathcal{M}_k^{\perp}$ as

$$\mathcal{M}_k = \{(i,j) \mid \Theta^{(k)}_{0;i,j} \neq 0\} \quad \text{and} \quad \mathcal{M}_k^{\perp} = \{(i,j) \mid \Theta^{(k)}_{0;i,j} = 0\} \, .$$

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:

$$E_{\mathcal{M}}(\hat{\Theta}^{(k)}; \Theta^{(l)}_0) := \{\text{sign}(\hat{\Theta}^{(k)}_{ij}) = \text{sign}(\Theta^{(l)}_{0;ij}) \; \forall i,j \in \mathcal{M}_l\} \, .$$

In order to derive bounds on model recovery, one must make assumptions on the true structure of Θ_0. Such conditions are often 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 $Z_{(i,j)} = X_i X_j - \mathbb{E}[X_i X_j]$ for $i,j \in \{1,\ldots,p\}$.

Assumption 5. *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 compliment. 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 $k = 1,\ldots,K+1$ there exists some $\alpha_k \in (0,1]$ such that

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

In the multivariate Gaussian case we have $\max_{e \in \mathcal{M}^{\perp}} \|\mathbb{E}[Z_e Z_{\mathcal{M}}^{\top}]\mathbb{E}[Z_{\mathcal{M}} Z_{\mathcal{M}}^{\top}]^{-1}\|_1 \leq (1-\alpha_k)$ for each k, in the theory below we denote and track $\alpha = \min_k\{\alpha_k\}$. One can therefore interpret the incoherence condition as a statement on the correlation between edge variables which are outside the model subspace $Z_{(i,j)}$ such that $(i,j) \notin \mathcal{E}$, with those contained in the true model $(i,j) \in \mathcal{E}$. 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 Sec. 3.1.1 Ravikumar et al. [2011] and Meinshausen [2008]). The theory 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 size of the operator norms $\left|\left|\left|\Sigma^{(k)}_0\right|\right|\right|_{\infty} := \max_i \sum_{j=1}^{p} |\Sigma^{(k)}_{0;ij}|$ and $\left|\left|\left|\Gamma^{(k)}_0\right|\right|\right|_{\infty}$,

we simplify our analysis by tracking the upper bound $K_{\Sigma_0} := \max_k \left|\left|\left|\Sigma_0^{(k)}\right|\right|\right|_\infty$ and $K_{\Gamma_0} := \max_k \left|\left|\left|\Gamma_0^{(k)}\right|\right|\right|_\infty$.

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 proof 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 3.3. *Bounds on the estimation error*

Consider the GFGL estimator with Assumptions (4,5). Assume $\lambda_1 = 16\alpha^{-1}\epsilon$, $\lambda_2 = \rho\lambda_1$ *for some finite* $\rho > 0$ *and* $\epsilon > 2^4\sqrt{2}c_{\sigma_\infty}\sqrt{\log(4p^\beta)/\gamma_{\min}T}$ *where* $\beta > 2$. *Let* d *be the maximum degree of each node in the true model, and define*

$$v_{\mathcal{C}} = 6\{1 + 16\alpha^{-1}(1 + 2\hat{n}_k^{-1}\rho)\}d\max\{K_{\Sigma_0}K_{\Gamma_0}, K_{\Gamma_0}^2K_{\Sigma_0}^3\} \; . \tag{5}$$

Then for $T \geq 2^9\gamma_{\min}^{-1}\max\{1/8c_{\sigma_\infty}, v_{\mathcal{C}}\}^2c_{\sigma_\infty}^2\log(4p^\beta)$, *we have*

$$\|\hat{\Theta}^{(k)} - \Theta_0^{(k_{\max})}\|_\infty \leq 2K_{\Gamma_0}\{1 + 16\alpha^{-1}(1 + 2\rho\hat{n}_k^{-1})\}\epsilon \; , \tag{6}$$

in probability greater than $1 - (1/p^{\beta-2} + f_\tau(p, T))$.

Theorem 3.4. *Model-selection consistency*

In addition to the assumptions in Theorem 3.3. Let $\theta_{\min}^{(k)} := \min_{ij}|\Theta_{0;ij}^{(k)}|$ *for all* $(i, j) \in \mathcal{M}_k$ *and for each* $k = 1, \ldots, B$. *Let*

$$v_\theta = 2K_{\Sigma_0}\{1 + 16\alpha^{-1}(1 + 2\rho\hat{n}_k^{-1})\}\theta_{\min}^{-1} \; .$$

If $T \geq 2^9\gamma_{\min}^{-1}\max\{(8c_{\sigma_\infty})^{-1}, v_{\mathcal{C}}, v_\theta\}^2c_{\sigma_\infty}^2\log(4p^\beta)$ *then GFGL attains sign-consistency*

$$P\{E_{\mathcal{M}}(\hat{\Theta}^{(k)}; \Theta_0^{(k_{\max})})\} \geq 1 - \{1/p^{\beta-2} + f_\tau(p, T)\} \; ,$$

with probability tending to one as $(p, T) \to \infty$.

Proof. Theorem 3.3 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. Theorem 3.4, is a corollary of Theorem 3.3 on the condition that the true entries in the precision matrix are sufficiently large. Full details of the proof are found in Appendix . □

The above bounds suggest that indeed, if regularisation is appropriately set one can obtain a consistent estimate of the precision matrices from GFGL. However, there are several important insights we can take from the results. Firstly, we clearly see the effect of the smoothing regulariser in Eq. 6. In particular, a larger ρ will result in

a larger upper bound for the error. In the analagous results from the i.i.d. graphical lasso case Ravikumar et al. [2011], the bound here 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 $\lambda_2(\lambda_1\hat{n}_k)^{-1}$, if λ_2/λ_1 does not grow faster than $\hat{n}$ then estimation consistency can be achieved. If we assume, for example that $\lambda_1 = \mathcal{O}(\sqrt{\log(p^\beta)/T})$ and $\hat{n}_k = \mathcal{O}(T)$, then this gives λ_2 the flexibility to grow with increasing T, but at a rate $\lambda_2 = \mathcal{O}(\sqrt{T\log p^\beta})$. We note that under such scaling it is possible to satisfy the conditions of Assumption 3 thus both changepoint and structure estimation consistency is achieved.

4 Discussion

4.1 A note on minimum detectable jumps

When compared to the neighbourhood based estimators, GFGL considers changepoints at the full precision matrix scale as opposed to seperately for each node. One might therefore expect that the minimum jump size required in our result (Theorem 1) is greater than that utilised in neighbourhood selection case (c.f. Kolar and Xing [2012]). For example, in the neighbourhood selection case, one may consider the analogous quantity $\eta^{\text{NS}}_{\text{min}}(a) := \min_{k\in[K]} \|\Sigma^{(k+1)}_{a,\cdot} - \Sigma^{(k)}_{a,\cdot}\|_2$, for nodes $a = 1,\ldots,p$. Summing over nodes, the neighbourhood selection jump size can now be related to the mixed (group) norm $\|\Sigma^{(k+1)} - \Sigma^{(k)}\|_{2,1} = \sum_a \|\Sigma^{(k+1)}_{a,\cdot} - \Sigma^{(k)}_{a,\cdot}\|_2$. Furthermore, if the smallest jump occurs at the same block for each neighbourhood, i.e. $\arg\min_{k\in[K]} \|\Sigma^{(k+1)}_{a,\cdot} - \Sigma^{(k)}_{a,\cdot}\|_2 = \arg\min_{k\in[K]} \|\Sigma^{(k+1)}_{a',\cdot} - \Sigma^{(k)}_{a',\cdot}\|_2$ for all $a \neq a' \in [p]$, then $\sum_a \eta^{\text{NS}}_{\text{min}}(a) = \min_k \|\Sigma^{(k+1)} - \Sigma^{(k)}\|_{2,1}$. Using the inequality (for $x \in \mathbb{R}^n$) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$, the jumps as measured through the group-norm can be related to those measured in a Frobenius sense, such that $\eta_{\text{min}} \leq \sum_a \eta^{\text{NS}}_{\text{min}}(a) \leq \sqrt{p}\eta_{\text{min}}$.

Thus, even though the minimum jump size in the GFGL case is greater, i.e. $\eta_{\text{min}} > \eta^{\text{NS}}_{\text{min}}(a)$, 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 trade-off between the minimum jump-size η_{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 $\eta_{\text{min}} \leq \sum_{a=1}^{p} \eta^{\text{NS}}_{\text{min}}(a)$ 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 η_{min} and $\eta^{\text{NS}}_{\text{min}}$ 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, although this is generally not the case with the block-constant GGM model (1).

4.2 Consistency with increasing dimensionality

The work of Kolar and Xing [2012] presents results in a standard asymptotic $T \to \infty$ setting. As such, they do not assess how fast changepoint consistency is achieved in relation to the number of data-streams p. In the case of GFGL, we give an upper bound on $P(E_\tau)$ (Theorem 3.1) that directly relates to the dimensionality of the problem. Of particular note, is that the convergence rate increases with the number of data-streams used. This aligns with what one may intuitively expect, i.e. if changepoints are shared across most data-streams, then increasing the number of data-streams will increase the "signal" associated with changepoints. We may thus improve changepoint detection performance by performing joint inference for changepoints.

In our high-dimensional analysis (Theorems 3.2, 3.3,3.4) we consider what happens when $p > T\delta_T$. Our upper bound for changepoint error has a requirement that $\eta_{\min} = \Omega(p^{-1}\sqrt{\log(p^\beta/2)/T\delta_T})$. We note that this lower bound allows for the calculation of a minimal T for the bound to hold. However, one must also take care that conditions in Assumption 3 are met. While λ_1 can be seen somewhat as a free parameter in the proof of changepoint consistency, it plays a key role in bounding the estimation error of the block-wise precision matrices. Specifically, Theorem 3.3 mandates $\lambda_1 = \Omega(\sqrt{\log(p^\beta)/T})$. We note, that this lower bound is generally sufficient to meet Assumption 3, for changepoint consistency, which requires $\lambda_1 = \mathcal{O}(\eta_{\min}/p) = \mathcal{O}(\sqrt{\log(p^(\beta/2))/T\delta_T})$ when $\eta_{\min}$ is guided by the lower bound in Theorem 3.2.

A Numerical Optimisation and optimality conditions

A.1 Multi-block ADMM algorithm for GFGL

Algorithm 2 provides extended detail relating to the pseudo-code presented in the main paper. The particular steps for each of the proximity updates (minimisation steps in the augmentd lagrangian) are given. An alternative ADMM algorithm for solving the GFGL problem is described in Hallac et al. [2017] and Gibberd and Nelson [2017].

A.2 Proof of Proposition 1 (Optimality Conditions)

In GFGL we have a set of conditions for each time-point which must be met jointly. Unlike non-fused estimators, we also have to consider the stationarity conditions due to a differenced term. The GFGL objective can then be re-written in terms of this difference, where one may equivalently minimise

$$\sum_{t=1}^{T}\Bigg(-\log\det(\sum_{s\leq t}\Gamma^{(s)})+\mathrm{tr}(\hat{S}^{(t)}\sum_{s\leq t}\Gamma^{(s)})\Bigg)+\lambda_1\sum_{t=1}^{T}\|\sum_{s\leq t}\Gamma^{(s)}_{\setminus ii}\|_1+\lambda_2\sum_{t=2}^{T}\|\Gamma^{(t)}\|_F\ ,$$

for $\sum_{s\leq t}\Gamma^{(s)} \succ 0$ for $t = 1,\ldots,T$. Setting the derivative to zero we obtain:

$$0=\sum_{t=l}^{T}\Bigg(-(\sum_{s\leq t}\hat{\Gamma}^{(s)})^{-1}+\hat{S}^{(t)}\Bigg)+\lambda_1\sum_{t=l}^{T}\hat{R}_1^{(t)}+\lambda_2\hat{R}_2^{(l)}\ .$$

Algorithm 2. ADMM algorithm for the minimisation of the GFGL cost function.

Input: Data and regulariser parameters: $\boldsymbol{X}, \lambda, \rho$, convergence thresholds: $t_p,\ t_d$

Output: Precision matrices: $\{\hat{\boldsymbol{\Theta}}^{(t)}\}_{t=1}^{T}$ and changepoints: $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\}$

Init: $U^{(t)} = V^{(t)} = W^{(t)} = I \in \mathbb{R}^{p \times p}$

while $\epsilon_p > t_p$, $\epsilon_d > t_d$ **do**

  for *t=1,...,T* **do**

    $$P^{(t)} = \begin{cases} V_{1;n}^{(t)} - \mathcal{V}_{1;n}^{(t)} & t = T \\ (V_{1;n}^{(t)} - \mathcal{V}_{1;n}^{(t)}) + (V_{2;n}^{(t)} - \mathcal{V}_{2;n}^{(t)}) & \text{otherwise} \end{cases};$$

    Perform eigen-decomposition: $\{\eta_n^{\top}, L\} = \text{eig}\{(\hat{S}^{(t)} - P^{(t)})\}$;

    Update eigenvalues, for each $i = 1, \ldots, p$:

    $\eta_{i;n+1} = -(\eta_{i;n} - \sqrt{\eta_{i;n}^2 + 8})/4$;

    $U_{n+1}^{(t)} = L\text{diag}(\eta_{n+1})L^{\top}$;

  Update Auxiliary Estimate:

  $V_{1;n+1}^{(1)} = \text{soft}([U_{n+1}^{(1)} + \mathcal{V}_{1;n}^{(1)}]_{\setminus\setminus ii};\ \lambda)$;

  for $t = 2, \ldots, T$ **do**

    $P^{t} = [(U_{n+1}^{(t)} + \mathcal{V}_{1;n}^{(t)}) + (V_2^{(t-1)} + W_n^{(t-1)} - \mathcal{W}_n^{(t-1)})]/2$;

    $V_{1;n+1}^{(t)} = \text{soft}([P]_{\setminus\setminus ii};\ \lambda)$;

    $V_{2;n+1}^{(t-1)} = (1/2)[(U_{n+1}^{(t-1)} + \mathcal{V}_n^{(t-1)}) + (V_{1;n}^{(t)} - W_n^{(t-1)} + \mathcal{W}_n^{(t-1)})]$;

    Threshold for jumps:

    $Q^{(t)} = V_{1;n+1}^{(t)} - V_{2;n+1}^{(t-1)} + \mathcal{W}_n^{(t-1)}$;

    $W_{n+1}^{t-1} = (Q^{(t)}/\|Q^{(t)}\|_F)\max(\|Q^{(t)}\|_F - \lambda\rho, 0)$;

  Update Dual variables:

  $\mathcal{V}_{1;n+1}^{(t)} = \mathcal{V}_{1;n}^{(t)} + U_{n+1}^{t} - V_{1;n}^{(t)}$;

  $\mathcal{V}_{2;n+1}^{(t)} = \mathcal{V}_{2;n}^{(t)} + U_{n+1}^{t} - V_{2;n}^{(t-1)}$;

  $\mathcal{W}_{n+1}^{(t)} = \mathcal{W}_n^{(t)} + V_{1;n+1}^{(t)} - V_{2;n}^{(t-1)} - W_{n+1}^{(t)}$;

return $\{\hat{\Theta}^{(t)}\}$

The above derivative, is linked through the data via the function $\hat{S}^{(t)} = X^{(t)}(X^{(t)})^{\top}$. Recalling that $X^{(t)} \sim \mathcal{N}_p(0, \Sigma^{(t)})$, we can then link $\Sigma^{(t)}$ to $\Gamma^{(l)}$ via $\Sigma^{(t)} = (\sum_{s \leq t} \Gamma^{(s)})^{-1}$.

We now have a way to relate the estimated precision matrices $\hat{\Gamma}^{(t)}$ and the corresponding ground-truth. Let us write the sampling error for the covariance matrix at time point t as $\Psi^{(t)} := \hat{S}^{(t)} - \Sigma^{(t)}$. Substituting $\Psi^{(t)}$ into the above stationarity conditions for GFGL we obtain;

$$\sum_{t=l}^{T} \left((\sum_{s \leq t} \Gamma^{(s)})^{-1} - (\sum_{s \leq t} \hat{\Gamma}^{(s)})^{-1} \right) - \sum_{t=l}^{T} \Psi^{(t)} + \lambda_1 \sum_{t=l}^{T} \hat{R}_1^{(t)} + \lambda_2 \hat{R}_2^{(l)} \ ,$$

and thus equivalently obtaining the result in Proposition 1. □

B 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

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

The compliment of the event on the LHS is equivalent to the target of proof; we wish to demonstrate $P[\max_{k\in[K]} |\tau_k - \hat{\tau}_k| \leq T\delta_T] \to 1$. 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 complimentary 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:

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

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 $T \to 0$.

B.1 Stationarity induced bounds

As a first step, let us introduce some bounds based on the optimality conditions which occur in probability one. From here, a set of events can be constructed that occur when the stationarity conditions are met. 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 stationarity equations (Prop. 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

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

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

$$\| \sum_{t=\hat{\tau}_k}^{\tau_k - 1} (\Sigma_0^{(t)} - \hat{\Sigma}^{(t)}) \|_F - \| \sum_{t=\hat{\tau}_k}^{\tau_k - 1} \Psi^{(t)} \|_F \leq 2\lambda_2 + \lambda_1 \sqrt{p(p-1)}(\tau_k - \hat{\tau}_k) \ . \tag{10}$$

The next step is to replace the time indexed inverse precision matrices $\Theta^{(t)}$ with the block-covariance matrices indexed $\Sigma^{(k)}$ and $\Sigma^{(k+1)}$. We can re-express the difference in precision matrices as the sum of a difference between true values before τ_k , i.e. $\Sigma^{(k+1)} - \Sigma^{(k)}$, and the difference between the next $(k+1)$st true block and estimated block, i.e. $\hat{\Sigma}^{(k+1)} - \Sigma^{(k+1)}$ to obtain:

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

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. $P[E_1 \cup E_2 \cup E_3] = 1$.

B.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. Similarly 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 $\|R_1\|_F = \| \sum_{t=\hat{\tau}_k}^{\tau_k - 1} \Sigma^{(k)} - \Sigma^{(k+1)} \|_F \geq (\tau_k - \hat{\tau}_k)\eta_{\min}$ 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 3. 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

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

and thus

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

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 $\|R_2\|_F \leq (\tau_k - \hat{\tau}_k)\|\hat{\Sigma}^{k+1} - \Sigma^{k+1}\|_F$. Substituting in (13) we have

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

Splitting the probability into three components, we obtain

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

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 $\eta_{\min} \leq 3\| \sum_{t=\tau_k}^{\bar{\tau}_k - 1} \Psi^{(t)} \|_F / (\bar{\tau}_k - \tau_k)$ 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 $\Sigma^{(t)} = \Sigma^{(k)}$ for all $t \in [\tau_k, \tau_{k+1}]$. One can now write the sampling error across time into one which relates to blocks k as

$$\| \sum_{t=\tau_k}^{\bar{\tau}_k - 1} \Psi^{(t)} \|_F \equiv (\bar{\tau}_k - \tau_k) \| W_{k;\bar{\tau}_k - \tau_k} \|_F \ ,$$

where $W_{k;n} = [n^{-1} \sum_{t=1}^{n} X^{(t)} (X^{(t)})^\top] - \Sigma_0^{(k)}$. A control on this quantity is given in the following lemma:

Lemma 1. *Sample Error Bound in High-Dimensions*

Let $\hat{W}_k^{(n)} = [n^{-1} \sum_{t=1}^{n} X^{(t)} (X^{(t)})^\top] - \Sigma_0^{(k)}$, then for any $\epsilon \in (0, 2^3 c_\sigma p)$, with sub-Gaussian noise $\{X^{(t)}\}_{t=1}^{T}$, the error is bounded according to

$$P(\|\hat{W}_{k;n}\|_F > \epsilon) \leq 4p^2 \exp\left(-\frac{n\epsilon^2}{2^7 c_\sigma^2 p^2} \right) \ ,$$

where $c_\sigma = (1+4\sigma^2)\max_{ii}\{\Sigma^{(k)}_{0;ii}\}$. *Furthermore, if* $\epsilon > \epsilon^{\alpha}_{\text{conv}} := c_\sigma 2^4\sqrt{2}p\sqrt{\log(p^{\beta/2})/n}$ *and* $\beta > 2$ *then* $P(\|\hat{W}_{k;n}\|_F > \epsilon) \leq p^{(2-\beta)} \to 0$ *as* $p \to \infty$.

Proof. See section B.5. □

In the specific setting of $A_{T,k,2}$ the probability we desire to bound is $P[\|W_{k,\bar{\tau}_k-\tau_k}\|_F > \eta_{\min}/3]$, applying Lemma 1 gives convergence if $\eta_{\min} > 15 \times 2^5 p\sqrt{\log(p^{\beta/2})/d_{\min}}$, where we note $\bar{\tau}_k - \tau_k > d_{\min}/2$ and for Gaussian sampling $c_\sigma = 5$. Finally, to ensure that the bound in Lemma 1 holds, we must ensure that $\eta_{\min}/3 \in (0, 2^3 c_\sigma/p)$. If $d_{\min} = \gamma_{\min}T$ and $\eta_{\min}$ follows the form $cp\sqrt{\log(p^{\beta/2}/\gamma_{\min}T)}$ for some constant c, then we require $T > \gamma^{-1}_{\min}(c/c_\sigma)^2 2^{-6}\log(p^{\beta/2})$.

Alternatively, we may study the sampling error in the standard asymptotic setting where $\bar{\tau}_k - \tau_k > p$.

Lemma 2. *Sample Error Bound in Standard Dimensions*

Let $p \leq n$ *and* $X^{(t)} \sim \mathcal{N}(0, \Sigma^{(k)}_0)$ *for* $t = 1, \ldots, n$. *If the covariance matrix* $\Sigma^{(k)}_0$ *has maximum eigenvalues* $\phi_{\max} < +\infty$, *then for all* $t > 0$ *and* $p > 2$ *we have*

$$P\left(\|\hat{W}_{k;n}\|_F \geq 4\phi_{\max}e^{-1/2}\sqrt{p\log n/n}\right) < 2n^{-p/2}\ . \tag{16}$$

Proof. The result is a corollary of bounds derived in [Wainwright, 2009], see B.6. □

In the context of $A_{T,k,2}$, Lemma 2 holds with $\eta_{\min} \geq 12\phi_{\max}e^{-1/2}\sqrt{p\log n/n}$ and $n = \gamma_{\min}T$. While the probability bound in the Lemma is not sharp, it does relate the convergence rate to the dimensionality p, which in this context is fixed as $T \to \infty$.

Finally, let us turn to $A_{T,k,3}$. Recall $P(A_{T,k,3}) := P(A_{T,k} \cap C_T \cap E_3) := P(A_{T,k} \cap C_T \cap \{\|\sum_{t=\hat{\tau}_k}^{\tau_k-1}\Psi^{(t)}\|_F \geq \|R_1\|_F/3\})$. 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 (Lemmas 1, 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 $n > T\delta_T$.

B.3 Bounding the Bad Cases

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

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

where $C^c_T = \{\max_{k\in[K]}|\hat{\tau}_k - \tau_k| \geq d_{\min}/2\}$ is the compliment of the good event. The events above correspond to estimating a changepoint; a) before the previous true changepoint ($D^{(l)}_T$); b) between the previous and next true changepoint ($D^{(m)}_T$), and c) after the next true changepoint ($D^{(r)}_T$). The events $D^{(l)}_T$ and $D^{(r)}_T$ 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.

Lemma 3. *Upper bound for* $P[A_{T,k} \cap D_T^{(m)}]$

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

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

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

In particular, one can demonstrate that:

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

Proof. The result follows from expanding events based on neighbouring changepoints (see Appendix B.7 for detail). □

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 $l = \hat{\tau}$, $l = \tau_k$ and $l = \hat{\tau}_k, l = \tau_{k+1}$. We respectively obtain:

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

and

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

The next step is to define an event that can bound $P(A_{T,k} \cap E_k^{'} \cap D_T^{(m)})$. Using the triangle inequality we bound $\|\Sigma^{(k+1)} - \Sigma^{(k)}\|_F$ 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

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

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

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

The stochastic error terms (containing $W_{k;\tau_k - t\hat{a}u_k}$) can then be shown to converge similarly to $P(A_{T,k} \cap C_T)$ c.f. Eq. (14). 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:

Lemma 4. *The probability of $D_T^{(l)}$ is bounded by*

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

Proof. This is based on a combinatorial argument for the events that can be considered on addition of each estimated changepoint. For details see Appendix B.8. □

In order to bound the above probabilities we relate the events $E_l^{''}$ and $E_l^{'}$ to the stationarity conditions as before (via Eqs. 20, 21). Setting $k = l$ and invoking the triangle inequality gives us

$$\begin{aligned} \{\|\Sigma_{l+1} - \Sigma_l\|_F &\leq 2\lambda_1\sqrt{p(p-1)} + \overbrace{2\lambda_2(|\tau_l - \hat{\tau}_l|^{-1} + |\hat{\tau}_{l+1} - \tau_l|^{-1})}^{A} \\ &\quad + \|W_{l;\tau_l - \hat{\tau}_l}\|_F + \|W_{l+1;\hat{\tau}_{l+1} - \tau_l}\|_F\} \ . \end{aligned}$$

Conditioning on the event $E_l^{''} \cap E_l^{'}$ implies that $A = 8\lambda_2/d_{\min}$. We can thus write

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

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

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

B.4 Summary

The bounds derived above 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 $k_\infty = \arg\max_k \{\max_{ii} \Sigma_{ii}^{(k)}\}$ select the block which results in the largest expected covariance error. Summing the probabilities, one obtains the upper bound:

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

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

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

where $C_K = K(K^2 2^{K+1} + 4)$. We thus arrive at the result of Theorem 3.1.

□

B.5 Lemma 1. High-Dimensional Bound on Empirical Covariance Error

Lemma. *Let $W_k^{(n)} = [n^{-1} \sum_{t=1}^n X^{(t)} (X^{(t)})^\top] - \Sigma_0^{(k)}$, then for any $\epsilon \in (0, 2^3 c_\sigma/p)$, with Gaussian noise $X^{(t)} \sim \mathcal{N}(0, \Sigma^{(k)})$, the error is bounded according to*

$$P(\|\hat{W}_{k;n}\|_F > \epsilon) \leq 4p^2 \exp\left(-\frac{n\epsilon^2}{2^7 c_\sigma^2 p^2}\right) \ ,$$

where $c_\sigma = (1+4\sigma^2) \max_{ii} \{\Sigma_{0;ii}^{(k)}\}$. Furthermore, if $\epsilon > \epsilon_{\text{conv}}^\alpha := c_\sigma 2^4 \sqrt{2} p \sqrt{\log(p^{\alpha/2})/n}$ and $\alpha > 2$ then

$$P(\|\hat{W}_{k;n}\|_F > \epsilon) \leq p^{(2-\alpha)} \to 0 \ .$$

Proof. From Ravikumar et al. [2011] Lemma 1, for $X^{(t)}$ with sub-Gaussian tails (parameter σ) we have

$$P(|\hat{W}_{k;n;i,j}| > \delta) \leq 4 \exp\left(-\frac{n\delta^2}{2^7 c^2}\right) ,$$

where $c = (1+4\sigma^2)\max_{ii}\{\Sigma_{0;ii}\}$ for all $\delta \in (0, 2^3 c)$. Take the union bound to obtain max norm

$$P(\|\hat{W}_{k;n}\|_\infty > \delta) \leq 4p^2 \exp\left(-\frac{n\delta^2}{2^7 c^2}\right) .$$

Now use the fact that $p^2\|X\|_\infty \geq p\|X\|_F$ to control the event relating to Frobenius norm. Note that if $P(\|X\|_\infty > \delta/p) = A$ and $P(\|X\|_F > \delta) = B$, then $A \geq B$ and hence we can use the larger probability A to bound the Frobenius event

$$P(\|\hat{W}_{k;n}\|_F) > \delta) \leq 4p^2 \exp\left(-\frac{n\delta^2}{p^2 2^7 c^2}\right)$$

The bound converges to zero for all $\epsilon > \epsilon_{\text{conv}}$ where

$$\epsilon^2_{\text{conv}} = \frac{p^2 2^2 2^7 c^2}{n} \log(4p^2)$$

giving $\epsilon_{\text{conv}} := c2^4\sqrt{2}p\sqrt{\log(p)/n}$.

□

B.6 Proof of Lemma . Standard-dimensional Bounds for Empirical Covariance Error

Lemma. *Concentration of spectral and Frobenius norm*

Let $p \leq n$ and $X^{(t)} \sim \mathcal{N}(0, \Sigma_0^{(k)})$ for $t = 1, \ldots, n$. If the covariance matrix Σ has maximum eigenvalues $\phi_{\max} < +\infty$, then for all $a > 0$

$$P(\left|\left|\left|\hat{W}_{k;n}\right|\right|\right|_2 \geq \phi_{\max}\delta(n,p,a)) \leq 2\exp(-na^2/2) , \tag{24}$$

where $\delta(n,p,a) := 2((p/n)^{1/2}+a)+((p/n)^{1/2}+a)^2$. Furthermore, with $\delta(n,p,\sqrt{p\log(n)/n})$ and $p > 2$ we have

$$P\left(\|\hat{W}_{k;n}\|_F \geq \frac{4\phi_{\max}}{\sqrt{e}}\sqrt{\frac{p\log n}{n}}\right) < 2n^{-p/2} . \tag{25}$$

Proof. The proof of Eq. 24 is given in Lemma 9 [Wainwright, 2009] and is based on a result for Gaussian ensembles from Davidson and Szarek [2001]. For the specific bound in (25) set $a = \sqrt{p\log(n)/n}$ such that

$$\begin{aligned}\delta(n,p,\sqrt{p\log(n)/n}) &= \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{p\log(n)}{n}}\right)\left(2 + \sqrt{\frac{p}{n}} + \sqrt{\frac{p\log(n)}{n}}\right) \\ &< \sqrt{\frac{p\log(n)}{n}}(2 + 2\sqrt{\frac{p\log(n)}{n}}) .\end{aligned}$$

Noting that $\sqrt{p\log(n)/n}$ is maximised for $n = e$, we obtain

$$\begin{aligned}\delta(n, p, \sqrt{p\log(n)/n}) &< \sqrt{\frac{p\log n}{n}}(2(1+\sqrt{\frac{p}{e}})) \\ &< \frac{4p}{\sqrt{e}}\sqrt{\frac{\log n}{n}} \ ,\end{aligned}$$

where the last inequality holds for $p > 2$. Note that $\|X\|_F \leq \sqrt{\text{rank}(X)}|||X|||_2$. Given we are in the setting $p < n$ so the matrix is full rank, we thus obtain the stated result (24)

$$P(\|\hat{W}_{k;n}\|_F \geq \phi_{\max}\delta(n, p, \sqrt{p\log n/n})/\sqrt{p}) \leq 2n^{-p/2} \ .$$

□

B.7 Proof of Lemma 3

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

$$\begin{aligned}E_k^{'} &:= \{(\hat{\tau}_{k+1} - \tau_k) \geq d_{\min}/2\} \ , \\ E_k^{''} &:= \{(\tau_k - \hat{\tau}_k) \geq d_{\min}/2\} \ .\end{aligned}$$

In particular, one can demonstrate that:

$$P[A_{T,k} \cap D_T^{(m)}] \leq P[A_{T,k} \cap E_k^{'} \cap D_T^{(m)}] + \sum_{j=k+1}^{K} P[E_j^{''} \cap E_j^{'} \cap D_T^{(m)}] \ . \quad (26)$$

Proof. The strategy is to expand the probability in terms of exhaustive events (relating to the estimated changepoint positions), under a symmetry argument, we assume $\hat{\tau}_k < \tau_k$. Noting $P[E_k^{'} \cup E_{k+1}^{''}] = 1$, then expanding the original event, we find

$$\begin{aligned}P[A_{T,k} \cap D_T^{(m)}] &\leq P[A_{T,k} \cap D_T^{(m)} \cap E_k^{'}] + P[A_{T,k} \cap D_T^{(m)} \cap E_{k+1}^{''}] \\ &\leq P[A_{T,k} \cap D_T^{(m)} \cap E_k^{'}] + P[D_T^{(m)} \cap E_{k+1}^{''}] \ .\end{aligned}$$

Now consider the event $D_T^{(m)} \cap E_{k+1}^{''}$ corresponding to the second term. One can then expand the probability of this intersection over the events $E_{k+1}^{'}$ and $E_{k+2}^{''}$ relating to the next changepoint, i.e

$$P[D_T^{(m)} \cap E_k^{''}] \leq P[D_T^{(m)} \cap E_{k+1}^{''} \cap E_{k+1}^{'}] + P[D_T^{(m)} \cap E_{k+1}^{''} \cap E_{k+2}^{''}] \ .$$

Again, $P[D_T^{(m)} \cap E_k^{''} \cap E_{k+2}^{''}]$ may be upper bounded by $P[D_T^{(m)} \cap E_{k+2}^{''}]$ such that $P[D_T^{(m)} \cap E_k^{''} \cap E_{k+2}^{''}] \leq P[D_T^{(m)} \cap E_{k+2}^{''}]$. Cascading this over all changepoints $j = k+1, \ldots, K$ we have

$$P[D_T^{(m)} \cap E_k^{''}] \leq \sum_{j=k+1}^{K} P[D_T^{(m)} \cap E_j^{''} \cap E_{j+1}^{'}] \ .$$

□

B.8 Proof of Lemma 4

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

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

Proof. We present here an expanded version of the proof given in Harchaoui and Lévy-Leduc [2010]. Recall the definitions of the different events:

$$E_k^{'} := \{(\hat{\tau}_{k+1} - \tau_k) \geq \frac{d_{\min}}{2}\} \quad \text{and} \quad E_k^{''} := \{(\tau_k - \hat{\tau}_k) \geq \frac{d_{\min}}{2}\} \; .$$

For each new changepoint in the model, there is an extra option for this (latest changepoint) to trigger the event

$$\{\exists k \in [K],\ \hat{\tau}_k \leq \tau_{k-1}\} \; . \tag{27}$$

In particular, the total number of combinations (of changepoints) which could trigger this event doubles on the addition of an extra changepoint. Lemma 4 considers the probability of each of the changepoints being estimated to the left of τ_{k-1}. To start, we remark that the probability of $D_T^{(l)}$ is bounded by

$$P[D_T^{(l)}] \leq \sum_{k=1}^{K} 2^{k-1} P[\max\{l \in [K] \,|\, \hat{\tau}_l \leq \tau_{l-1}\} = k] \; . \tag{28}$$

The term $P[\max\{l \in [K] \mid \hat{\tau}_l \leq \tau_{l-1}\} = k]$ describes the probability that the last changepoint (such that $\hat{\tau}_l$ is to the left, i.e. before τ_{l-1}) is k. On increasing k by one (for $k \geq 2$), the number of combinations of left/right estimates for previous changepoints doubles. For example, consider the case for $k = 3$ such that the event $S_3 := \{\hat{\tau}_3 \leq \tau_2\}$ is triggered, see Fig. 2. The possible results for previous changepoints are then $S_2 := \{\hat{\tau}_2 \leq \tau_1\}$, it's compliment S_2^c, and the event $S_1 := \{\hat{\tau}_1 \leq 1\}$ or S_1^c. In total, there are 2^2 ways that the event S_3 can occur[1]. In general for the changepoint k there are 2^{k-1} combinations of events that allow S_k to be triggered. However, since these events are not mutually exclusive, this only provides an upper bound.

Harchaoui and Lévy-Leduc [2010] and Kolar and Xing [2012] remark that an event where the kth changepoint is the largest to satisfy $\{\hat{\tau}_l \leq \tau_{l-1}\}$, is a subset of events relating to later changepoints $l \geq k$. Correspondingly, we have

$$\{\max\{l \in [K] \,|\, \hat{\tau}_l \leq \tau_{l-1}\} = k\} \subseteq \cup_{l=k}^{K} \{\tau_l - \hat{\tau}_l \geq d_{\min}/2\} \cap \{\hat{\tau}_{l+1} - \tau_l \geq d_{\min}/2\} \; . \tag{29}$$

The union bound applied to (29) provides us with the bound:

$$P[\max\{l \in [K] \,|\, \hat{\tau}_l \leq \tau_{l-1}\} = k] \leq \sum_{l \geq k} P[\{\tau_l - \hat{\tau}_l \geq \frac{d_{\min}}{2}\} \cap \{\hat{\tau}_{l+1} - \tau_l \geq \frac{d_{\min}}{2}\}] \; ,$$

[1] Arguably, there are actually 3 combinations of changepoint event that can cause S_3 as $\hat{\tau}_1 > \hat{\tau}_0 = 1$ by definition. However, this does not effect the upper bound.

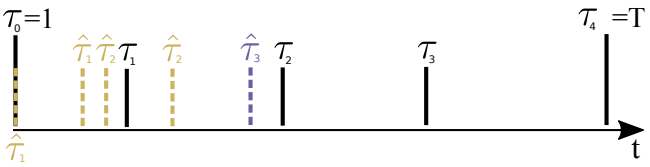

Figure 2: The gold changepoint estimates indicate examples of allowable positions for the changepoints $l < k = 3$ which satisfy $\{\hat{\tau}_l \leq \tau_{l-1}\}$. Note: for the case displayed $K = 3$ and $k = 3$ thus there are 4 combinations of changepoints (in gold) that permit the purple event $\max\{l \in [4] \mid \hat{\tau}_l \leq \tau_{l-1}\} = 3$.

and thus

$$P[D_T^{(l)}] \leq \sum_{k=1}^{K} 2^{k-1} \sum_{l \geq k}^{K} P[\{\tau_l - \hat{\tau}_l \geq \frac{d_{\min}}{2}\} \cap \{\hat{\tau}_{l+1} - \tau_l \geq \frac{d_{\min}}{2}\}] \, .$$

Since we want an upper bound, the largest factor (2^K) can be taken out the summation. The term $k = K$ contains the event $\{\hat{\tau}_{K+1} - \tau_K \geq d_{\min}/2\}$, this occurs with probability one as the last changepoint $\hat{\tau}_{K+1} = T + 1$. We can thus truncate the final term and obtain the bound:

$$\begin{aligned} P[D_T^{(l)}] \quad \leq \quad & 2^K \sum_{k=1}^{K-1} \sum_{l \geq k}^{K-1} P[\{\tau_l - \hat{\tau}_l \geq \frac{d_{\min}}{2}\} \cap \{\hat{\tau}_{l+1} - \tau_l \geq \frac{d_{\min}}{2}\}] \\ & + 2^K P[\{\tau_K - \hat{\tau}_K \geq \frac{d_{\min}}{2}\}] \, . \end{aligned}$$

The above can be written in a shortened form by relating it to the events $E_k^{'}, E_k^{''}$ defined in (18), such that

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

□

C Proof of Model-Selection Consistency

C.1 Proof Overview

Let us define a set of pairs $\mathcal{M}_k = \{(i,j) \mid \Theta_{0;ij}^{(k)} \neq 0\}$ to be the support set of the true model in block k and let $\mathcal{M}_k^{\perp} = \{(i,j) \mid \Theta_{ij}^{(k)} = 0\}$ be its compliment.

Theorem. *Bounds on the estimation error*

Consider the GFGL estimator with Assumptions (4,5). Assume $\lambda_1 = 16\alpha^{-1}\delta$, $\lambda_2 = \rho\lambda_1$ *for some finite* $\rho > 0$ *and* $\delta > 2^4\sqrt{2}c_{\sigma_\infty}\sqrt{\log(4p^\beta)/\gamma_{\min}T}$ *where* $\beta > 2$.

Let d be the maximum degree of each node in the true model, and define

$$v_{\mathcal{C}} = 6\{1 + 16\alpha^{-1}(1+2\rho)\}d\max\{K_{\Sigma_0}K_{\Gamma_0}, K_{\Gamma_0}^2 K_{\Sigma_0}^3\} \ .$$

Then for $T \geq 2^9\gamma_{\min}^{-1}\max\{1/8c_{\sigma_\infty}, v_{\mathcal{C}}\}^2 c_{\sigma_\infty}^2\log(4p^\beta)$*, we have*

$$\|\hat{\Theta}^{(k)} - \Theta_0^{(k_{\max})}\|_\infty \leq 2K_{\Gamma_0}\{1 + 16\alpha^{-1}(1+2\rho\hat{n}_k^{-1})\}\delta \ ,$$

in probability greater than $1 - 1/p^{\beta-2}$*.*

Proof. A popular and fairly general approach to demonstrating such consistency is known as the *primal-dual witness* method [Wainwright, 2009]. Principally, this method works by deconstructing the KKT conditions of an M-estimator into two blocks. Let us label these conditions $\mathrm{KKT}(\mathcal{M}, \partial r(\Theta_{\mathcal{M}}))$ and $\mathrm{KKT}(\mathcal{M}^\perp, \partial r(\Theta_{\mathcal{M}^\perp}))$, such that they respectively concern components of the the true model $\Theta_{0;\mathcal{M}}$ and the compliment $\Theta_{0;\mathcal{M}^\perp}$. The primal-dual witness approach consists of the following steps:

1. Solve a restricted problem; $\bar{\Theta} := \arg\min_\Theta l(\Theta; X) + \lambda_T r(\Theta)$, such that $\Theta_{\mathcal{M}^\perp} = 0$. This constitutes a restricted estimation problem, whereby the estimate is only supported on the true model-subspace. It verifies that the $\mathrm{KKT}(\mathcal{M}, \partial r(\Theta_{\mathcal{M}}))$ is satisfied under the block corresponding to the true support.

2. Select $\bar{R}$ as the sub-differential of the regulariser $r(\cdot)$ evaluated at $\bar{\Theta}$. Find the subgradient over the components outside the model-subspace $\mathcal{M}^\perp$ via $\mathrm{KKT}(\bar{\Theta}, \bar{R})$

3. Check that the sub-gradient in step (2) is sufficiently small to demonstrate the solution is dual feasible.

In what follows we wll dissect the GFGL estimator according to the above steps, a similar approach in the stationary i.i.d setting is discussed in Ravikumar et al. [2011]. In our case, we will require some care to take account of the smoothing regulariser.

Let $\hat{S}_{k;\hat{n}_k} := \hat{n}_k^{-1}\sum_{t=1}^{\hat{n}_k} X^{(t)}(X^{(t)})^\top$ represent the empirical covariance matrix calculated by taking $\hat{n}_k$ samples from $X^{(t)}$ with covariance matrix $\Sigma_0^{(k)}$. 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

$$\sum_{l \neq k \in [\hat{B}]} \hat{n}_{lk}(W_{l;\hat{n}_{lk}}) + \hat{n}_{kk}W_{k;\hat{n}_{kk}} - \hat{n}_k\hat{\Sigma}^{(k)} + \lambda_1\hat{n}_k\hat{R}_1^{(\hat{\tau}_{k-1})} + \lambda_2(\hat{R}_2^{(\hat{\tau}_{k-1})} - \hat{R}_2^{(\hat{\tau}_k)}) = 0 \ , \tag{30}$$

where $\hat{n}_{lk}$ describes the proportion of overlap between the lth true block and the kth estimated block, and $W_{k;n} := \Sigma_0^{(k)} - \hat{S}_{k;n}$ represents the error accrued in the empirical covariance estimate. The term $\sum_{l \neq k \in [\hat{B}]} \hat{n}_{lk}(W_{l;\hat{n}_{lk}})$ can be though of as providing a sampling bias due to estimation error in the changepoints, wheras the term $\hat{n}_{kk}W_{k;\hat{n}_{kk}}$ compares samples and the ground-truth of the same underlying covariance matrix.

We will now proceed to construct an oracle estimator $\bar{\Theta} \in \bar{\mathbb{R}}^{\hat{B}p \times \hat{B}p}$. The oracle is

constructed through solving the restricted problem

$$\bar{\Theta} := \underset{\{U^{(k)} \succ 0 |\, U^{(k)}_{\mathcal{M}^\perp} = 0\}_{k=1}^{\hat{B}}}{\text{Argmin}} \Bigg[\sum_{k=1}^{\hat{B}} \Big\{ \sum_{l=1}^{\hat{B}} \hat{n}_{lk} \text{tr}(\hat{S}^{(l)} U^{(k)}) - \hat{n}_k \log\det(U^{(k)}) \Big\} + \lambda_1 \sum_{k=1}^{\hat{B}} \hat{n}_k \|U^{(k)}\|_1 + \lambda_2 \sum_{k=2}^{\hat{B}} \|U^{(k)} - U^{(k-1)}\|_F \Bigg] .$$

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 (30). 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 max of 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

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

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

Lemma 5. *The out-of-subspace parameters are dual feasible such that* $\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty < 1$ *if*

$$\max \left\{ \|\text{ave}(W^{(k)})\|_\infty, \|\mathcal{E}(\Delta)\|_\infty, \frac{\lambda_2}{\hat{n}_k} \|\bar{R}_2^{(\hat{\tau}_{k-1})}\|_\infty, \frac{\lambda_2}{\hat{n}_k} \|\bar{R}_2^{(\hat{\tau}_k)}\|_\infty \right\} \leq \alpha \lambda_1 / 16 , \tag{31}$$

where $\text{ave}(W^{(k)}) := \hat{n}_k^{-1} (\sum_{l \neq k}^{\hat{B}} \hat{n}_{lk} W_{l;\hat{n}_{lk}} + \hat{n}_{kk} W_{k;\hat{n}_{kk}})$.

We note at this point, that the condition (31) 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 $\text{ave}(W^{(k)}) \to W^{(k)}_{\text{exact}}$. To make this argument more specific, we construct a loose bound on the sampling error accumulated over an estimated block.

Lemma 6. *The sampling error over a block is almost surely bounded according to*

$$\sum_{l \in \hat{\mathcal{B}}^{(k)}} \hat{n}_{lk} \|W_{l;\hat{n}_{lk}}\|_\infty \leq \max\{\hat{n}_k, d_{\min}\} \|W_{l_\infty; d_{\min}/2}\|_\infty ,$$

and thus the average sampling error is bounded according to

$$\|\mathrm{ave}(W^{(k)})\|_\infty \leq \max\{1, d_{\min}/\hat{n}_k\}\|W_{\infty;d_{\min}/2}\|_\infty \ .$$

If changepoint estimation is consistent (according to Assumption 4) then $P(E_\tau) = f_\tau(T,p)$, and thus have $d_{\min}/\hat{n}_k < d_{\min}/(d_{\min} - \delta_T T) \to 1$, as $T \to \infty$. As a result, we bound $\|\mathrm{ave}(W^{(k)})\|_\infty \leq \|W_{\infty;d_{\min}/2}\|_\infty$ from above and then analyse the conditions (31) on condition of the intersection $E_\tau \cap \mathcal{C}$ where

$$\mathcal{C} := \{\|\hat{W}_{k_\infty;d_{\min}/2}\|_\infty \leq \delta\} \ .$$

Through choice of regulariser $\lambda = 16\alpha^{-1}\delta$, the condition $\|\mathrm{ave}(W^)\|_\infty \leq \alpha\lambda_1/16$ is automatically satisfied. 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 7. *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 to map $\|\Delta\|_\infty$ to $\|\mathcal{E}(\Delta)\|_\infty$ we need to modify our argument for the actual control on $\|\Delta\|_\infty$.

Lemma 8. *The elementwise ℓ_∞ norm of the error is bounded such that $\|\bar{\Delta}\|_\infty = \|\bar{\Theta} - \Theta_0\|_\infty \leq r$ if*

$$r := 2K_{\Gamma_0}\{\|\mathrm{ave}(W_k)\|_\infty + \lambda_1 + \lambda_2\hat{n}_k^{-1}(\|\bar{R}_2^{(\hat{\tau}_{k-1})}\|_\infty + \|\bar{R}_2^{(\hat{\tau}_k)}\|_\infty)\} \ , \qquad (32)$$

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

We now propogate the results of Lemma 8 through Lemma 7, while conditioning on event $\{\mathcal{C} \cap E_\tau\}$. First, let us note 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 (32) with the stated form of λ_1 such that

$$r \leq r_{\mathcal{C}} := 2K_{\Gamma_0}\{\delta + \lambda_1(1 + 2\rho\hat{n}_k^{-1})\} = 2K_{\Gamma_0}\{1 + 16\alpha^{-1}(1 + 2\rho\hat{n}_k^{-1})\}\delta \ .$$

The condition in Lemma 8 is now met, if $\delta \in (0, 1/\max\{1/8c_{\sigma_\infty}, v_{\mathcal{C}}\})$ where

$$v_{\mathcal{C}} = 6\{1 + 16\alpha^{-1}(1 + 2\rho\hat{n}_k^{-1})\}d\max\{K_{\Sigma_0}K_{\Gamma_0}, K_{\Gamma_0}^2K_{\Sigma_0}^3\} \ .$$

Using Lemma B.5 we have the probabilistic bound on $\mathcal{C}^c$ given as $P(\|W_{k_\infty;d_{\min}/2}\|_\infty > \delta) \leq 1/p^{(\beta-2)}$, where we need $\delta > 2^4\sqrt{2}c_{\sigma_\infty}\sqrt{\log(4p^\beta)/\gamma_{\min}T}$. Remember $c_{\sigma_\infty} = \max_k\{c_{\sigma_k}\}$, i.e. we assume the slowest concentration possible over all blocks $k = 1, \ldots, B$. This results in a lower bound for the sample size of

$$T \geq 2^9\gamma_{\min}^{-1}\max\{1/8c_{\sigma_\infty}, v_{\mathcal{C}}\}^2 c_{\sigma_\infty}^2\log(4p^\beta) \quad ; \quad \beta > 2 \ . \qquad (33)$$

The restricted solution is hence dual-feasible under the conditions of the proof such that $\bar{\Theta} = \hat{\Theta}$. The true edges are included in the estimate with probability at least

$1 - 1/p^{\beta-2}$. 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 9. *For matrices $\Theta_T \in \mathcal{S}_{++}^T := \{\{A^{(t)}\}_{t=1}^T \mid A^{(t)} \succ 0\ ,\ A^{(t)} = A^{(t)\top}\}$ the GFGL cost function is strictly convex.*

□

Theorem. *Model-selection consistency*

In addition to the assumptions in Theorem 3.3. Let $\theta_{\min}^{(k)} := \min_{ij} |\Theta_{0;ij}^{(k)}|$ for all $(i,j) \in \mathcal{M}_k$ and for each $k = 1, \ldots, B$. Let

$$v_\theta = 2K_{\Sigma_0}\{1 + 16\alpha^{-1}(1 + 2\rho\hat{n}_k^{-1})\}\theta_{\min}^{-1}\ .$$

If $T \geq 2^9\gamma_{\min}^{-1}\max\{(8c_{\sigma_\infty})^{-1}, v_{\mathcal{C}}, v_\theta\}^2 c_{\sigma_\infty}^2 \log(4p^\beta)$ then GFGL attains sign-consistency with probability tending to one:

$$P\{E_{\mathcal{M}}(\hat{\Theta}^{(k)}; \Theta_0^{(k_{\max})})\} \geq 1 - 1/p^{\beta-2} \to 1 \quad ; \quad \text{as } (p,T) \to \infty\ .$$

Proof. This result is a simple corollary of Theorem 3.3 with the additional condition on a known and appropriately sized $\theta_{\min}$. The proof follows from Lemma 8 where we require $r = r_\theta \geq \theta_{\min}/2$. The element-wise error incurred by the estimate $|\hat{\Theta}_{ij}^{(k)} - \Theta_0^{(k)}|$ is not enough to change the sign of the estimate and the result follows. An exact analogy is given (in the static case) by Lemma 7 Ravikumar et al. [2011]. □

C.2 Dual-feasibility with sampling and mis-specification error (Proof of Lemma 5)

Lemma. *Dual Feasibility*

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

$$\max\left\{\|\text{ave}(W)\|_\infty, \|\mathcal{E}(\Delta)\|_\infty, \frac{\lambda_2}{\hat{n}_k}\|\bar{R}_2^{(\hat{\tau}_{k-1})}\|_\infty, \frac{\lambda_2}{\hat{n}_k}\|\bar{R}_2^{(\hat{\tau}_k)}\|_\infty\right\} \leq \alpha\lambda_1/16\ .$$

Proof. We can write the block-wise optimality conditions (30) for the restricted estimator as

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

As pointed out in Ravikumar et al. [2011], this equation may be written as an ordinary linear equation by vectorising the matrices, for instance $\text{vec}\{(\Theta_0^{(k)})^{-1}\Delta^{(k)}(\Theta_0^{(k)})^{-1}\} =$

$\{(\Theta_0^{(k)})^{-1} \otimes (\Theta_0^{(k)})^{-1}\}\mathrm{vec}(\Delta^{(k)}) \equiv \Gamma_0 \mathrm{vec}(\Delta)$. Utilising the fact $\Delta_{\mathcal{M}^\perp} = 0$ we can split the optimality conditions into two blocks of linear equations

$$\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}}\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}) + \mathrm{vec}(G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}}) = 0 \tag{34}$$

$$\Gamma^{(k)}_{0;\mathcal{M}^\perp\mathcal{M}}\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}) + \mathrm{vec}(G^{(k)}_{\hat{n}_k}(X;\lambda_1,\lambda_2)_{\mathcal{M}^\perp}) = 0 \ , \tag{35}$$

where

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

and the average empirical covariance error over the block is given by

$$\mathrm{ave}(W^{(k)}) := \hat{n}_k^{-1}(\sum_{l\neq k}^{\hat{B}} \hat{n}_{lk} W_{l;\hat{n}_{lk}} + \hat{n}_{kk} W_{k;\hat{n}_{kk}}) \ .$$

Solving (34) for $\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}})$ we find

$$\mathrm{vec}(\Delta^{(k)}_{\mathcal{M}}) = -(\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\{G(X;\lambda_1,\lambda_2)_{\mathcal{M}}\} \ .$$

Substituting this into (35) and re-arranging for $\bar{R}_{1;\mathcal{M}^\perp}$ gives

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

and thus

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

where $H^{(k)} = \Gamma^{(k)}_{0;\mathcal{M}^\perp\mathcal{M}}(\Gamma^{(k)}_{0;\mathcal{M}\mathcal{M}})^{-1}$. Taking the ℓ_∞ norm of both sides gives

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

Lemma 10. *The error in the model-space dominates that outside such that*

$$\|\mathrm{ave}(W_{\mathcal{M}^\perp})\|_\infty \leq \|\mathrm{ave}(W_{\mathcal{M}})\|_\infty \ , \tag{36}$$

$$\|\mathcal{E}_{\mathcal{M}^\perp}(\Delta)\|_\infty \leq \|\mathcal{E}_{\mathcal{M}}(\Delta)\|_\infty \ . \tag{37}$$

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

Via the results above, we obtain $\|H^{(k)}\mathrm{vec}(\bar{R}^{(k)}_{1;\mathcal{M}})\|_\infty \leq 1-\alpha$ and

$$\|\bar{R}^{(k)}_{1;\mathcal{M}^\perp}\|_\infty \leq \lambda_1^{-1}(2-\alpha)\{\|\mathrm{ave}(W_{\mathcal{M}})\|_\infty + \|\mathcal{E}_{\mathcal{M}}(\Delta)\|_\infty + \lambda_2 \hat{n}_k^{-1}(\|\bar{R}^{(\hat{\tau}_{k-1})}_{2;\mathcal{M}}\|_\infty + \|\bar{R}^{(\hat{\tau}_k)}_{2;\mathcal{M}}\|_\infty)\} + \|H^{(k)}\mathrm{vec}(\bar{R}^{(k)}_{1;\mathcal{M}})\|_\infty \, .$$

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

□

C.3 Control on the remainder term $\|\mathcal{E}(\Delta)\|_\infty$ (Proof of Lemmas 7,8)

Lemma. *Lemma 5 Ravikumar et al. [2011]*

If the bound $\|\Delta\|_\infty \leq (3K_{\Sigma_0}d)^{-1}$ holds and d is the maxmimum node degree, then

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

Proof. The reader is directed to Ravikumar et al. [2011] for full details. The proof relies on representing the remainder term of the log-det function as

$$\mathcal{E}(\Delta) = \Theta_0^{-1}\Delta\Theta_0^{-1}\Delta J\Theta_0^{-1} \, , \tag{38}$$

for matrix

$$J := \sum_{m=0}^{\infty}(-1)^m(\Theta_0^{-1}\Delta)^m \, .$$

Given the stated control on $\|\Delta\|_\infty$ the norm of this matrix can be bounded such that $\left|\left|\left|J^\top\right|\right|\right|_\infty \leq 3/2$, the result follows by working through (38) with a maximum degree size d.

□

Lemma. *Control of Estimation error*

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

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

and

$$r \leq \min\left\{\frac{1}{3K_{\Sigma_0}d}, \frac{1}{3K^3_{\Sigma_0}K_{\Gamma_0}d}\right\} \, .$$

Proof. Note that $\bar{\Theta}_{\mathcal{M}^\perp} = \Theta_{0;\mathcal{M}^\perp} = 0$ and thus $\|\Delta\|_\infty = \|\Delta_{\mathcal{M}}\|_\infty$. We follow Ravikumar et al. [2011] (Lemma 6) 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])

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

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

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

such that $F\{\mathrm{vec}(\Delta_{\mathcal{M}})\} = \mathrm{vec}(\Delta_{\mathcal{M}})$ iff $Q(\Theta^{(k)}_{0;\mathcal{M}} + \Delta_{\mathcal{M}}) = Q(\Theta^{(k)}_{\mathcal{M}}) = 0$.

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 $\mathbb{B}(r) := \{\Theta_{\mathcal{M}} |\, \|\Theta_{\mathcal{M}}\|_\infty \leq r\}$ onto itself. Expanding $F(\mathrm{vec}(\Delta_{\mathcal{M}}))$, we find

$$\begin{aligned}
F(\mathrm{vec}(\Delta_{\mathcal{M}})) &= -(\Gamma_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\{Q(\Theta^{(k)}_{0;\mathcal{M}} + \Delta_{\mathcal{M}})\} + \mathrm{vec}(\Delta_{\mathcal{M}}) \\
&= (\Gamma_{0;\mathcal{M}\mathcal{M}})^{-1}\mathrm{vec}\big[\{(\Theta^{(k)}_0 + \Delta)^{-1} - (\Theta^{(k)}_0)^{-1}\}_{\mathcal{M}} - \mathrm{ave}(W_{k;\mathcal{M}}) - \lambda_1 \bar{R}^{(k)}_{1;\mathcal{M}} \\
&\quad - \lambda_2 \hat{n}_k^{-1}(\bar{R}^{(\hat{\tau}_{k-1})}_{2;\mathcal{M}} - \bar{R}^{(\hat{\tau}_k)}_{2;\mathcal{M}})\big] + \mathrm{vec}(\Delta_{\mathcal{M}}) \\
&= T_1 - T_2 \, ,
\end{aligned}$$

where

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

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

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

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 $\|F(\mathrm{vec}(\Delta_{\mathcal{M}}))\|_\infty \leq r$. Through the construction of F, we have $\|\Delta_{\mathcal{M}}\|_\infty \leq r$ iff $Q(\Theta^{(k)}_{0;\mathcal{M}} + \Delta_{\mathcal{M}}) = Q(\Theta^{(k)}_{\mathcal{M}}) = 0$ and since $Q(\bar{\Theta}^{(k)}_{\mathcal{M}}) = 0$ for any $\bar{\Theta}^{(k)}_{\mathcal{M}}$ we obtain $\|\bar{\Delta}_{\mathcal{M}}\|_\infty \leq r$ where $\bar{\Delta}_{\mathcal{M}} := \bar{\Theta} - \Theta_0$. Finally, the existence of a solution $\bar{\Theta}^{(k)}_{\mathcal{M}}$ corresponding to $\mathrm{vec}(\bar{\Delta}_{\mathcal{M}}) \in \mathbb{B}(r)$ is guranteed by Brouwer's fixed point theorem (cite).

□

C.4 Strict Convexity of GFGL (Proof of Lemma 9)

Lemma. *The GFGL cost function is strictly convex*

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

Proof. The neagtive log-det barrier $-\log\det(\Theta^{(t)})$ is strictly convex on $\Theta^{(t)} \in \mathcal{S}^1_{++}$. While the frobenius norm is stictly convex on a given matrix $\|A\|_F$ for $A \in \mathcal{S}^1_{++}$ it is not striclty 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

$$\min_{\Theta_T \in \mathcal{S}^{(T)}_{++}} \Big\{ \sum_{t=1}^T \big[\langle \Theta^{(t)}, \hat{S}^{(t)} \rangle - \log\det(\Theta^{(t)}) \big] \Big\}$$
$$\text{such that} \quad \sum_{t=1}^T \|\Theta^{(t)}_{-ii}\|_1 + \frac{\lambda_2}{\lambda_1} \sum_{t=2}^T \|\Theta^{(t)} - \Theta^{(t-1)}\|_F \leq C(\lambda_1) \; .$$

We can alternatively write

$$\min_{\Theta_T \in \mathcal{S}^{(T)}_{++}} \Big\{ \sum_{t=1}^T \big[\langle \Theta^{(t)}, \hat{S}^{(t)} \rangle - \log\det(\Theta^{(t)}) \big] \Big\}$$
$$\text{such that} \quad \sum_{t=1}^T \|\Theta^{(t)}_{-ii}\|_1 \leq C_{\text{sparse}}(\lambda_1, \lambda_2)$$
$$\sum_{t=2}^T \|\Theta^{(t)} - \Theta^{(t-1)}\|_F \leq C_{\text{smooth}}(\lambda_1, \lambda_2) \; .$$

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 $\hat{S}^{(t)}_{ii} > 0$ 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

$$\min_{\Theta^{(t)} \in \mathcal{S}^{(1)}_{++}} \big\{ \langle \Theta^{(t)}, \hat{S}^{(t)} \rangle - \log\det(\Theta^{(t)}) \big\}$$
$$\text{such that} \quad \|\Theta^{(t)}_{-ii}\|_1 \leq C_t(\lambda_1, \lambda_2) \; .$$

The constraint due to smoothing allows exactly this, for instance, one may obtain a bound

$$\|\Theta^{(t)}_{-ii}\|_1 \leq C_{\text{sparse}}(\lambda_1, \lambda_2) - \sum_{s \neq t} \|\Theta^{(s)}_{-ii}\|_1 \; .$$

Writing $\Theta^{(s)} = \Theta^{(1)} + \sum_{q=2}^s (\Theta^{(q)} - \Theta^{(q-1)})$ for $s \geq 2$ we obtain

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

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

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

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

□

C.5 Bounding the block-wise sampling error (Proof of Lemma 6)

Let $\hat{\mathcal{B}}^{(k)}$ be a set containing all true block indexes l which overlap with the estimated block of interest k and let $\hat{n}_{lk} = \min\{\min\{\hat{\tau}_k, \tau_{l+1}\} - \max\{\hat{\tau}_{k-1}, \tau_l\}, 0\}$ be the overlap of the lth true block and kth estimate block. The average sampling error over block k can then be written as

$$\text{ave}(W^{(k)}) := \frac{1}{\hat{n}_k} \sum_{l\in\hat{\mathcal{B}}^{(k)}} \hat{n}_{lk}(\hat{S}_{l;\hat{n}_{lk}} - \Sigma^{(l)})\ .$$

Given a lower bound on the true size of blocks $d_{\min} = \min_k\{\tau_k - \tau_{k-1}\}$ we can bound the average sampling error (even in the case where $\delta_T T > d_{\min}$).

Lemma. *Let $\hat{n}_k := \hat{\tau}_k - \hat{\tau}_{k-1}$. The sampling error over a block is almost surely bounded according to*

$$\sum_{l\in\hat{\mathcal{B}}^{(k)}} \hat{n}_{lk}\|W_{l;\hat{n}_{lk}}\|_\infty \leq \max\{\hat{n}_k, d_{\min}\}\|W_{l_\infty;d_{\min}/2}\|_\infty\ ,$$

Proof. We consider to bound cases where $\hat{\tau}_k - \hat{\tau}_{k-1} < d_{\min}$ differently according to

$$\sum_{l\in\mathcal{B}^{(k)}} \hat{n}_{lk}\|W_l^{(\hat{n}_{lk})}\|_\infty \leq \begin{cases} (\hat{\tau}_k - \hat{\tau}_{k-1})\|W_\infty^{(d_{\min}/2)}\|_\infty & \text{if } \hat{\tau}_k - \hat{\tau}_{k-1} \geq d_{\min}\ (A) \\ d_{\min}\|W_\infty^{(d_{\min}/2)}\|_\infty & \text{if } \hat{\tau}_k - \hat{\tau}_{k-1} < d_{\min}\ (B) \end{cases}$$

Case A

To show the first bound where $\hat{\tau}_k - \hat{\tau}_{k-1} \geq d_{\min}$, consider the simple case where we have two true blocks that overlap the estimated block k, namely l and $l+1$. Note that $W_l^{(n_1)}$ is independent of $W_{l+1}^{(n_2)}$, thus we can write

$$P[n_1\|W_l^{(n_1)}\|_\infty + n_2\|W_l^{(n_2)}\|_\infty > \epsilon] = P[\|W_l^{(n_1)}\|_\infty > \epsilon/2n_1] + P[\|W_{l+1}^{(n_2)}\|_\infty > \epsilon/2n_2]\ .$$

If we further assume that the structure in the blocks is the same, i.e. $\Sigma_0^{(l)} = \Sigma_0^{(l+1)}$ then we can use the bound of Ravikumar et al. (c.f. Lemma 1) to state

$$P[n_1\|W_l^{(n_1)}\|_\infty + n_2\|W_l^{(n_2)}\|_\infty > \epsilon] \propto p^2\big(\exp(-c_l^{-1}\epsilon^2 n_1^{-1}) + \exp(-c_l^{-1}\epsilon^2 n_2^{-1})\big)\ . \tag{39}$$

Letting $n_2 = |\hat{\tau}_k - \hat{\tau}_{k-1}| - n_1$ and differentiating with respect to n_1 we find that the probability of exceeding ϵ is maximised at $n_1 = n_2 = |\hat{\tau}_k - \hat{\tau}_{k-1}|/2$. To arrive at the result (in the two block setting), we simply note that the block $\Sigma_0^{(l)}$ that we sample from should be taken to be the one with the largest c_σ.

Case B

In this case $\hat{\tau}_k - \hat{\tau}_{k-1} < d_{\min}$, the interval over which the sampling error occurs is relatively small. We desire to find a multiplier in terms of $d_{\min}$ such that we can still maintain

$$\sum_{l \in B^{(k)}} \hat{n}_{lk} \|W_{l;\hat{n}_{lk}}\|_\infty \leq n_{\text{eff}} \|W_{l_\infty;d_{\min}/2}\|_\infty \ . \tag{40}$$

In the extreme case where $\hat{n}_{lk} = 1$ and $\hat{\tau}_k - \hat{\tau}_{k-1} = 1$. From Ravikumar, we obtain

$$P[\|W_{l;1}\|_\infty > \epsilon d_{\min}] \leq p^2 \exp(-c_l^{-1} \epsilon^2 d_{\min}^2) \ ,$$

and

$$P[\|W_{l;d_{\min}/2}\|_\infty > \epsilon] \leq p^2 \exp(-2c_l^{-1} \epsilon^2 d_{\min}^{-1}) \ , \tag{41}$$

thus in high-probability $n_{\text{eff}} = d_{\min}$ maintains the bound (40). From the bound on the first part (A) we noted that the worst case error was obtained at $n_1 = n_2 = \hat{\tau}_k - \hat{\tau}_{k-1}/2$. This result still holds in the case where $\hat{\tau}_k - \hat{\tau}_{k-1} < d_{\min}$. For all $2n_1 = 1, \ldots, d_{\min}$, we find

$$P[\|W_{l;n_1}\|_\infty > \epsilon d_{\min} n_1^{-1}] \leq p^2 \exp(-c_l^{-1} \epsilon^2 d_{\min}^2 n_1^{-1}) \ .$$

The probability (41) is obtained at the limit $n_1 = d_{\min}/2$ and therefore the bound holds for all $n_1, n_2 < d_{\min}/2$.

Multiple Blocks

For cases with more blocks the result still holds by noting that $d_{\max} > \min_{l \in K^*}\{n_{lk}\}$, thus the error from any blocks fully contained in the interval $\{\hat{\tau}_{k-1}, \ldots, \hat{\tau}_k\}$ will be dominated by the blocks at either end. However, we do need to reassess the size of the boundary blocks n_1, n_2. Rather than combining over an interval $|\hat{\tau}_k - \hat{\tau}_{k-1}|$ this interval is reduced according to

$$n_1 = n_2 = \frac{1}{2}\big(|\hat{\tau}_k - \hat{\tau}_{k-1}| - \sum_{l \in \tilde{\mathcal{B}}^{(k)}} |\tau_l - \tau_{l-1}|\big) \ ,$$

where $\tilde{\mathcal{B}}^{(k)}$ is the set $\mathcal{B}^{(k)}$ without the first and last elements (which represent lengths n_1 and n_2 respectively). From assuption we have a lower bound on the changepoint distance $d_{\min}$, we can thus upper bound the size of the boundary regions according to

$$n_1 + n_2 \leq |\hat{\tau}_k - \hat{\tau}_{k-1}| - (|\mathcal{B}^{(k)}| - 2) d_{\min} \ .$$

Setting $n_1 = n_2 = \{|\hat{\tau}_k - \hat{\tau}_{k-1}| - (|\mathcal{B}^{(k)}| - 2)d\}/2$, substituting into (39) and then maximising with respect to d, suggests that the probability of exceeding level ϵ is maximised when $d = d_{\min}$. We can thus use this setting of n_1, n_2 to upper bound the error. The residual interval length $n_1 + n_2$ either satisfies the two block analysis for $n_1 + n_2 < d_{\min}$ or otherwise. The stated result therefore follows in generality. $\square$

References

D Angelosante and GB Giannakis. Sparse Graphical Modeling of Piecewise-Stationary Time-series. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1960–1963, 2011. ISBN 9781457705397. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=5946893.

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

Onureena Banerjee and Laurent El Ghaoui. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9:485–516, 2008.

K Bleakley and J P Vert. The group fused Lasso for multiple change-point detection. Technical Report HAL-00602121, HAL, 2011.

Tony Cai, Weidong Liu, and Xi Luo. A Constrained l1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011. doi: 10.1198/jasa.2011.tm10155.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 77(2):475–507, 2015. ISSN 14679868. doi: 10.1111/rssb.12079.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

Kenneth R. Davidson and Stanislaw J. Szarek. Local Operator Theory, Random Matrices and Banach Spaces. *Handbook of the Geometry of Banach Spaces*, 1:317–366, 2001. ISSN 18745849. doi: 10.1016/S1874-5849(01)80010-3.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014. ISSN 00905364. doi: 10.1214/14-AOS1245.

Alexander J. Gibberd and James 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. ISSN 15206149. doi: 10.1109/ICASSP.2014.6854087. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6854087.

Alexander J Gibberd and James D B Nelson. Regularized Estimation of Piecewise Constant Gaussian Graphical Models : The Group-Fused Graphical Lasso. *Journal of Computational and Graphical Statistics*, 2017.

David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network Inference via the Time-Varying Graphical Lasso. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017. ISBN 9781450348874. URL http://arxiv.org/abs/1703.01958.

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, dec 2010. ISSN 0162-1459. doi: 10.1198/jasa.2010.tm09181. URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2010.tm09181.

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, dec 2012. ISSN 0162-1459. doi: 10.1080/01621459.2012.737745. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.737745.

Mladen Kolar and Eric P Xing. On Time Varying Undirected Graphs. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 15:407–415, 2011.

Mladen Kolar and Eric P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6:2069–2106, 2012. ISSN 19357524. doi: 10.1214/12-EJS739.

Yan Lan, Moulinath Banerjee, and George Michailidis. Change-point estimation under adaptive sampling. *The Annals of Statistics*, 37(4):1752–1791, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS602. URL http://projecteuclid.org/euclid.aos/1245332832.

Marc Lavielle and Gilles Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006. ISSN 0363-1672. doi: 10.1007/s10986-006-0028-9.

Sokbae Lee, Myung Hwan Seo, and Youngki 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. ISSN 14679868. doi: 10.1111/rssb.12108. URL http://arxiv.org/abs/1209.4875.

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

C. Loader. Change point estimation using nonparametric regression. *The Annals of Statistics*, 24(4):1667–1678, 1996. URL http://projecteuclid.org/DPubS?service=UI{&}version=1.0{&}verb=Display{&}handle=euclid.aos/1032298290.

N. Meinshausen. A note on the Lasso for Gaussian graphical model selection. *Statistics & Probability Letters*, 78:880–884, 2008. doi: 10.1016/j.spl.2007.09.014.

Nicolai Meinshausen and Peter Bühlmann. High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000281. URL http://www.jstor.org/stable/25463463.

Ricardo Pio Monti, Peter Hellyer, David Sharp, Robert Leech, Christoforos Anagnostopoulos, and Giovanni Montana. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, 103:427–443, aug 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.07.033. URL http://www.ncbi.nlm.nih.gov/pubmed/25107854.

Hans-Georg Muller. Change-Points in Nonparametric Regression Analysis. *The Annals of Statistics*, 20(2):737–761, 1992. ISSN 0090-5364. doi: 10.1214/aos/1176348654. URL http://projecteuclid.org/euclid.aos/1176348654.

Hernando Ombao, Rainer von Sachs, and Wensheng Guo. SLEX Analysis of Multivariate Nonstationary Time Series. *Journal of the American Statistical Association*, 100(470):519–531, 2005. ISSN 0162-1459. doi: 10.1198/016214504000001448.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1/2):100, 1954. ISSN 00063444. doi: 10.2307/2333009. URL http://www.jstor.org/stable/2333009?origin=crossref.

Marc Raimondo. Minimax estimation of sharp change points. *Annals of Statistics*, 26 (4):1379–1397, 1998. ISSN 00905364. doi: 10.1214/aos/1024691247.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(January 2010):935–980, 2011. ISSN 19357524. doi: 10.1214/11-EJS631.

Sandipan Roy, Yves Atchadé, and George Michailidis. Change point estimation in high dimensional Markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. ISSN 13697412. doi: 10.1111/rssb.12205. URL http://doi.wiley.com/10.1111/rssb.12205.

Takumi Saegusa and Ali Shojaie. Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics*, 10(1):1341–1392, 2016. ISSN 19357524. doi: 10.1214/16-EJS1137.

Martin 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.

Tengyao Wang and Richard J. Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80000, 2017. ISSN 13697412. doi: 10.1111/rssb.12243. URL http://doi.wiley.com/10.1111/rssb.12243.

Yao Xie, Jiaji Huang, and Rebecca Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal on Selected Topics in Signal Processing*, 7(1):12–27, 2013. ISSN 19324553. doi: 10.1109/JSTSP.2012.2234082.