COMMUNITY DETECTION IN 
PARTIAL CORRELATION NETWORK MODELS

Christian Brownlees†  Guðmundur Stefán Guðmundsson†  Gábor Lugosi†,‡,♦

May 15, 2018

Abstract

We introduce a class of partial correlation network models with a community structure for large panels of time series. In the model, series are partitioned into latent groups such that correlation is higher within groups than between them. We then propose an algorithm that allows us to detect the communities using the eigenvectors of the sample covariance matrix. We study the properties of the procedure and establish its consistency. The methodology is used to study real activity clustering in the U.S. and Europe.

Keywords: Community Detection, Graphical Models, Partial Correlation Networks, Random Graphs, Spectral Clustering

JEL: C3, C33, C55

† Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE; e-mail: christian.brownlees@upf.edu; stefan.gudmundsson@upf.edu; gabor.lugosi@upf.edu.
‡ ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
♦ Barcelona Graduate School of Economics

Christian Brownlees and Gábor Lugosi acknowledge financial support from the Spanish Ministry of Science and Technology (Grant MTM2015-67304) and from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2011-0075). We have benefited from discussions with Majid Al-Sadoon, Bamu Demir Pakel, Jin-Chuan Duan, Geert Mesters, Fabrizio Lillo, Cavit Pakel, Nicolas Verzelen, Pedro Souza, Michael Wolf and Piotr Zwiernik.
1 Introduction

Network analysis has become an active field of research in time series econometrics (Billio, Getmansky, Lo, and Pellizzon 2012; Hautsch, Schaumburg, and Schienle 2015; Diebold and Yilmaz 2014; Demirer, Diebold, Liu, and Yilmaz 2018). Contributions to network analysis in this area typically focus on developing methodologies to learn the interdependence structure of large multivariate systems. This is often achieved by assuming the interdependence structure of the system is sparse and then relying on sparse estimation methods (Meinshausen and Bühlmann 2006; Belloni and Chernozhukov 2011; Kock and Callot 2015; Medeiros and Mendes 2016; Basu and Michailidis 2015; Kock 2016). Network techniques have been applied successfully for dimensionality reduction and regularisation.

In the analysis of real-world networks it is often found that vertices are partitioned into groups such that the concentration of linkages is high among vertices in the same group and low otherwise. This empirical regularity is called community structure or clustering (Girvan and Newman 2002). A popular class of models introduced to study networks with such a community structure are Stochastic Block Models (SBMs) (Holland, Laskey, and Leinhardt 1983). SBMs are random graphs, that is, graphs in which vertices are fixed and the edges are determined randomly and independently by Bernoulli trials (Chung and Lu 2006; van der Hofstad 2015). In the simplest version of these models, the vertices are partitioned into \( k \) communities and the probability of an edge between two vertices is \( p \) if they belong to the same community and \( q \) otherwise, with \( p > q \).

In this work we introduce a class of partial correlation network models with a community structure for large panels of time series. In the model, the series are partitioned into latent groups such that correlation is higher within groups than between them. We achieve this by introducing a model where the partial correlation structure of the \( n \) series in the panel is determined by a latent graph defined over \( n \) vertices: The \( i \)-th and \( j \)-th time series have zero partial correlation if and only if vertices \( i \) and \( j \) in the graph are not connected by an edge. The graph that determines the partial correlation structure of the panel is then assumed to be an SBM.
A natural question that arises in this setting is how to detect the communities of the model from a sample of \( T \) observations. An extensive literature (Fortunato, 2010) deals with the problem of community detection when the network structure of the data is observable. A popular algorithm to carry out community detection is a spectral clustering algorithm that uses the eigenvectors of the graph Laplacian to detect community structure. In our framework, community detection is more challenging as the network structure of the data is not observed. Motivated by spectral clustering methods (Ng, Jordan, and Weiss, 2001; von Luxburg, 2007), we propose a community detection procedure called Blockbuster in which spectral clustering is applied to the sample covariance matrix of the observations. In particular, the algorithm applies \( k \)-means clustering to a matrix whose columns are the rescaled eigenvectors corresponding to the \( k \) largest eigenvalues of the sample covariance matrix. The \( k \)-means partition of the rows of the rescaled eigenvector matrix is the estimate of the community partition.

We study the properties of the Blockbuster community detection procedure and show that it consistently detects the communities when the number of observations and dimension of the panel are sufficiently large. More precisely, our key result establishes a bound on the fraction of vertices that the algorithm misclusters in a similar fashion to Rohe, Chatterjee, and Yu (2011) and Qin and Rohe (2013), and shows that it is close to zero with high probability when \( n \) and \( T \) are large enough, provided that \( n/T \) is small.

We consider an extension of our baseline model in which the time series in the panel are influenced by a set of common factors. Panels of economic and financial time series typically exhibit evidence of a factor structure and an extensive literature has developed around factor models (Bai and Ng, 2013; Forni, Hallin, Lippi, and Reichlin, 2000). We introduce a variant of our algorithm that detects communities in this setting and establish its consistency.

Moreover, we introduce a regularised covariance estimator based on the Blockbuster algorithm, which is motivated by the block covariance structure of the model we introduce in this work. The estimator takes a block-structured form where the diagonal blocks are the sample covariance matrices of each community, while thresholding is applied to the
off-diagonal blocks. The analysis of the theoretical properties of this estimator is beyond
the scope of the paper, but we use it in an out-of-sample forecast validation exercise in
the empirical application.

A natural application of our methodology is to study business cycle synchronisation
(Hamilton and Owyang 2012, Leiva-Leon 2017), where it is of interest to find regions
that co-move closely over the business cycle (Francis, Owyang, and Savascin 2017). We
apply our methodology to analyse two panels of real activity growth measures for the
states of the U.S. and the regions of Europe. For the U.S. we use a dataset constructed by
Hamilton and Owyang (2012) comprising quarterly employment growth rates at the state-
level from 1956-Q2 to 2007-Q4. For Europe we use data from Eurostat comprising yearly
gross regional product for the European Nomenclature of Territorial Units for Statistics
(NUTS 1) regions from 2000 to 2013. Blockbuster delivers a meaningful partition of
the states and regions of the two panels. In particular, the U.S. communities bear close
resemblance to previously published results by Hamilton and Owyang (2012). In the
U.S. dataset we also carry out an out-of-sample validation exercise and show that the
Blockbuster covariance regularisation procedure improves covariance prediction compared
to a number of alternative procedures.

This work is related to several different strands of the literature. First, the literature
on estimation of sparse high-dimensional network and graphical models, see for example
Peng, Wang, Zhou, and Zhu (2009), Belloni, Chen, and Chernozhukov (2016), Barigozzi
and Brownlees (2013), Lauritzen (1996). Second, the literature on stochastic block model
and community detection, see amongst others Abbe, Bandeira, and Hall (2016), Sarkar
and Bickel (2015), Newman (2006) and Arias-Castro and Verzelen (2014). Third, the
literature on identifying latent group structures in econometrics. Some examples of pa-
pers in this area are Hamilton and Owyang (2012), Ando and Bai (2016), Francis et al.
of the methodology we introduce are the straightforward implementation, the fact that
we do not rely on regularised estimation methods (which typically depend on the choice
of penalty parameters) and that we obtain large sample classification consistency guar-
antees. Fourth, the literature on identification of network effects in panel models, see, among others Graham (2015) and de Paula (2017). Last, this paper is related to Gudmundsson (2018a) where the community detection problem is studied in the context of large dimensional Vector Autoregressions (see also Gudmundsson 2018b).

The rest of the paper is organised as follows. Section 2 presents the framework. Section 3 introduces extensions to the baseline methodology. Section 4 carries out a simulation study to assess the finite-sample properties of the procedure. Section 5 uses our community detection methodology to study real activity clustering in the U.S. and Europe. Concluding remarks follow in Section 6.

2 Methodology

2.1 Model

Let \( \{Y_t\} \) with \( Y_t = (Y_{1t}, \ldots, Y_{nt})' \) be an \( n \)-dimensional stationary stochastic process with mean zero and covariance matrix \( \Sigma = \mathbb{E}[Y_t Y_t'] \). We denote its inverse covariance matrix as \( K = \Sigma^{-1} \), which we refer to as the concentration matrix hereafter. In this work we characterise the interdependence structure of the panel with partial correlations. The partial correlation between \( Y_{it} \) and \( Y_{jt} \) measures the linear dependence between the two variables after partialling out the influence of the remaining variables in the panel. We formally define it as

\[
\rho^{ij} = \text{Corr}(e_{it}, e_{jt}),
\]

where \( e_{it} \) and \( e_{jt} \) are the prediction errors of the best linear predictors of \( Y_{it} \) and \( Y_{jt} \), respectively, based on \( \{Y_{st} : 1 \leq s \neq i, j \leq n\} \). It is well known that the linear partial dependence structure of the system is embedded in the concentration matrix (Dempster 1972). It can be shown (Pourahmadi 2013) that the elements of the concentration matrix \( K_{ij} \) are related to the partial correlations through the identity

\[
\rho^{ij} = -\frac{[K]_{ij}}{\sqrt{[K]_{ii}[K]_{jj}}}, \quad i \neq j.
\]
The concentration matrix and the partial correlations of $Y_t$ thus share the same sparsity structure: The $(i,j)$-th element of $K$ is zero if and only if $Y_{it}$ and $Y_{jt}$ have zero partial correlation. If $Y_t$ is Gaussian, a zero partial correlation between $Y_{it}$ and $Y_{jt}$ implies their conditional independence given the remaining components of $Y_t$.

In this work we assume that the partial correlation structure of $Y_t$ is determined by a latent graph. We denote an undirected weighted graph as $G = (V, E, W)$ where $V = \{1, \ldots, n\}$ is the vertex set, $E \subset V \times V$ is the edge set and $W$ is the set of edge weights. For an edge weight $w_{ij} \in W$, we have $w_{ij} \neq 0$ if $(i, j) \in E$ and $w_{ij} = 0$ otherwise. As the graph is undirected we have $w_{ij} = w_{ji}$. It is useful to introduce the adjacency, degree and Laplacian matrices to represent the structure of a graph. The adjacency matrix $A$ of a graph is defined as an $n \times n$ matrix with the $(i,j)$-th element $[A]_{ij}$ equal to $w_{ij}$ if there is an edge between vertices $i$ and $j$ and zero otherwise. Let $d_i$ denote the degree of vertex $i$, that is, $d_i = \sum_{j=1}^{n} [A]_{ij}$. We define the degree matrix $D$ as an $n \times n$ diagonal matrix with element $[D]_{ii} = d_i$. Finally, the symmetric degree-normalised Laplacian is defined as $L = I_n - D^{-1/2}AD^{-1/2}$ where $I_n$ is the $n \times n$ identity matrix. Note that both the adjacency matrix and the Laplacian are symmetric. Throughout the paper we refer to undirected weighted graphs simply as graphs.

We assume that the concentration matrix $K$ is a function of a latent graph $G$ through its Laplacian $L$ and takes the form

$$K = \frac{1}{\sigma^2} I_n + \frac{\phi}{\sigma^2} L,$$  \hspace{1cm} (1)

where $\sigma^2 > 0$ is called the network-variance parameter and $\phi \geq 0$ the network-dependence parameter. Note that this definition guarantees that $[K]_{ij}$ is zero if and only if $i$ and $j$ are not joined by an edge in $G$. The graph thus determines the partial correlation network structure of the random vector $Y_t$. Note that the model also guarantees that the concentration matrix is symmetric and positive definite.

It is informative to consider the regression representation implied by the model (Pourahmadi 2013). If we disallow self-edges for ease of exposition, we may express the $i$-th
variable of $Y_t$ as

$$Y_{it} = \sum_{j \in N(i)} \beta_{ij} Y_{jt} + \epsilon_{it},$$

where $N(i)$ denotes the set vertices that have edges with $i$ and

$$\beta_{ij} = \frac{\phi}{1 + \phi \sqrt{d_i d_j}}, \quad \text{Var}(\epsilon_{it}) = \frac{\sigma^2}{1 + \phi}.$$  

In other words, $Y_{it}$ is given as a linear combination of the series it is partially correlated with, plus an idiosyncratic innovation. Notice that the coefficients $\beta_{ij}$ depend on the edge-weights $w_{ij}$ which allows for a flexible dependence structure in the panel decided by the graph. Notice, that our model has similarities with spatial autoregressive processes Anselin (2007).

In this work we assume that the latent graph $G$ is generated from a random graph model which allows for a community structure. An (inhomogeneous) random graph is defined as a graph in which the vertex set $\mathcal{V}$ is fixed and the existence of an edge in $\mathcal{E}$ is determined by a Bernoulli trial, independent of all other edges. In particular, we rely on a weighted version of the degree-corrected stochastic block model, which is a generalisation of the popular stochastic block model [Holland et al. (1983)].

The stochastic block model is an extension of the Erdős-Rényi random graph (Erdős and Rényi, 1959) in which the vertex set $\mathcal{V}$ is partitioned into $k$ subsets $\mathcal{V}_1, \ldots, \mathcal{V}_k$, typically referred to as communities. An edge is present between vertices $i$ and $j$ with probability $p_s$ if both vertices belong to $\mathcal{V}_s$ and probability $q_{sv}$ if they belong to $\mathcal{V}_s$ and $\mathcal{V}_v$, respectively, with $s \neq v$. Figure 1 gives an example of a stochastic block model with $n = 100$, $k = 5$, $p_s = p = 0.25$ and $q_{sv} = q = 0.01$. In the stochastic block model, all vertices within a given community have the same expected degree. This is often an unrealistic assumption. We therefore assume that $G$ is generated by a degree-corrected stochastic block model [Karrer and Newman (2011)], an extension of the stochastic block model that allows for a general degree distribution. The classical stochastic block model as well as its degree-corrected variant typically generate unweighted graphs where all the non-zero weights $w_{ij}$ are equal to one. This would impose a restrictive structure on the concentration matrix.
To make the model more flexible, we allow the underlying graph $G$ to be weighted so that each off-diagonal element depends on an edge-specific random weight. We call the model presented here a Generalised Stochastic Block Model (GSBM) and formally define it as follows.

**Definition 1 (Generalised Stochastic Block Model).** Let $Z$ be the $n \times k$ community membership matrix, such that $[Z]_{is} = 1$ if vertex $i$ belongs to community $s$ and zero otherwise. Let $B$ be the symmetric $k \times k$ matrix of community-specific edge probabilities and let $\Theta$ be the $n \times n$ diagonal matrix of non-negative, fixed and unknown vertex-specific probability weights. Let $W$ be an $n \times n$ symmetric matrix where each element $[W]_{ij}$ is a random variable supported on the interval $[w, \bar{w}]$ with $0 < w \leq \bar{w}$, and expected value $\mu$.

In a Generalised Stochastic Block Model the probability of an edge between vertices $i$ and $j$ that belong to communities $s$ and $v$, respectively, is $[\Theta]_{ii}[B]_{sv}[\Theta]_{jj} \leq 1$ and all edges are independent. Furthermore, each edge $(i, j)$ is associated with a weight $[W]_{ij}$ which is drawn independently of all other weights, and all edges. It is convenient to write

$$G \sim \text{GSBM}(Z, B, \Theta, W), \quad (2)$$

to indicate that a random graph $G$ is a Generalised Stochastic Block Model.

Notice that the matrix $Z$ defines the community partition $V^k = \{V_1, \ldots, V_k\}$ of the vertex set $V$. We let $n_s = \vert V_s \vert$ denote the size of community $s$. We may assume that the elements of $V_1$ are the first $n_1$ elements of $V$, the elements of $V_2$ the next $n_2$, and so on. The probability of an edge between vertices $i$ and $j$ that belong, respectively, to communities $s$ and $v$ is $[\Theta]_{ii}[B]_{sv}[\Theta]_{jj}$. The matrix $B$ contains the community-specific component of the edge probabilities whereas the matrix $\Theta$ contains the vertex-specific component. This allows vertices that belong to the same community to have different expected degrees. Notice that $\Theta$ and $B$ are only unique up to a multiplicative constant.

We follow Karrer and Newman (2011) and normalise $[\Theta]_{ii}$ such that they sum to one.

---

1. That is, the off-diagonal elements would only depend on the degrees and the parameters $\phi$ and $\sigma^2$.
2. We assume with no loss of generality that there are no empty communities, so that each column of $Z$ has at least one non-zero entry.
within communities, that is, \( \sum_{i \in \mathcal{V}_s} \Theta_{ii} = 1 \) for all \( s = 1, \ldots, k \). Then \( [B]_{sv} \) can be interpreted as the expected number of edges between communities \( s \) and \( v \) if \( s \neq v \) and twice the expected number of edges within community \( s \) if \( s = v \). Last, the matrix \( W \) contains the weights that are associated with the edges in the resulting graph. Notice that we allow the random variables that are the entries of \( W \) to be heterogeneous, but we constrain them to have the same expected value \( \mu \). It is important to emphasise that we assume that the weights are positive. It is possible to allow them to be negative, however, in this case the fraction of negative weights has to be appropriately controlled to ensure the positive definiteness of \( K \). We leave this extension for future research. We also point out that Barigozzi and Brownlees (2013) and Brownlees, Nualart, and Sun (2018) document that the fraction of negative partial correlations estimated in network models for economic panels is negligible.

Figure 1: THE STOCHASTIC BLOCK MODEL

![The figure shows a realisation of a stochastic block model with \( n = 100 \), \( k = 5 \), \( p = 0.25 \) and \( q = 0.01 \).](image)

We conclude by introducing the formal definition of our partial correlation network model which we name the Stochastic Block Partial Correlation Model (SBPCM).
Definition 2 (Stochastic Block Partial Correlation Model). Let \( G \sim \text{GSBM}(Z, B, \Theta, W) \) be a Generalised Stochastic Block Model as in Definition 1. Let \( K \) be the \( n \times n \) concentration matrix corresponding to the random graph \( G \), defined as in (1). In a Stochastic Block Partial Correlation Model, the \( n \)-dimensional stationary stochastic process \( \{Y_t\} \) is such that, for all \( t \), \( Y_t \) has mean zero and covariance matrix \( \Sigma = K^{-1} \).

We also need an assumption on the joint distribution of the process \( \{Y_t\} \) that guarantees that empirical averages are sufficiently well behaved. This is implied by appropriate mixing and tail conditions that are spelled out in Assumption 2 below.

2.2 The Blockbuster Algorithm

Suppose we observe a sample \( Y_1, \ldots, Y_T \) of observations from the model in Definition 2 arranged into the \( T \times n \) matrix \( Y = [Y_1, \ldots, Y_T]' \), and let \( k \) be the number of communities. The community structure of the model is assumed to be unknown, although we assume that \( k \) is known. We adapt spectral clustering techniques (Ng et al., 2001) to detect the communities of the model from the sample in a procedure we call Blockbuster. The proposed algorithm detects the communities using the rescaled eigenvectors of the sample covariance matrix. An important feature of the algorithm is that it allows us to detect the communities without estimating the network structure.

We provide a description of Blockbuster in Algorithm 1. Given the sample and the number of communities \( k \), first construct the \( n \times k \) matrix \( \hat{U} \) of the eigenvectors corresponding to the \( k \) largest eigenvalues of the sample covariance matrix \( \hat{\Sigma} = (1/T) YY' \).

We refer to this matrix as the matrix of partitioning eigenvectors. Then form the matrix \( \hat{X} \) by normalising the rows of \( \hat{U} \) to unit length, that is, \( \hat{X} = \hat{N} \hat{U} \) where \( \hat{N} \) is an \( n \times n \) diagonal matrix with its \( i \)-th element \( [\hat{N}]_{ii} = 1/\| [\hat{U}]_{i\bullet} \| \) and \( \| [\hat{U}]_{i\bullet} \| \) is the Euclidean norm of \( [\hat{U}]_{i\bullet} \), the \( i \)-th row of \( \hat{U} \). The algorithm then applies \( k \)-means clustering to the rows of \( \hat{X} \). The \( k \)-means algorithm partitions a set of data points into \( k \) clusters by solving the optimisation problem

\[
\min_{\{m_1, \ldots, m_k\}} \sum_{i=1}^{n} \min_{s} \| [\hat{X}]_{i\bullet} - m_s \|^2.
\]  

If the \( k \)-th and \( k + 1 \)-th are tied, we may take both.
Let $\mathbf{m}_s^* \in \mathbb{R}^k$ be the vectors that solve it. These vectors are called the centroids of the clusters returned by $k$-means. Each row $[\hat{\mathbf{X}}]_i$ is then assigned to the cluster corresponding to the centroid closest to it. This yields a partition $\hat{\mathcal{V}}^k = \{\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_k\}$ of the vertex set, which is our estimator of the community partition $\mathcal{V}^k$.

Algorithm 1: The Blockbuster Algorithm

**Input:** Sample $\mathbf{Y}_t$ for $t = 1, \ldots, T$, number of communities $k$.

**Procedure:**

1. Compute the sample covariance matrix $\hat{\Sigma}$.

2. Construct the $[n \times k]$ eigenvector matrix $\hat{\mathbf{U}}$ such that its columns are the eigenvectors corresponding to the $k$ largest eigenvalues of $\hat{\Sigma}$.

3. Standardise each row of $\hat{\mathbf{U}}$ by its norm and denote the row-normalised eigenvector matrix by $\hat{\mathbf{X}}$, so that $[\hat{\mathbf{X}}]_{ij} = [\hat{\mathbf{U}}]_{ij}/\|\hat{\mathbf{U}}_i\|$. 

4. Apply the $k$-means algorithm to the rows of $\hat{\mathbf{X}}$.

**Output:** Return the $k$-means partition $\hat{\mathcal{V}}^k = \{\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_k\}$ as the estimate of the community structure.

### 2.3 Theory

In this section we bound the fraction of vertices that are incorrectly clustered by Blockbuster. We show that the fraction is close to zero with probability close to one when the cross-sectional dimension $n$ and the number of observations $T$ are large enough, provided that $n/T$ is small, under the assumption that the process $\{\mathbf{Y}_t\}$ is strongly mixing and has exponentially small tail probabilities.

We introduce additional notation that is used throughout this work. Let $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$ denote the spectral and Frobenius norms of the $n \times n$ matrix $\mathbf{A}$, respectively. We denote the $i$-th smallest eigenvalue of $\mathbf{A}$ as $\lambda_i(\mathbf{A})$ and the set of eigenvalues of $\mathbf{A}$ that are in the interval $S$ as $\lambda_S(\mathbf{A})$. Notice that we use the convention $\lambda_1(\mathbf{A}) \leq \cdots \leq \lambda_n(\mathbf{A})$. We

---

4 The solution is not necessarily unique. We let $\mathbf{m}_s^*$ be some set of vectors that achieve the minimum.

5 It is well known that the exact solution is NP hard and that the standard approximation to it is prone to finding local minima, which may be arbitrarily far off from the exact solution. We use a variant called $k$-means++ (Arthur and Vassilvitskii 2007), which guarantees an approximation within a factor of $O(\log k)$ of the optimum. For simplicity of discussion, we assume that the minimum is computed. All theoretical findings remain true for $k$-means++.
refer to the eigenvectors corresponding to the $k$ largest (smallest) eigenvalues of $A$ as its $k$ top (bottom) eigenvectors, counting multiplicities. By $f(n) = O(g(n))$ we mean that there exist a real number $M > 0$ and a real number $n_0$ such that for all $n \geq n_0$ we have $|f(n)| \leq M|g(n)|$. Similarly, $f(n) = \Omega(g(n))$ means $|f(n)| \geq M|g(n)|$ for all $n \geq n_0$ and $M > 0$. We refer to an event that happens with probability approaching one as an event that happens with high probability.

Define $A = \mathbb{E}[A]$ as the population adjacency matrix of the graph $G$. Notice that if $G \sim \text{GSBM}(Z, B, \Theta, W)$, we may decompose the population adjacency matrix as $A = \mu \Theta Z B Z' \Theta$. We also define the population degree matrix $D$ as the diagonal matrix with $[D]_{ii} = \sum_{j=1}^{n} [A]_{ij}$, the population normalised Laplacian $L = I_n - D^{-1/2} A D^{-1/2}$ and the population concentration matrix $K = \frac{1}{\sigma^2} I_n + \frac{\phi}{\sigma^2} L$, analogously to $D$, $L$ and $K$. In this section we state results in terms of the sample concentration matrix $\hat{K} = \hat{\Sigma}^{-1}$, with $\hat{U}$ being the matrix of its bottom $k$ eigenvectors. Let $U$ and $U'$ be the $n \times k$ matrices of the bottom $k$ eigenvectors of $K$ and $K'$, respectively. Define $N$ and $N'$ as $n \times n$ diagonal matrices with $[N]_{ii} = \|[U]_i\|^{-1}$ and $[N']_{ii} = \|[U']_i\|^{-1}$, respectively. Finally, let $X = NU$ and $X' = N'U'$ be the row-normalised counterparts of $U$ and $U'$.

To bound the fraction of misclustered vertices, we follow the strategy of Rohe et al. (2011) and Qin and Rohe (2013). We begin by noting that the $k$-means objective function of (3) can be written as

$$\min_{\{m_1, \ldots, m_k\}} \sum_{i=1}^{n} \min_s \|\hat{X}_i - m_s\|^2 = \min_{M \in \mathcal{M}(n,k)} \|\hat{X} - M\|^2_F,$$

where $\mathcal{M}(n,k) = \{M \in \mathbb{R}^{n \times k} : M \text{ has no more than } k \text{ different rows}\}$. Let the estimated centroid matrix $\hat{C}$ be defined as

$$\hat{C} = \arg\min_{M \in \mathcal{M}(n,k)} \|\hat{X} - M\|^2_F. \quad (4)$$

Its $i$-th row $[\hat{C}]_i$ is equal to the $k$-means centroid that is closest to row $i$ of the eigenvector matrix $\hat{X}$, so that $[\hat{C}]_i \in \{m^*_1, \ldots, m^*_k\}$. It is clear that the $k$-means centroid matrix has no more than $k$ different rows. Let the population centroid matrix $C$ be defined
analogously to (4) with \( \hat{X} \) replaced with \( XO \), where \( O \) is a \( k \times k \) orthonormal rotation matrix that depends on \( \hat{X} \) and \( X \). This matrix is discussed in detail in Theorem 2. We adopt the same definition of misclustered vertices as [Rohe et al. (2011)] and say that vertex \( i \) is correctly clustered if \([\hat{C}]_{i \bullet} \) is closer to \([C]_{i \bullet} \) than any other population centroid \([C]_{j \bullet} \) for \( j \neq i \). Define the set

\[
\mathcal{M} = \left\{ i : \| [\hat{C}]_{i \bullet} - [C]_{i \bullet} \| \geq \sqrt{1/2} \right\}.
\] (5)

The condition \( \| [\hat{C}]_{i \bullet} - [C]_{i \bullet} \| \geq \sqrt{1/2} \) is necessary for vertex \( i \) to be misclustered, as shown and discussed in more detail in Lemma 2. The size of the set \( \mathcal{M} \) is thus an upper bound of the number of misclustered vertices.

Theorem 1 bounds the fraction \( |\mathcal{M}|/n \). The theorem is in the spirit of Theorem 3.1 from [Rohe et al. (2011)] and Theorem 4.4 of [Qin and Rohe (2013)]. The theorem shows that the fraction of misclassified vertices is close to zero with high probability when \( \log n/(n\rho_n) \) and \( n/T \) are close to zero. Some technical conditions under which the theorem is established are detailed in Assumptions 1 and 2 below.

**Theorem 1 (Misclustered Vertices).** Consider an SBPCM as in Definition 2. Let \( \mathcal{M} \) be as in (5) and define \( \rho_n = \min_{i,j} |B|_{i,j} \). Suppose Assumptions 1 and 2 hold.

If \( T = \Omega \left( n^{2/\gamma - 1} \right) \), we have

\[
\frac{|\mathcal{M}|}{n} = O \left( \frac{n}{T} + \frac{\log n}{n\rho_n} \right),
\]

with high probability, where \( \gamma \) is defined in Assumption 2.

The strategy we employ to establish Theorem 1 is the following. Lemma 1 abstracts from all randomness and shows that Blockbuster recovers the partition \( V^k \) trivially from the eigenvectors of the population concentration matrix \( \mathcal{K} \). Theorem 2 shows the concentration of the first \( k \) eigenvectors of \( \hat{\mathcal{K}} \) around those of \( \mathcal{K} \) with high probability. To this end, we make use of an extension of the random matrix concentration results from [Oliveira (2009)] to allow for weighted graphs and a generalisation of the sample covariance matrix concentration inequalities from [Vershynin (2012)] which allow for strongly mixing
data with generalised exponential tails, by virtue of the results of Merlevêde, Peligrad, and Rio (2011). Using these results, we can prove Theorem 1. Proofs of the statements in this section can be found in Appendix A.

To establish Theorem 1 we require the following random graph assumptions.

**Assumption 1.** Let $K$ be as in (1) with the random graph $G$ as in Definition 1. We assume the following.

(i) $B$ is symmetric and positive definite.

(ii) $\rho_n = \Omega(\log(n)/n)$ where $\rho_n = \min_{ij}|B|_{ij}$.

Assumption (i) requires $B$ to be positive definite, and hence invertible so that all $k$ communities are distinguishable. As an example, consider the simple version of the model where $p_s = p$ and $q_{sv} = q$ for all $s \neq v$. Then this assumption simply requires $p > q$. It is natural to characterise the behaviour of a random graph when $n$ grows as a function of the edge probabilities in $B$. Assumption (ii) requires the probabilities to be at least of the order $\log(n)/n$, which is termed the semi-sparse regime by Sarkar and Bickel (2015). The regime where the probabilities are constant results in a dense graph, whereas regimes where the probabilities are falling in $n$ yield less dense graphs. We make a distinction between the sparse regime where the probabilities are proportional to $1/n$ and the semi-sparse regime, which is among the sparsest regimes where exact recovery of the communities is possible (Abbe et al., 2016). Furthermore, norm convergence of graph matrices is well established in this regime, see (Oliveira, 2009; Chung and Radcliffe, 2011).

The following lemma is an extension of Lemma 3.3 from Qin and Rohe (2013). It establishes the form of the row-normalised eigenvector matrix $X$.

**Lemma 1.** Let $G \sim \text{GSBM}(Z, B, \Theta, W)$ be a Generalised Stochastic Block Model as in Definition 1. Let $K$ be the population concentration matrix, $U$ the matrix of its bottom $k$ eigenvectors and $X$ the row-normalised counterpart of $U$.

---

6 In fact, Algorithm 1 can easily handle the case where $B$ is negative definite. This corresponds to a bipartite graph with a heterophilic community structure, where there are few edges within the communities but many across them (Rohe et al., 2011), but we do not pursue this here.
Then \( \lambda_i(K) = (1 + \phi)/\sigma^2 \) for all \( i = k + 1, \ldots, n \) and \( \lambda_i(K) \in [1/\sigma^2, (1 + \phi)/\sigma^2) \) for \( i = 1, \ldots, k \). Furthermore, there exists a \( k \times k \) orthonormal matrix \( V \) such that \( X = ZV \).

In particular, the lemma implies \( [X]_{ij} = [X]_{lj} \) if and only if \( [Z]_{ij} = [Z]_{lj} \), so the rows corresponding to two vertices that belong to the same community are equal in \( X \). Hence there are only \( k \) different rows in \( X \) and \( k \)-means can trivially recover the partition \( \mathcal{V}^k \) from \( X \) by selecting each of the different rows as centroids.

To derive the concentration results for the sample concentration matrix \( \hat{K} \) of the process \( \{Y_t\} \) and its eigenvectors, we require appropriate mixing and distributional assumptions. We derive the results conditional on the random graph \( G \). This is justified by the fact that observing a single random graph suffices in random graph concentration results like those of Oliveira (2009), provided that \( n \) is large enough. We formulate our assumptions on the sequence of isotropic random vectors \( \Sigma^{-1/2}Y_t \).

Let \( B_{-\infty}^r \) and \( B_{r+m} \) be the \( \sigma \)-algebras generated by \( \{\Sigma^{-1/2}Y_t : -\infty \leq t \leq r\} \) and \( \{\Sigma^{-1/2}Y_t : r+m \leq t \leq \infty\} \), respectively. Define the \( \alpha \)-mixing coefficients of the process as

\[
\alpha(m) = \sup_{r} \sup_{A \in B_{-\infty}^r, B \in B_{r+m}} \left| \Pr(A \cap B|\mathcal{G}) - \Pr(A|\mathcal{G}) \Pr(B|\mathcal{G}) \right|.
\]

We assume the \( \{Y_t\} \) process satisfies the following assumptions.

**Assumption 2.** Let \( \{Y_t\} \) be a zero-mean stationary process with covariance matrix \( \Sigma = \mathbb{E}[Y_tY_t'] \).

(i) \( \{\Sigma^{-1/2}Y_t\} \) is strongly mixing with mixing coefficients satisfying \( \alpha(m) \leq e^{-c_1m\gamma_1} \) for any positive integer \( m \) where \( \gamma_1, c_1 > 0 \) are constants.

(ii) For any vector \( x \) with \( \|x\| = 1 \) and for any \( s > 0 \) we have

\[
\sup_{t > 0} \Pr\left( |x'\Sigma^{-1/2}Y_t| > s |\mathcal{G} \right) \leq c_3e^{-(s/c_2)^{\gamma_2}},
\]

where \( \gamma_2, c_2, c_3 > 0 \) are constants.

(iii) \( \gamma < 1 \) where \( 1/\gamma = 1/\gamma_1 + 1/\gamma_2 \).

Notice that \( \{\Sigma^{-1/2}Y_t\} \) is a sequence of isotropic random vectors as \( \mathbb{E}[\Sigma^{-1/2}Y_t\Sigma^{-1/2}Y_t'] = I_n \).
Assumptions (i)-(iii) allow us to apply the concentration inequality of Merlevède et al. (2011) to the sample covariance matrix \( \hat{\Sigma} \), which plays a key role in the following theorem. Assumption (ii) imposes generalised exponential tails on the distribution of the isotropic vectors \( \Sigma^{-1/2} Y_t \). This can be seen as a generalisation of the definition of sub-Gaussian vectors in Vershynin (2012). In particular, the assumption implies that all the elements of the vector have generalised exponential tails similar to what Fan, Liao, and Mincheva (2013) assume.

Theorem 2 establishes the concentration of the eigenvectors of \( \hat{K} \) around those of \( K \).

**Theorem 2 (Concentration).** Let \( Y_t \) for \( t = 1, \ldots, T \) be observations from an SBPCM as in Definition 2. Let \( \hat{U} \) the matrix of the bottom \( k \) eigenvectors of the sample concentration matrix \( \hat{K} \) and \( \hat{X} \) its row-normalised counterpart. Let \( U \) and \( X \) be as in Lemma 1.

If \( T = \Omega \left( \frac{n^2}{\gamma - 1} \right) \), there exists a \( k \times k \) orthonormal rotation matrix \( O \), that depends on \( \hat{X} \) and \( X \), such that

\[
\left\| \hat{X} - XO \right\| = O \left( \frac{n}{\sqrt{T}} + \sqrt{\frac{\log n}{\rho_n}} \right),
\]

with high probability.

The theorem shows how the sample row-normalised eigenvectors \( \hat{X} \) concentrate around their population analogue \( X \) with high probability, up to a rotation.\(^8\)\(^9\)

The estimated centroid matrix \( \hat{C} \) was defined in 4. We similarly define

\[
\mathcal{C} = \arg\min_{M \in \mathcal{M}(n,k)} \| XO - M \|_F^2,
\]

where the orthonormal rotation is included because \( \hat{X} \) concentrates around the rotated population eigenvectors \( XO \) as shown in Theorem 2. For the GSBM of Definition 1, we have \( C = XO = ZV O \). It follows that \( [C]_{i*} = [Z]_{i*}V O \) and we may thus equivalently define vertex \( i \) as correctly clustered if \( [\hat{C}]_{i*} \) is closer to \( [Z]_{i*}V O \) than any other rotated

\(^8\)The rotation is required because \( \hat{K} = K \) does not necessarily imply \( \hat{U} = U \). At the very least, the columns might be permuted.

\(^9\)The row-normalised eigenvectors do not necessarily converge under our assumptions, but this result is sufficient to establish Theorem 1.
population centroid \( [Z]_j \cdot V \cdot O, j \neq i \). The next lemma provides a condition that implies that vertex \( i \) is correctly clustered. This statement is established in Theorem 4.4 in Qin and Rohe (2013) (see also Rohe et al. (2011)).

**Lemma 2.** Let \( \hat{C} \) be the estimated centroid matrix as in (4). Let \( V \) and the community membership matrix \( Z \) be as in Lemma 1 and let the orthonormal rotation \( O \) be as in Theorem 2.

Then \( \| \hat{C}_{i\cdot} - [Z]_{i\cdot} V \cdot O \| < \| \hat{C}_{j\cdot} - [Z]_{j\cdot} V \cdot O \| \) holds for all \( j \neq i \) that satisfy the condition \( \| \hat{C}_{i\cdot} - [Z]_{i\cdot} V \cdot O \| < \sqrt{1/2} \).

The lemma justifies bounding the number of misclustered vertices by the number of nodes that do not satisfy the condition \( \| \hat{C}_{i\cdot} - [Z]_{i\cdot} V \cdot O \| < \sqrt{1/2} \), which is sufficient for vertex \( i \) to be correctly clustered. The set \( M \) from (5) hence contains the set of all misclustered nodes. With this, all the tools for proving Theorem 1 have been established.

### 2.4 Discussion

A number of comments on the SBPCM of Definition 2 and the Blockbuster algorithm are in order.

The model bears some similarities to a factor model with a block structure, that is, a model in which the variables are generated by a factor model and each variable loads on exactly one factor (Hallin and Liska 2011; Jackson, Kose, Otrok, and Owyang, 2016). However, a block factor model is not equivalent to a SBPCM in general. It is typically assumed that correlation across blocks is negligible so that the covariance matrix is approximately diagonal in block factor models. Assume, for ease of exposition, that \( p_s = p \) and \( q_{sv} = q \). Then, a block factor model corresponds to \( q \) close to or equal to zero. On the contrary, we only require \( p > q \) in our framework implying that there may be a high correlation between blocks. Furthermore, notice that a block factor model is associated with a partial correlation network in which each community is a clique, that is, a sub-graph where every possible pair of vertices is joined by an edge. Our model can then replicate such network structures with \( p = 1 \) and \( q = 0 \), but it also allows for structures where the communities are much more sparse.
Alternative approaches to detect communities when the graph structure is unknown have been considered in the literature. One way is to estimate the partial correlation network and then apply spectral clustering to the estimated network Laplacian. The estimation of the graphical structure of the data is typically carried out using LASSO techniques, which requires the selection of a tuning parameter that determines the sparsity of the estimated network (Bühlmann and van de Geer, 2011; Meinshausen and Bühlmann, 2006; Peng et al., 2009). Note that Meinshausen (2008) shows that LASSO may not recover the sparsity pattern consistently for some classes of graphical models, see also Meinshausen and Yu (2009). A highlight of our approach is that it allows us to learn the community structure without estimating the structure of the network.

Finally, it should be noted that the sample covariance estimator in Blockbuster may be substituted with any other covariance matrix estimator, for example a shrinkage estimator. The first term in the bound in Theorem 1 is a function of the convergence rate of the sample covariance matrix. By resorting to a shrinkage estimator for the covariance, one may obtain a faster rate of convergence. Notice however that sparsity in the concentration matrix does not imply sparsity in the covariance matrix.

3 Extensions

3.1 Community Detection in the Presence of Common Factors

We introduce an extension of the SBPCM in which the components of the panel $Y_t$ are influenced by a set of common factors, and a modification of the Blockbuster algorithm that consistently detects communities in this setting.

Let $F_t$ be an $R$-dimensional vector of common factors with mean zero and covariance matrix $I_R$ and let $q_r$ be $n$-dimensional fixed vectors of factor loadings for $r = 1,\ldots, R$. We assume that the random vector $Y_t$ is generated as

$$Y_t = QF_t + \epsilon_t, \quad (7)$$
where \( Q = [q_1, q_2, \ldots, q_R] \) is an \( n \times R \) matrix of factor loadings and \( \epsilon_t \) is generated by an SBPCM as in Definition 2, with \( \mathbb{E}[\epsilon_t | F_t] = 0 \) and covariance matrix \( \text{Cov}(\epsilon_t | F_t) = \Sigma_\epsilon \).

We may assume without loss of generality that \( \|q_1\| \geq \|q_2\| \geq \ldots \geq \|q_R\| \).

We define the SBPCM with common factors as follows.

**Definition 3.** Let \( G \sim \text{GSBM}(Z, B, \Theta, W) \) be a Generalised Stochastic Block Model as in Definition 1. Let \( K_\epsilon \) be the \( n \times n \) concentration matrix corresponding to the random graph \( G \), defined as in (1). In an SBPCM with common factors, the \( n \)-dimensional stationary stochastic process \( \{Y_t\} \) defined as in (7) has mean zero and covariance matrix

\[
\Sigma = K_\epsilon^{-1} + \sum_{r=1}^{R} q_r q_r'.
\] (8)

Consider a sample drawn from the model in Definition 3 satisfying Assumption 2, arranged into a \( T \times n \) matrix \( Y = [Y_1, \ldots, Y_T]' \). Under appropriate conditions, with high probability, the top \( R \) eigenvectors of \( \hat{\Sigma} = (1/T)Y'Y \) correspond to the \( R \) factors and the next \( k \) eigenvectors are the partitioning eigenvectors. This motivates Algorithm 2:

Given \( Y \) and \( k \), first construct the \( n \times k \) matrix \( \hat{U} \) such that it contains the eigenvectors corresponding to the \((R+1)\)-th largest to the \((R+k)\)-th largest eigenvalues of the sample covariance matrix \( \hat{\Sigma} \). Then form the row-normalised eigenvector matrix \( \hat{X} \) as in Algorithm 1 and apply \( k \)-means clustering to its rows.

**Algorithm 2** The Blockbuster Algorithm in the Presence of Common Factors

**INPUT:** Sample \( Y_t \) for \( t = 1, \ldots, T \), number of communities \( k \), number of factors \( R \).

**PROCEDURE:**

1. Compute the sample covariance matrix \( \hat{\Sigma} \).
2. Construct the \( n \times k \) eigenvector matrix \( \hat{U} \) such that its columns are the eigenvectors corresponding to the \((R+1)\)-th to the \((R+k)\)-th largest eigenvalues of \( \hat{\Sigma} \).
3. Standardise each row of \( \hat{U} \) by its norm and denote the row-normalised eigenvector matrix by \( \hat{X} \), so that \( [\hat{X}]_{ij} = [\hat{U}]_{ij} / \| [\hat{U}]_{i*} \| \).
4. Apply the \( k \)-means algorithm to the rows of \( \hat{X} \).

**OUTPUT:** Return the \( k \)-means partition \( \hat{V}^k = \{\hat{V}_1, \ldots, \hat{V}_k\} \) as the estimate of the community structure.
It is convenient to state the results of this section in terms of the concentration matrices rather than covariances. Let $K\epsilon$ be the population concentration matrix of $\epsilon$ from (7) and $K$ be the population concentration matrix of $Y$. Let $u_i(K)$ and $u_i(K\epsilon)$ for $i = 1,\ldots,n$ be their eigenvectors. Define $U\epsilon$ as the $n \times k$ matrix of the bottom $k$ eigenvectors of $K\epsilon$ and $U$ as the matrix of the eigenvectors corresponding to the ($R+1$)-th smallest to the ($R+k$)-th smallest eigenvalues of $K$, so that $[U]_{ii} = u_{R+i}(K)$ for $i = 1,\ldots,k$, where $[U]_{ii}$ refers to the $i$-th column of $U$.

We impose the following assumptions on the factor loadings.

**Assumption 3.** Let the process $\{Y_t\}$ be an SBPCM as in Definition 3 satisfying Assumption 2. We assume the following.

(i) $q^r_qv = 0$ for all $r,v = 1,\ldots,R$, $r \neq v$.

(ii) $q^r[u_i]_{ii} = 0$ for all $r = 1,\ldots,R$ and $i = 1,\ldots,k$.

(iii) $\|q_R\|^2 > \frac{\sigma^2_\phi}{1+\phi}$.

The first assumption requires the factor loadings to be mutually orthogonal. The second implies that the factor loading vectors carry no information on the community structure nor the degrees of the underlying graph $G$. The third guarantees that the factors are strong enough to be dominant in the spectrum of $K$ by assuming that $\|q_R\|^2$ is larger than a constant which does not depend on $n$. Notice that this assumption is trivially satisfied for large enough $n$ when the factors are strong and the norms diverge to infinity, as in for example Bai and Ng (2002) and Bai (2003). This assumption is not restrictive in the case of weak factors either, where the norms of the factor loading vectors are bounded, see for example Onatski (2012). For further discussion on the difference between strong and weak factors, refer to Chudik, Pesaran, and Tosetti (2011).

We now show that Algorithm 2 recovers the communities with high probability under suitable conditions. Theorem 3 extends Theorem 1 to bound the fraction $|\mathcal{M}|/n$ under the model in Definition 3.
Theorem 3 (Misclustered Vertices II). Consider an SBPCM with common factors as in Definition 3. Let $\mathcal{M}$ be defined analogously to (5) for this model and define $\rho_n = \min_{ij} |B|_{ij}$. Suppose Assumptions 1, 2 and 3 hold.

If $T = \Omega \left( n^{2/\gamma - 1} \right)$, we have

$$\frac{|\mathcal{M}|}{n} = O \left( \frac{n}{T} \|\Sigma\|^2 + \frac{\log n}{n\rho_n} \right),$$

with high probability.

We follow a similar strategy to Section 2 to establish Theorem 3. Lemma 3 extends the population results of Lemma 1 to the model in Definition 3.

Lemma 3. Let $\mathcal{G} \sim \text{GSBM} (Z, B, \Theta, W)$ be a Generalised Stochastic Block Model as in Definition 1. Let $\mathcal{K}$ be the population concentration matrix, $U$ be the matrix of its $(R+1)$-th to $(R+k)$-th bottom eigenvectors and $\mathcal{X}$ the row-normalised counterpart of $U$.

Then $\lambda_i(\mathcal{K}) = 1/(\sigma^2/(1+\phi) + \|q_i\|^2)$ for $i = 1, \ldots, R$, $\lambda_i(\mathcal{K}) \in [1/\sigma^2, (1+\phi)/\sigma^2)$ for $i = R+1, \ldots, R+k$ and $\lambda_i(\mathcal{K}) = (1+\phi)/\sigma^2$ for all $i = R+k+1, \ldots, n$. Furthermore, there exists a $k \times k$ orthonormal matrix $V$ such that $\mathcal{X} = \mathcal{Z}V$.

The eigenvectors $\mathcal{X}$ thus allow $k$-means to recover the community partition, by a similar logic to before. Arguments analogous to those in Section 2 then allow us to prove Theorem 3.

3.2 Community Structure and Covariance Estimation

The presence of a community structure in a panel suggests a natural covariance estimation strategy. The covariance matrix can be estimated by a block covariance matrix estimator where the diagonal blocks contain the sample covariance matrices of each community, while thresholding is applied to the off-diagonal blocks (Pourahmadi 2013). We provide a description of an estimator based on this idea in Algorithm 3. Notice that in our implementation we propose to choose a different thresholding parameter for each block of the covariance matrix. This is a reasonable compromise between universal thresholding.
(where a single threshold level is chosen for all entries of the covariance matrix) and
adaptive thresholding (where a different threshold level is chosen for each entry) for
panels exhibiting a block covariance structure. In the empirical application we choose the
threshold parameters by using an adaptation of the cross-validation criterion of [Bickel and
Levina (2008)] adapted to dependent data. It is important to emphasise that such criteria
does not ensure that the resulting covariance estimator is positive definite. However, in
our empirical exercise we found this not to be a concern for reasonable ranges of the
thresholding parameters.

Algorithm 3 Blockbuster Covariance Estimator

**INPUT:** Sample $Y_t$ for $t = 1, \ldots, T$, number of communities $k$, threshold parameters $\lambda_{ij}$
for $1 \leq i \neq j \leq k$.

**PROCEDURE:**

1. Run the Blockbuster algorithm and obtain an estimate of the community partition
   of the panel $\hat{V}^k = \{\hat{V}_1, \ldots, \hat{V}_k\}$.

2. Reorder the series in the panel so that the first $\hat{n}_1 = |\hat{V}_1|$ series are the ones in
   community $\hat{V}_1$, the following $\hat{n}_2$ series are the ones in $\hat{V}_2$, and so on.

3. Let $\hat{\Sigma}_{sv}$ denote the $\hat{n}_s \times \hat{n}_v$ sample covariance matrix of the series in community $\hat{V}_s$
   with the series in community $\hat{V}_v$

4. The Blockbuster covariance estimator $\hat{\Sigma}_B$ is defined as

\[
\hat{\Sigma}_{Bsv} = \begin{cases} 
\hat{\Sigma}_{sv} & s = v, \\
T_{\lambda_{sv}}(\hat{\Sigma}_{sv}) & \text{otherwise,}
\end{cases}
\]

where $\hat{\Sigma}_{Bsv}$ denotes the $(s, v)$-th $\hat{n}_s \times \hat{n}_v$ block of $\hat{\Sigma}_B$, and $T_{\lambda_{sv}}(\hat{\Sigma}_{sv})$ returns $\hat{\Sigma}_{sv}$
with the elements that are less than $\lambda_{sv}$ in absolute value set to zero.

**OUTPUT:** Return the Blockbuster covariance $\hat{\Sigma}_B$.

In case there are common factors present in the panel, as in the model in Definition 3,
we may employ a regularisation approach similar to POET (Fan et al., 2013). Let $\lambda_i(\hat{\Sigma})$
and $u_i(\hat{\Sigma})$ denote the $i$-th eigenvalue and eigenvector of the sample covariance matrix.
Then we can regularise the sample covariance of the data using

\[
\hat{\Sigma}_B = \sum_{i=n-R+1}^{n} \lambda_i(\hat{\Sigma})u_i(\hat{\Sigma})u_i'(\hat{\Sigma}) + \hat{R}_B,
\]
where $R$ is the number of factors and $\hat{R}_B$ is the Blockbuster covariance estimator applied to the residual matrix $\hat{R} = \sum_{i=1}^{n-R} \lambda_i(\hat{\Sigma}) u_i(\hat{\Sigma}) u_i'(\hat{\Sigma})$. Notice that we apply block-by-block thresholding to the off diagonal blocks of the residual covariance matrix whereas the standard POET estimator applies universal thresholding to all the off diagonal elements.

In the empirical application section of this paper we use this estimator to validate the community partition estimated by Blockbuster out-of-sample. Analysing the theoretical properties of the estimator is not the focus of this work however, so we leave it open for future research.

4 Simulation Study

In this section we carry out a simulation study to investigate the finite-sample properties of the Blockbuster algorithm.

For illustrative purposes, we begin by drawing a single sample from the SBPCM with $R = 1$ common factors and $k = 2$ communities for $T = 500$ and $n = 100$. We draw a GSBM with $k = 2$ communities where each community has size $n/k$. The edge probabilities are set to $p_s = p = 0.5$ for all $s = 1, \ldots, k$ and $q_{vr} = q = 0.01$ for all $v, r = 1, \ldots, k$, $v \neq r$. The network-dependence parameter $\phi$ is set to 20 while the network variance $\sigma^2$ is 1. We draw $[\Theta]_{ii}$ from a power law distribution $f(x) = \alpha x_m^\alpha / x^{\alpha+1}$ for $x \geq x_m$ with parameters $x_m = 0.75$ and $\alpha = 2.5$. The edge-weights are drawn uniformly in the interval $[0.3, 1]$. We draw the data identically and independently from a multivariate Gaussian with covariance matrix given as in (8), where the factor loadings $q$ are generated from a standard normal.

The first panel of Figure 2 displays a heatmap of the correlation matrix of the panel conditional on the factor with the series ordered by the true community partition. The second and third panels display the corresponding sample correlation matrix when the series are randomly shuffled and when they are ordered by the community partition estimated by Blockbuster. The figure shows that the Blockbuster algorithm detects the communities accurately. In particular, in this specific replication there are no misclustered
Figure 2: HEATMAPS OF SIMULATED DATA

The first panel of the figure displays a heatmap of the correlation matrix (conditional on the factor) with the series ordered by the true community partition. The second and third panels display heatmaps of the sample correlation matrix where the series are randomly shuffled and where they are ordered according to the estimated community partition, respectively. The data are simulated with $n = 100$, $T = 500$ and $k = 2$.

Next, we draw repeated samples of size $T$ on $n$ vertices from the SBPCM with $R = 1$ common factors and $k = 5$ communities. The edge probabilities are set to $p_s = p = c_p (\log n)^{1.01} / n$ for all $s = 1, \ldots, k$ and $q_{vr} = q = c_q (\log n)^{1.01} / n$ for all $v, r = 1, \ldots, k$, $v \neq r$. We calibrate $c_p$ and $c_q$ so that when $n = 100$ the $(p, q)$ pair is equal to $(0.25, 0.01)$, $(0.25, 0.05)$ or $(0.50, 0.01)$. Note that as the edge probabilities are functions of $n$, varying $n$ changes the probabilities. We draw $[\Theta]_{ii}$ once from a power law distribution with parameters as before and keep the weights fixed across replications. We similarly draw the factor loadings $q$ as before and keep them fixed over repetitions. The rest of the parameters are set as before.

We then apply the Blockbuster algorithm to the sample to recover the communities. To measure the quality of the Blockbuster partition $\hat{V}^k$ we compare it to $V^k$ by calculating the fraction of correctly classified vertices, which we call hit ratio. Note that $\hat{V}^k$ only estimates $V^k$ up to a permutation. As the number of communities $k$ is low in our simulations, we calculate the hit percentage for every possible permutation and select the maximum as the final hit ratio figure.

We repeat the Monte Carlo experiment 1000 times for different values of $n$ (50, 100 and 200) and $T$ (50, 100, 200, 500 and 1000). The results are summarised in Table I.

---

10 We have also simulated from the model without common factors and the results are similar.
Table 1: Hit Ratio of Blockbuster

<table>
<thead>
<tr>
<th>Panel A: $\phi = 5$</th>
<th>Panel B: $\phi = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 50$</td>
</tr>
<tr>
<td>$T = 50$</td>
<td>100</td>
</tr>
<tr>
<td>$p/q$</td>
<td>$n = 50$</td>
</tr>
<tr>
<td>0.25/0.01</td>
<td>65.5%</td>
</tr>
<tr>
<td>0.50/0.01</td>
<td>82.4%</td>
</tr>
<tr>
<td>0.25/0.05</td>
<td>40.9%</td>
</tr>
<tr>
<td></td>
<td>$n = 100$</td>
</tr>
<tr>
<td>0.25/0.01</td>
<td>49.0%</td>
</tr>
<tr>
<td>0.50/0.01</td>
<td>63.6%</td>
</tr>
<tr>
<td>0.25/0.05</td>
<td>33.8%</td>
</tr>
<tr>
<td></td>
<td>$n = 200$</td>
</tr>
<tr>
<td>0.25/0.01</td>
<td>37.2%</td>
</tr>
<tr>
<td>0.50/0.01</td>
<td>46.1%</td>
</tr>
<tr>
<td>0.25/0.05</td>
<td>29.2%</td>
</tr>
</tbody>
</table>

The table reports hit ratios for Blockbuster applied to simulated data from the SBPCM with common factors of Definition 7. The probabilities in the rows correspond to $n = 100$.

Panel A of Table 1 shows the results with network dependence parameter $\phi = 5$, whereas panel B shows the results when $\phi = 50$. The first two rows of panel A in Table 1 show that Blockbuster performs quite well when $n = 50$. If the probability of within-community edges is high as in the second row, the algorithm has a 82% hit ratio even with a small sample of $T = 50$. Comparing the second to third rows, it is clear that the results are better if $p$ is large relative to $q$. It is also clear that the results with $n = 100$ and $n = 200$ are only better than those with $n = 50$ when the sample size is larger, as in the last column of Panel B. For larger $n$, more samples are required to consistently estimate the covariance matrix and thus the community partition as the results of Theorem 1 suggest. Comparing the results in panels B with A shows that the results improve across the board with stronger network dependence.

5 Empirical Application

We use the methodology developed in the paper to study two panels of real activity growth measures for the U.S. and Europe. Our objective is to partition the series in the panels into communities characterised by a high degree of interdependence in real activity growth.
5.1 Community Detection in the U.S.

For our first application we consider a dataset constructed by Hamilton and Owyang (2012). The data consists of quarterly growth rates of payroll employment for the U.S. states (excluding Alaska and Hawaii) from the second quarter of 1956 to the fourth of 2007, which results in a panel of $n = 48$ time series over $T = 207$ periods. The data are seasonally adjusted and annualised. See Hamilton and Owyang (2012) for further details on the dataset.

Hamilton and Owyang (2012) use their dataset to study business cycle synchronicity. We summarise their findings for the purpose of comparison, even though they deal with a research question different from the one considered here. The authors introduce a Bayesian methodology based on a Markov-switching model to cluster states into communities with similar business cycle timing. They also propose a cross-validation procedure to estimate the number of communities and find evidence of three clusters. We focus on the results where they use a set of exogenous state level characteristics in addition to the employment growth rate. They find (see right column of figure 3 in Hamilton and Owyang, 2012) that the states are partitioned, roughly speaking, into a cluster of oil-producing and agricultural states, a cluster containing several East Coast states together with California, Arizona and Colorado, and a cluster containing the remaining states. Note that our results are not directly comparable with theirs, as we provide point estimates while they provide community membership posterior probability. Moreover, they use more information than us as they also take advantage of exogenous explanatory variables. It is also worth pointing out that in their results some states have a low posterior probability of belonging to any cluster, whereas in our algorithm every state is assigned to a community.

11 We winsorise the data when the growth rates are larger than 20% in absolute value. There are only three such observations in the sample and they all belong to West Virginia. See also Hamilton and Owyang (2012).
5.1.1 In-Sample Community Detection

We show the results of applying Blockbuster to the entire sample in Figure 3. A scree plot suggests that the data has one common factor, so we apply Algorithm 2 with $R = 1$. The number of communities is set to three as in Hamilton and Owyang (2012). The left panel of the figure shows that our results bear interesting similarities to their findings. The red community roughly matches the first cluster of Hamilton and Owyang (2012) and contains oil-producing and agricultural states. The green cluster contains East Coast states, California and Arizona, which roughly corresponds to their third cluster. Finally, the remaining states are Mid West states together with Alabama. Notice that the communities estimated by Blockbuster mostly form geographically contiguous areas even though no spatial information is given in the algorithm. The right panel of Figure 3 shows a heatmap of the correlation matrix of the panel conditional on the factor where the series are ordered according to their community membership whereas Table 2 reports the average correlations by block. We see that series in the same community are positively correlated in the vast majority of cases and that the intra-community correlation is larger than the inter-community correlation. We also calculate the proportion of variance explained by the principal components. The first principal component in the panel explains 50% of the total variance. The proportion of the variance explained by the principal components associated with the communities is sizeable and explains 16% of the total variation of the panel.
The figure displays the communities detected by Blockbuster in the U.S. dataset with the number of communities set to $k = 3$. The left panel shows the communities on a map and the right panel displays a heatmap of the correlations (conditional on the factor) ordered by the Blockbuster partition. The top group in the heatmap corresponds to the Red community in the map, the middle group to the Blue community and the bottom group to the Green community.

Table 2: U.S. Real Activity Correlation Matrix Block Averages ($k = 3$)

<table>
<thead>
<tr>
<th>Community</th>
<th>Red</th>
<th>Blue</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>0.143</td>
<td>-0.078</td>
<td>-0.107</td>
</tr>
<tr>
<td>Blue</td>
<td>-0.078</td>
<td>0.207</td>
<td>-0.076</td>
</tr>
<tr>
<td>Green</td>
<td>-0.107</td>
<td>-0.076</td>
<td>0.222</td>
</tr>
</tbody>
</table>

The table shows the averages of the elements of the correlation matrix (conditional on the factor) within the estimated community blocks on the diagonal and the average of the elements in the across-community blocks on the off-diagonal, for the community partition estimated by Blockbuster with $k = 3$.

We also run the Blockbuster algorithm with the number of communities set to two and four. The results are reported in Figure 4. When the number of communities is set to two, the algorithm partitions the U.S. into East Coast states together with California, Arizona, Missouri and Tennessee, and a residual cluster containing all remaining states. When the number of communities is set to four, in comparison to the baseline case, the oil-producing and agricultural states community gets split into two separate clusters and California and Arizona are absorbed into the cluster containing oil-producing states. Note that the community corresponding to the East Coast is relatively stable across different choices of the number of clusters.
5.1.2 Out-of-Sample Community Validation

We carry out an out-of-sample validation exercise to assess the performance of the Blockbuster community detection algorithm on the basis of the covariance regularisation procedure described in Algorithm 3. The validation exercise is designed as follows. We split the sample in half and label the first half in-sample and the second half out-of-sample.\footnote{We split the sample in half to have enough observations to calculate the out-of-sample covariance matrix with sufficient accuracy. For example, a 75%/25% sample split would yield only 51 observations to estimate a 48-dimensional covariance matrix.} We estimate the Blockbuster covariance matrix in-sample and then evaluate the estimator out-of-sample by predicting the out-of-sample sample covariance matrix of the panel. Since there is evidence of one factor in the data, we apply the Blockbuster covariance estimator which accounts for one factor. We measure the precision of the forecast on the basis of the Kullback-Leibler (KL) loss proposed by Stein (Stein, 1956; Pourahmadi, 2013). The loss is given by

$$L = \text{tr} \left( \hat{\Sigma}_C \Sigma_B^{-1} \right) - \log \left( \left| \hat{\Sigma}_C \Sigma_B^{-1} \right| \right) - n$$

where $\hat{\Sigma}_C$ is a candidate estimator and $\Sigma_B$ is a benchmark to compare against. We use the sample covariance matrix of the out-of-sample period as the benchmark. We run the procedure with the number of communities ranging from two to four. We compare the performance of the Blockbuster covariance estimator with the (standard) sample covariance as well as the POET covariance estimator (Fan \textit{et al.}, 2013), which is a regul-
larisation procedure based on an approximate factor model representation of the panel, and the (linear) Ledoit and Wolf (LW) shrinkage covariance estimator (Ledoit and Wolf, 2004). Table 3 reports the relative KL loss gains over the sample covariance, POET and LW estimators. The table shows that the Blockbuster estimator performs favourably relative to the set of alternative estimators considered. In particular, the best out-of-sample performance is obtained for \( k = 2 \).

Table 3: U.S. Real Activity Kullback-Leibler Losses

<table>
<thead>
<tr>
<th>( k )</th>
<th>SCM</th>
<th>LW</th>
<th>POET</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.735%</td>
<td>13.436%</td>
<td>4.313%</td>
</tr>
<tr>
<td>3</td>
<td>3.127%</td>
<td>11.041%</td>
<td>1.521%</td>
</tr>
<tr>
<td>4</td>
<td>1.537%</td>
<td>9.581%</td>
<td>1.950%</td>
</tr>
</tbody>
</table>

The table reports the relative gain of the Blockbuster covariance estimator with respect to the sample covariance estimator, the POET covariance estimator and the (linear) Ledoit and Wolf shrinkage covariance estimator in terms of the Kullback-Leibler loss for different choices of the number of communities \( k \).

5.2 Community Detection in Europe

For our second application we construct a dataset using data from Eurostat. The data consists of yearly growth rates of the gross regional product (GRP) for the first-level European Nomenclature of Territorial Units for Statistics (NUTS 1) regions of Europe (excluding Bulgaria, Croatia, Cyprus, Estonia, Iceland and Turkey) from 2000 to 2013, which results in a panel of \( n = 99 \) time series over \( T = 13 \) periods. The GRP data is in purchasing power standard units. The analysis of this panel is clearly more challenging, as the large sample theory developed in the paper is not applicable in this setting. However, simulations suggest that even when \( T \) is smaller than \( n \), the Blockbuster algorithm performs satisfactorily provided that the probability of the intra-community linkages is sufficiently strong relative to the probability of inter-community linkages and/or the network dependence parameter is sufficiently large. Due to size of the sample, we only carry out an in-sample analysis.
5.2.1 In-Sample Community Detection

We show the results for Europe in Figures 5 and 6. A scree plot suggests that the data has one common factor and we use Algorithm 2 setting the number of factors $R$ to one. As we have little guidance from the literature on how to choose the number of clusters, we let the number of communities range from six to eight. In the following discussion, we focus on the clustering results when the number of communities is set to seven as in Figure 5. Blockbuster finds mostly geographically contiguous areas. The four largest European economies (France, Germany, Italy and the UK) essentially form clusters of their own. In some cases, these clusters also include smaller neighbouring economies that have strong economic ties with these large economies. For instance, Ireland is in the UK cluster. On the other hand, there are clusters that consist of multiple smaller economies. For instance, there is evidence of a Southern European cluster, containing Greece, Portugal and Spain (with the exception of Catalonia), as well as a Central-Eastern European cluster containing (among others) Austria, Finland, Poland and parts of Sweden. The right panel of Figure 5 shows a heatmap of the correlation matrix of the panel conditional on the factor where the series are ordered according to their community membership whereas Table 4 reports the average correlations by block. We see that series in the same community are positively correlated in the vast majority of cases. Intra-community correlations are quite strong and in the neighbourhood of 0.5. Unlike the US dataset, there is more inter-community correlation as we observe blocks exhibiting strong correlation. For instance, the Cyan community (that essentially corresponds to Germany) is highly correlated with the Orange and Blue communities; these three consist of essentially Austria, Belgium, Denmark, Finland, Germany, Hungary, Poland and Sweden. Overall, the clustering results provide interesting insight into the community structure of Europe, despite the small sample size.
Figure 5: Europe Real Activity Clustering \((k = 7)\)

The figure displays the communities detected by Blockbuster in the European dataset with the number of communities set to \(k = 7\). The left panel shows the communities on a map and the right panel displays a heatmap of the correlations (conditional on the factor) ordered by the Blockbuster partition. From top to bottom the groups in the heatmap correspond to the Red, Blue, Green, Yellow, Pink, Cyan and Orange communities.

Table 4: Europe Real Activity Correlation Matrix Block Averages \((k = 7)\)

<table>
<thead>
<tr>
<th>Community</th>
<th>Red</th>
<th>Blue</th>
<th>Green</th>
<th>Yellow</th>
<th>Pink</th>
<th>Cyan</th>
<th>Orange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>0.594</td>
<td>0.048</td>
<td>–0.031</td>
<td>0.053</td>
<td>0.057</td>
<td>–0.182</td>
<td>0.050</td>
</tr>
<tr>
<td>Blue</td>
<td>0.048</td>
<td>0.604</td>
<td>–0.289</td>
<td>–0.246</td>
<td>–0.046</td>
<td>0.230</td>
<td>0.073</td>
</tr>
<tr>
<td>Green</td>
<td>–0.031</td>
<td>–0.246</td>
<td>0.646</td>
<td>0.134</td>
<td>–0.357</td>
<td>–0.202</td>
<td>–0.168</td>
</tr>
<tr>
<td>Yellow</td>
<td>0.053</td>
<td>–0.246</td>
<td>0.134</td>
<td>0.506</td>
<td>–0.041</td>
<td>–0.499</td>
<td>–0.153</td>
</tr>
<tr>
<td>Pink</td>
<td>0.057</td>
<td>–0.046</td>
<td>–0.357</td>
<td>–0.041</td>
<td>0.430</td>
<td>0.058</td>
<td>–0.014</td>
</tr>
<tr>
<td>Cyan</td>
<td>–0.182</td>
<td>0.230</td>
<td>–0.202</td>
<td>–0.499</td>
<td>0.058</td>
<td>0.628</td>
<td>0.165</td>
</tr>
<tr>
<td>Orange</td>
<td>0.050</td>
<td>0.073</td>
<td>–0.168</td>
<td>–0.153</td>
<td>–0.014</td>
<td>0.165</td>
<td>0.321</td>
</tr>
</tbody>
</table>

The table shows the averages of the elements of the correlation matrix (conditional on the factor) within the estimated community blocks on the diagonal and the average of the elements in the across-community blocks on the off-diagonal, for the community partition estimated by Blockbuster with \(k = 7\).
6 Conclusion

In this paper we consider the problem of community detection in partial correlation network models. We begin by introducing a class of partial correlation network models in which the underlying network structure is random and determined by a latent random graph (Chung and Lu, 2006; van der Hofstad, 2015). We use this framework to introduce a class of partial correlation network models with a community structure by assuming that the latent random graph is a weighted and degree-corrected stochastic block model (Holland et al., 1983; Karrer and Newman, 2011).

A natural question that arises in this framework is how to detect communities from a sample of observations. We propose an algorithm called Blockbuster that uses the \( k \)-means clustering procedure on the top rescaled eigenvectors of the estimated covariance matrix to detect communities. We study the large sample properties of the algorithm and establish its consistency when the number of variables \( n \) and observations \( T \) is large.

Two extensions of the baseline methodology are introduced. First we extend the baseline methodology to the case where common factors affect the dependence structure of the data. Second, we introduce a covariance regularisation procedure based on Blockbuster.
which is motivated by the special block covariance structure of the model.

We apply the methodology to study real activity clustering in the U.S. and Europe. The objective of the application is to partition the two panels into communities characterised by a high degree of interdependence in real activity growth. For the U.S. we employ a dataset of quarterly employment growth rates for the states from 1956 to 2007 constructed by Hamilton and Owyang (2012), whereas for Europe we construct a dataset of yearly GRP growth rates for the NUTS 1 regions from 2000 until 2013 using data from Eurostat. Blockbuster detects a meaningful partition of the series in the panels. Interestingly, the procedure performs well in the European panel despite the short sample size. In the U.S. panel we design an out-of-sample validation exercise and show that the covariance regularisation procedure based on Blockbuster improves covariance prediction compared to a number of alternative procedures.
References


A Proofs of Main Results

Proof of Lemma 1. See Lemmas 3.2 and 3.3 of Qin and Rohe (2013) and let $\tau = 0$. Their proof is for the matrix $-L + I_n$ which is easily translatable to our case. Notice that $K = (1/\sigma^2)I_n + (\phi/\sigma^2)L$ implies that the eigenvectors of $K$ and $L$ are the same and their spectra are the closely related.

Proof of Theorem 2. By the definition of $K$ and $\Sigma$ and the fact that $\bar{d}_{\min} = \Omega(n^{\rho_n})$, we have

$$\|K - \Sigma\| = O\left(\|L - \Sigma\|\right) = O\left(\sqrt{\log n / n^{\rho_n}}\right),$$

with high probability from Theorem 4. By Theorem 5 if $T = \Omega\left(n^2/\gamma - 1\right)$, we have

$$\|\hat{\Sigma} - \Sigma\| = O\left(\sqrt{n / T}\|\Sigma\|\right),$$

with high probability where $C > 0$ is an absolute constant. To turn this into a statement about the concentration matrix, note that we may write

$$\|\hat{K} - K\| = \|\hat{K}(\Sigma - \hat{\Sigma})K\| \leq \|\hat{K}\|\|\Sigma - \hat{\Sigma}\||K| = O\left(\sqrt{n / T}\right),$$

which follows as $\|\Sigma\|$ and $\|K\|$ are bounded by a constant with high probability by (9). Weyl’s inequality and Lemma 1 for large enough $T$. The triangle inequality along with (10) and (9) then delivers

$$\|\hat{K} - K\| \leq \|\hat{K} - K\| + \|K - K\| = O\left(\sqrt{n / T} + \sqrt{\log n / n^{\rho_n}}\right).$$

We next apply the Davis-Kahan theorem (Davis and Kahan, 1970; Bhatia, 1997) to bound the angles between the subspaces of the bottom $k$ eigenvectors of $\hat{K}$ and $K$. We follow Sarkar and Bickel (2015) and the arguments of Appendix B of Rohe et al. (2011) closely. We then obtain

$$\|\hat{U} - U\| = O\left(\sqrt{n / T} + \sqrt{\log n / n^{\rho_n}}\right),$$

with high probability, where $O \equiv FG'$ is a $k \times k$ orthonormal rotation matrix based on the matrices of the singular value decomposition $U\hat{U} = F\Psi G'$. We now bound $\|\hat{N} - N\|$.

Notice that

$$|\hat{U}_{ji} - [U]_{ji}[O]_{ji}| \leq \|\hat{U} - U[O]\| \leq \|\hat{U} - U\|,$$

for all $i = 1, \ldots, k, j = 1, \ldots, n$, by the definition of the Euclidean and operator norms.
Combining this with the fact that the rows of \( \hat{U} \) and \( U \) have only \( k \) elements, we obtain

\[
\left\| \hat{U}_j \bullet - [U]_j \bullet \right\| = \sqrt{\sum_{i=1}^{k} \left( [\hat{U}]_{ji} - [U]_{ji} \bullet \right)^2} = O \left( \left\| \hat{U} - U \mathcal{O} \right\| \right),
\]

which bounds the row norms of \( \hat{U} - U \mathcal{O} \). Using the definition of \( \hat{N} \) and \( N \), the fact that the Euclidean norm is invariant under orthonormal transformations and the reverse triangle inequality gives

\[
[\hat{N}^{-1} - N^{-1}]_{jj} = \left\| [\hat{U}]_j \bullet \right\| - \left\| [U]_j \bullet \right\| \leq \left\| [\hat{U}]_j \bullet - [U]_j \bullet \right\|.\]

This implies

\[
\left\| \hat{N}^{-1} - N^{-1} \right\| = O \left( \left\| \hat{U} - U \mathcal{O} \right\| \right),
\]
as \( \hat{N}^{-1} - N^{-1} \) is a diagonal matrix. Furthermore

\[
\left\| \hat{N} - N \right\| \leq \left\| (N^{-1} - \hat{N}^{-1}) \right\| \left\| N \right\| = O \left( \left\| N \right\| \left\| \hat{U} - U \mathcal{O} \right\| \right),
\]

with high probability, as \( \left\| \hat{N} \right\| \) is close to \( \left\| N \right\| \) for large enough \( n \). As \( \left\| N \right\| = 1/ \min_j \left\| [U]_j \bullet \right\| \) and row \( i \) of \( U \) has length

\[
\left\| [U]_i \bullet \right\| = \left( \frac{[\Theta]_{ii}}{\sum_{j \in \mathcal{V}_s} [\Theta]_{jj}} \right)^{1/2},
\]

where \( s \) is the community that vertex \( i \) belongs to, we have \( \left\| N \right\| = O (\sqrt{n}) \). It then follows that

\[
\left\| \hat{N} - N \right\| = O \left( \sqrt{n} \left\| \hat{U} - U \mathcal{O} \right\| \right),
\]

with high probability.

To put the pieces together, we may write

\[
\left\| \hat{X} - X \mathcal{O} \right\| = \left\| \hat{N} \hat{U} - NU \mathcal{O} \right\| \leq \left\| \hat{N} - N \right\| \left\| \hat{U} - U \mathcal{O} \right\| + \left\| \hat{N} - N \right\| \left\| U \mathcal{O} \right\| \left\| N \right\| \left\| \hat{U} - U \mathcal{O} \right\|.
\]

Notice that the cross term is lower order and \( \left\| U \mathcal{O} \right\| \) is a constant, as the matrix \( U \mathcal{O} \) has \( k \) columns of unit length. This taken together with (12) and (13), yields

\[
\left\| \hat{X} - X \mathcal{O} \right\| = O \left( \sqrt{n} \left\| \hat{U} - U \mathcal{O} \right\| \right) = O \left( \frac{n}{\sqrt{T}} + \sqrt{\log n} \rho_n \right),
\]

with high probability. \( \Box \)

**Proof of Lemma 2.** This follows from a straightforward extension of the proof of Lemma 3.2 from Rohe et al. (2011). Furthermore, see the proof of Theorem 4.4 of Qin and Rohe (2013). \( \square \)
Proof of Theorem 1. The proof follows the approach in Rohe et al. (2011) closely. Note that as \( C \in \mathcal{M}(n, k) \)
\[
\left\| \hat{X} - \hat{C} \right\|_F \leq \left\| \hat{X} - C \right\|_F, \tag{14}
\]
and by the triangle inequality
\[
\left\| \hat{C} - C \right\|_F \leq \left\| \hat{C} - \hat{X} \right\|_F + \left\| \hat{X} - C \right\|_F \leq 2 \left\| \hat{X} - C \right\|_F, \tag{15}
\]
where the last inequality follows from (14).

We now bound the cardinality of the set of misclustered vertices. To this end, it suffices to bound the size of the set \( \mathcal{M} \). Notice that
\[
|\mathcal{M}| = \sum_{i \in \mathcal{M}} 1 \leq 2 \sum_{i \in \mathcal{M}} \left\| \hat{C}_{i, \bullet} - [Z]_{\bullet} V \right\|^2 \leq 2 \left\| \hat{X} - C \right\|_F \leq 8 \left\| \hat{X} - C \right\|_F^2,
\]
where the second inequality follows from the fact that
\[
\sum_{i=1}^n \left\| \hat{C}_{i, \bullet} - [Z]_{\bullet} V \right\|^2 = \left\| \hat{C} - C \right\|_F^2
\]
and that the sum only includes a subset of the vertices. Theorem 2 implies
\[
\left\| \hat{X} - X_0 \right\|_F = O \left( \frac{n}{\sqrt{T}} + \frac{\log n}{\rho_n} \right),
\]
as \( \left\| \hat{X} - X_0 \right\|_F \leq \sqrt{k} \left\| \hat{X} - X_0 \right\| \) for a rank \( k \) matrix. It follows that
\[
\frac{|\mathcal{M}|}{n} = O \left( \frac{n}{T} + \frac{\log n}{n \rho_n} \right).
\]

Proof of Lemma 2. We wish to find the eigenvalues and eigenvectors of the matrix \( K^{-1} = K_{\epsilon}^{-1} + \sum_{r=1}^R q_r q_r^T \). We then wish to show that \( U = U_1 \), which allows us to apply Lemma 1 and finish the proof.

We proceed by induction and consider \( R = 1 \) with \( K_1^{-1} = K_{\epsilon}^{-1} + q_1 q_1^T \) first. We then find the eigenvectors of \( K^{-1} \equiv K_R^{-1} = K_{R-1}^{-1} + q_R q_R^T \) given those of \( K_{R-1}^{-1} \). Notice that the eigenvectors \( u_i(K_\epsilon) \) for \( i = 1, \ldots, n \) form a basis in \( \mathbb{R}^n \). We may then write \( q_1 \) as the linear combination
\[
q_1 = \sum_{i=1}^n \gamma_{i,1} u_i(K_\epsilon),
\]
where \( \gamma_{i,1} \) are scalars. The assumption that \( q_1^T [U_1]_{\bullet i} = 0 \) from Assumption 3 then amounts to \( \gamma_{i,1} = 0 \) for \( i = 1, \ldots, k \), so that
\[
q_1 = \gamma_{k+1,1} u_{k+1}(K_\epsilon) + \cdots + \gamma_{n,1} u_n(K_\epsilon).
\]

We guess and verify the eigenvectors and eigenvalues of \( K_1 \) using the eigenvalue equation
\[
K_1^{-1} u_i(K_1) = \lambda_i(K_1)^{-1} u_i(K_1),
\]
and ensure they are mutually orthogonal. We begin with the bottom eigenvalue and eigenvector of \( K_1 \) and guess \( \lambda_1(K_1)^{-1} = \lambda_n(K_\epsilon)^{-1} + ||q_1||^2 \) and \( u_1(K_1) = q_1/||q_1|| \). We
have

\[ K^{-1}_1 q_1 = (K^{-1}_e + q_1 q_1^t) q_1 = K^{-1}_e q_1 + \|q_1\|^2 q_1 = (\lambda_n(K_e)^{-1} + \|q_1\|^2) q_1, \]

where the last equality follows from

\[ K^{-1}_e q_1 = \sum_{i=k+1}^{n} \gamma_{i,1} K^{-1}_e u_i(K_e) = \sum_{i=k+1}^{n} \gamma_{i,1} \lambda_i(K_e)^{-1} u_i(K_e) = \lambda_n(K_e)^{-1} q_1, \]

as \( \lambda_i(K_e) = \lambda_n(K_e) \) for \( i = k + 1, \ldots, n \) from Lemma 1. Dividing through by \( \|q_1\| \) we have the eigenvector. Next we consider the second to the \( k + 1 \)-th bottom eigenvalues and eigenvectors of \( K_1 \). We postulate \( \lambda_{i+1}(K_1)^{-1} = \lambda_i(K_e)^{-1} \) and \( u_{i+1}(K_1) = u_i(K_e) \) for \( i = 1, \ldots, k \). Consider

\[ K^{-1}_1 u_i(K_e) = K^{-1}_e u_i(K_e) + q_1 q_1^t u_i(K_e) = K^{-1}_e u_i(K_e) = \lambda_i(K_e)^{-1} u_i(K_e), \]

for \( i = 1, \ldots, k \), which follows from the fact that the bottom \( k \) eigenvectors of \( K_e^{-1} \) are orthogonal to \( q_1 \).

It remains to find the last \( n - k - 1 \) eigenvectors of \( K_1 \). Let

\[ u_n(K_1) = \gamma_{n-1,1} u_n(K_e) - \gamma_{n,1} u_{n-1}(K_e), \]

and

\[ u_{n-i}(K_1) = \gamma_{n,1} u_n(K_e) + \gamma_{n-1,1} u_{n-1}(K_e) + \ldots + \gamma_{n-i,1} u_{n-i}(K_e) - \left( \frac{\sum_{j=n-i}^{n} \gamma_{j,1}^2}{\gamma_{n-i-1,1}} \right) u_{n-i-1}(K_e), \]

for all \( i = 1, \ldots, n - k - 2 \). These vectors are orthogonal to \( q_1 \), as the last term always cancels out all the others. They are also orthogonal to each other by a similar argument. Returning to the eigenvalue equation, we have

\[ K^{-1}_1 u_i(K_1) = (K^{-1}_e + q_1 q_1^t) u_i(K_1) = K^{-1}_e u_i(K_1) = \lambda_n(K_e)^{-1} u_i(K_1), \]

for \( i = k + 2, \ldots, n \). Dividing through by the norm delivers the remaining eigenvectors of \( K_1 \), all with eigenvalue \( \lambda_n(K_e)^{-1} \).

Now assume that we know the eigenvectors \( u_i(K_{R-1}) \) of \( K_{R-1} \). We look for the eigenvectors of \( K^{-1} \equiv K^{-1}_R = K^{-1}_{R-1} + q_R q_R^t \) and note that \( K^{-1}_{R-1} = K^{-1}_e + \sum_{r=1}^{R-1} q_r q_r^t \). Proceeding as before, we begin with the bottom \( R \) eigenvalues and eigenvectors of \( K \) and guess \( \lambda_i(K)^{-1} = \lambda_n(K_e)^{-1} + \|q_i\|^2 \) and \( u_i(K) = q_i/\|q_i\| \) for \( i = 1, \ldots, R \). We have

\[ K^{-1}_i q_i = \left( K^{-1}_e + \sum_{r=1}^{R} q_r q_r^t \right) q_i = K^{-1}_e q_i + \|q_i\|^2 q_i = \left( \lambda_n(K_e)^{-1} + \|q_i\|^2 \right) q_i, \]

for all \( i = 1, \ldots, R \) as \( q_i^t q_v = 0 \) for all \( v = 1, \ldots, R, v \neq i \). We next postulate \( \lambda_{R+i}(K)^{-1} = \lambda_i(K_e)^{-1} \) and \( u_{R+i}(K) = u_i(K_e) \) for \( i = 1, \ldots, k \). As before

\[ K^{-1}_i u_i(K_e) = K^{-1}_e u_i(K_e) + \sum_{r=1}^{R} q_r q_r^t u_i(K_e) = K^{-1}_e u_i(K_e) = \lambda_i(K_e)^{-1} u_i(K_e), \]
for $i = 1, \ldots, k$. Finally, notice that $K$ is a rank-one update of $K_{R-1}$, so we may apply a similar logic to before: Notice that $u_i(K_{R-1})$ form a basis for $\mathbb{R}^n$ and write

$$q_R = \sum_{i=R+k}^n \gamma_{i,R} u_i(K_{R-1}).$$

Then let

$$u_n(K) = \gamma_{n-1,R} u_n(K_{R-1}) - \gamma_{n,R} u_{n-1}(K_{R-1}),$$

and

$$u_{n-1}(K) = \gamma_{n,R} u_n(K_{R-1}) + \ldots + \gamma_{n-i,R} u_{n-i}(K_{R-1}) - \left( \sum_{j=n-i}^n \frac{\gamma_j^2}{\gamma_{n-i-1,R}} \right) u_{n-i-1}(K_{R-1}),$$

for all $i = 1, \ldots, n - k - R - 1$. From the eigenvalue equation

$$K_{R-1}^{-1} u_i(K) = \left( K_{R-1}^{-1} + q_R q_R' \right) u_i(K) = K_{R-1}^{-1} u_i(K) = \lambda_n (K_{R-1})^{-1} u_i(K),$$

for all $i = 1, \ldots, n - k - R - 1$. Notice that by induction $\lambda_n (K_{R-1})^{-1} = \lambda_n (K_{\epsilon})^{-1}$. After normalising, we have the last $n - k - R$ eigenvectors, all with eigenvalue $\lambda_n (K_{\epsilon})^{-1}$. We thus have all the eigenvalues of $K$ and have shown that $U = U_{\epsilon}$. We may apply Lemma 1 to finish the proof.

**Proof of Theorem 3.** We have

$$\|K^{-1} - K^{-1}\| = \left\| K^{-1}_{\epsilon} + \sum_{r=1}^R q_r q_r' - \left( K^{-1}_{\epsilon} + \sum_{r=1}^R q_r q_r' \right) \right\| = \|K^{-1}_{\epsilon} - K^{-1}_{\epsilon}\| = O \left( \sqrt{\frac{\log n}{n\rho n}} \right),$$

where the last step follows from the fact that $\|K_{\epsilon} - K_{\epsilon}\| = O \left( \sqrt{\log n / (n\rho n)} \right)$ by Theorem 4 and that $\|K^{-1}_{\epsilon}\|$ is bounded by a constant by Lemma 1. This implies

$$\|K - K\| = O \left( \sqrt{\frac{\log n}{n\rho n}} \right),$$

as $\|K\|$ is bounded by a constant by Lemma 3. By Theorem 5 and similar arguments as in the proof of Theorem 2, we obtain

$$\|\hat{K} - K\| = O \left( \sqrt{\frac{n}{T}} \|\Sigma\| \right),$$

as $\|K\|$ is bounded with high probability. Then we have

$$\|\hat{K} - K\| \leq \|\hat{K} - K\| + \|K - K\| = O \left( \sqrt{\frac{n}{T}} \|\Sigma\| + \sqrt{\frac{\log n}{n\rho n}} \right),$$

by the triangle inequality. We may then apply the Davis-Kahan theorem in a similar way to before in the proof of Theorem 2. Finally, applying similar arguments as for Theorem 44.
we obtain
\[
\frac{|\mathcal{M}|}{n} = O\left(\frac{n}{\bar{T}} \|\Sigma\|^2 + \frac{\log n}{n \rho_n}\right).
\]

\[\square\]

B Additional Results

This appendix contains some results that are used in Sections 2 and 3. In Theorem 4 we extend the random graph concentration result of Theorem 3.1 in \cite{Oliveira2009} to allow for weighted graphs.

**Theorem 4.** Consider a random undirected and weighted graph on \(n\) vertices \(G = (V, E, W)\) where \((i, j) \in E\) with probability \(p_{i,j} = p_{j,i}\) independently and the edge weights \(W_{ij} \in \mathbb{W}\) are independent random variables supported on the interval \([\alpha_{ij}, \beta_{ij}]\) with \(\beta_{ij} > 0\), mean \(\mu > 0\) and variance \(\sigma_{ij}^2\) for all \(i, j\). Let \(L\) be the Laplacian matrix corresponding to \(G\) and \(\mathcal{L}\) be its population analogue. Let \(\bar{d}_{\min} = \min_i \bar{d}_i\) be the minimum expected degree of \(G\).

If \(\bar{d}_{\min} = \Omega(\log(n))\), then
\[
\|L - \mathcal{L}\| = O\left(\sqrt{\log(n)} \frac{\bar{d}_{\min}}{\bar{d}_{\min}}\right),
\]
with high probability.

**Proof.** We follow \cite{Oliveira2009}. We begin by controlling the following sum of independent zero-mean random variables
\[
d_i - \bar{d}_i = \sum_{j=1}^{n} (B_{ij} W_{ij} - p_{ij} \mu_{ij})
\]
Notice that \(|B_{ij} W_{ij} - p_{ij} \mu_{ij}|\) takes its maximum at one of the following: \(p_{ij} \mu_{ij}\), \(|\alpha_{ij}| + p_{ij} \mu_{ij}\) or \(\beta_{ij} - p_{ij} \mu_{ij}\). Hence
\[
B_{ij} W_{ij} - p_{ij} \mu_{ij} \leq |\alpha_{ij}| + \beta_{ij} \leq \max_i |\alpha_{ij}| + \max_i \beta_{ij}.
\]
as the the eigenvalues of \(A_{ij}\) are in \([-1, 0, 1]\). Next consider the variance statistic
\[
\sum_{j=1}^{n} \mathbb{E} [(B_{ij} W_{ij} - p_{ij} \mu_{ij})^2] = \sum_{j=1}^{n} \text{Var} \left(B_{ij} W_{ij}\right).
\]
We require the variance of the composite random variable \(B_{ij} W_{ij}\), where \(B_{ij}\) and \(W_{ij}\) are independent. We have
\[
\text{Var} \left(B_{ij} W_{ij}\right) = \text{Var} \left(B_{ij}\right) \mathbb{E} \left[W_{ij}\right]^2 + \mathbb{E} \left[B_{ij}\right]^2 \text{Var} \left(W_{ij}\right) + \text{Var} \left(B_{ij}\right) \text{Var} \left(W_{ij}\right)
\]
\[
= p_{ij} (1 - p_{ij}) \mu_{ij}^2 + \sigma_{ij}^2 p_{ij}^2 + p_{ij} (1 - p_{ij}) \sigma_{ij}^2 = p_{ij} (1 - p_{ij}) \mu_{ij}^2 + p_{ij} \sigma_{ij}^2
\]
\[
\leq p_{ij} \left(\mu_{ij}^2 + \sigma_{ij}^2\right) = p_{ij} \mu_{ij} \left(\mu_{ij} + \frac{\sigma_{ij}^2}{\mu_{ij}}\right) \leq p_{ij} \mu_{ij} \left(\max_i \mu_{ij} + \max_i \sigma_{ij}^2\right),
\]

using $\mathbb{E} [B_{ij}] = p_{ij}$ and $\text{Var} (B_{ij}) = p_{ij} (1 - p_{ij})$. It follows that

$$\sum_{j=1}^{n} \text{Var} (B_{ij} W_{ij}) \leq \left( \max_{ij} \mu_{ij} + \max_{ij} \sigma_{ij}^2 \right) \bar{d}_i,$$

as $\bar{d}_i = \sum_j p_{ij} \mu_{ij}$.

Define $\nu = \max_{ij} \mu_{ij} + \max_{ij} \sigma_{ij}^2 + \max_{ij} |\alpha_{ij}| + \max_{ij} \beta_{ij}$, fix $c > 0$ and assume $n^{-c} \leq \delta \leq 1/2$. We may then apply Corollary 7.1. from Oliveira (2009) with $n = 1$ to obtain for all $r > 0$ and all $i$

$$\mathbb{P} \left( |d_i - \bar{d}_i| \geq r \right) = \mathbb{P} \left( \left| \frac{d_i}{\bar{d}_i} - 1 \right| \geq \frac{r}{\bar{d}_i} \right) = \mathbb{P} \left( \left| \frac{d_i}{\bar{d}_i} - 1 \right| \geq t \right) \leq 2e^{-\frac{\nu^2 \bar{d}_i^2}{8s^2 + 4Mt_{d_i}}},$$

where we took $r = t \bar{d}_i$. Using the arguments from the previous paragraph, we set $M = \nu$ and $s^2 = \nu \bar{d}_i$ and take $t = 4\sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}^2}} \leq 2$, where the inequality follows as $d_{\min} > C \log(n)$ and we may choose $C$ high enough. Then

$$\mathbb{P} \left( \left| \frac{d_i}{\bar{d}_i} - 1 \right| \leq 4 \sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}}} \right) \geq 1 - \delta. \quad (16)$$

By the mean value theorem

$$\left| \sqrt{1 + x} - 1 \right| = \left( \frac{1}{2 \sqrt{1 + \gamma}} \right) |x| \leq \max_{\gamma \in [-3/4, 3/4]} \left( \frac{1}{2 \sqrt{1 + \gamma}} \right) |x| = |x|,$$

with $x \in [-3/4, 3/4]$ and the last equality follows as $\gamma = -3/4$ yields the maximum. Take $x = \frac{d_i}{\bar{d}_i} - 1$ and notice that we may choose $C$ to make $d_{\min} > C \log(n)$ large enough for (16) to imply $|\frac{d_i}{\bar{d}_i} - 1| \leq 3/4$. It follows that

$$\left| \sqrt{\frac{d_i}{\bar{d}_i}} - 1 \right| \leq \left| \frac{d_i}{\bar{d}_i} - 1 \right| \leq 4 \sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}}},$$

with probability greater than $1 - \delta$.

Define $T = D^{-1/2}$ and $\mathcal{T} = D^{-1/2}$. Notice that $\|T T^{-1} - I_n\| = \max_{i} \left| \sqrt{\frac{d_i}{\bar{d}_i}} - 1 \right|$ as this is a diagonal matrix, which yields $\|T T^{-1} - I_n\| \leq 4 \sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}}} \mathbb{P}$ with probability greater than $1 - \delta$. We define the intermediate operator $\mathcal{M} = I_n - \mathcal{T} \mathcal{A} \mathcal{T}$ and note that it satisfies $\mathcal{M} = I_n - (T T^{-1}) (I_n - L) (T T^{-1})$ as $L = I_n - T \mathcal{A}$. To compare $\mathcal{L}$ with $L$, we bound the distance of each from $\mathcal{M}$. We begin with $\|\mathcal{M} - L\|$ and write

$$\|\mathcal{M} - L\| = \| (T T^{-1}) (I_n - L) (T T^{-1}) + (I_n - L) (T T^{-1}) - (I_n - L) (T T^{-1}) - (I_n - L) \|$$

$$= \| (T T^{-1} - I_n) (I_n - L) (T T^{-1}) + (I_n - L) (T T^{-1} - I_n) \|$$

$$\leq \| T T^{-1} - I_n \| \| T T^{-1} \| + \| T T^{-1} - I_n \|$$

$$\leq 4 \sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}}} \left( 1 + 4 \sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}}} \right) + 4 \sqrt{\frac{\nu \log (2n/\delta)}{d_{\min}}}$$
\[ \leq 10 \sqrt{\frac{\nu \log(2n/\delta)}{d_{\min}}}, \]

where we used the fact that \( \|I_n - L\| \leq 1 \) (Chung, 1997).

We now control \( \|M - L\| \). Let \( B_{ij} \) for all \( 1 \leq i, j \leq n \) be independent Bernoulli variables that take the value 1 with probability \( p_{ij} \) and 0 otherwise. Let \( e_1, e_2, \ldots, e_n \) be the standard basis for \( \mathbb{R}^n \). Define for all \( 1 \leq i, j \leq n \) the \( n \times n \) matrices

\[
A_{ij} = \begin{cases} 
  e_i e'_j + e_j e'_i, & i \neq j, \\
  e_i e'_i, & i = j.
\end{cases}
\]

Then we may write the object of interest as the sum

\[ M - L = \sum_{i \leq j} Y_{ij}, \]

where

\[
Y_{ij} = \frac{B_{ij} W_{ij} - p_{ij} \mu_{ij}}{\sqrt{d_i d_j}} A_{ij}
\]

are mean-zero independent random matrices. As the eigenvalues of \( A_{ij} \) are in \( \{-1, 0, 1\} \), the eigenvalues of \( Y_{ij} \) are in

\[
\left\{ \pm \frac{(B_{ij} W_{ij} - p_{ij} \mu_{ij})}{\sqrt{d_i d_j}}, \pm p_{ij} \mu_{ij} \frac{\sqrt{\overline{d}_i \overline{d}_j}}{\sqrt{d_i d_j}} \right\},
\]

and thus \( \|Y_{ij}\| \leq \nu/\sqrt{d_i d_j} \leq \nu/\overline{d}_{\min} \) by similar arguments to before. Notice that

\[
A^2_{ij} = \begin{cases} 
  e_i e'_j + e_j e'_i, & i \neq j, \\
  e_i e'_i, & i = j.
\end{cases}
\]

The variance statistic is

\[
\sum_{1 \leq i \leq j \leq n} \mathbb{E} \left[ Y_{ij}^2 \right] = \sum_{1 \leq i \leq j \leq n} \text{Var} \left( B_{ij} W_{ij} \right) \frac{A^2_{ij}}{d_i d_j}
\]

\[
= \sum_{i=1}^{n} \text{Var} \left( B_{ii} W_{ii} \right) \frac{e_i e'_i}{d_i^2} + \sum_{1 \leq i < j \leq n} \text{Var} \left( B_{ij} W_{ij} \right) \frac{(e_i e'_i + e_j e'_j)}{d_i d_j}
\]

\[
= \sum_{i=1}^{n} \frac{1}{d_i} \left( \sum_{j=1}^{n} \frac{\text{Var} \left( B_{ij} W_{ij} \right)}{d_j} \right) e_i e'_i.
\]

Using \( \text{Var} \left( B_{ij} W_{ij} \right) \leq p_{ij} \mu_{ij} \nu \) from before and the fact that this is a diagonal matrix, we obtain

\[
\lambda_n \left( \sum_{1 \leq i \leq j \leq n} \mathbb{E} \left[ Y_{ij}^2 \right] \right) = \max_{i} \frac{1}{d_i} \left( \sum_{j=1}^{n} \frac{\text{Var} \left( B_{ij} W_{ij} \right)}{d_j} \right)
\]

\[
\leq \nu \max_{i} \frac{1}{d_i} \left( \sum_{j=1}^{n} \frac{p_{ij} \mu_{ij}}{d_j} \right) \leq \nu \frac{\max_{i} \frac{1}{d_i} \left( \sum_{j=1}^{n} p_{ij} \mu_{ij} \right)}{d_{\min}} = \frac{\nu}{d_{\min}},
\]

47
as \( \sum_j p_{ij} \mu_{ij} = \bar{d}_i \).

We may thus set \( M = s^2 = \nu/d_{\min} \) and apply Corollary 7.1 from Oliveira (2009), which yields for all \( t > 0 \)

\[
\mathbb{P} \left( \left\| \sum_{1 \leq i \leq j \leq n} Y_i \right\| \geq t \right) \leq 2ne^{-\frac{t^2d_{\min}}{\nu(\nu+\delta)}},
\]

Take \( t = 4\sqrt{(\nu \log(2n/\delta))/d_{\min}} \). When we bounded the degrees, we had ensured \( t \leq 3/4 \leq 2 \). Hence

\[
\mathbb{P} \left( \| M - \mathcal{L} \| \leq 4\sqrt{\frac{\nu \log(2n/\delta)}{d_{\min}}} \right) \geq 1 - \delta,
\]

and it follows by the triangle inequality that

\[
\mathbb{P} \left( \| L - \mathcal{L} \| \leq 14\sqrt{\frac{\nu \log(2n/\delta)}{d_{\min}}} \right) \geq 1 - \delta.
\]

Theorem 5 extends the matrix concentration inequality results for matrices with independent sub-Gaussian rows of Theorem 5.39 from Vershynin (2012). The rows are allowed to be strongly mixing with generalised exponential tails. A key result that we make use of is Theorem 1 (and Remark 1) of Merlevède et al. (2011), which is a Bernstein-type concentration inequality for strongly mixing random variables. We consider sample covariance matrices in particular.

**Theorem 5.** Consider a \( T \times n \) matrix \( Y \), whose rows are observations of centred \( n \)-dimensional stationary random vectors \( [Y]_t \), that satisfy (i), (ii) and (iii) of Assumption 2. Let their covariance matrix be \( \Sigma \) and define the sample covariance matrix as \( \hat{\Sigma} = \frac{1}{T} Y'Y \).

If \( T = \Omega \left( n^{2/\gamma - 1} \right) \), we have

\[
\left\| \hat{\Sigma} - \Sigma \right\| = O \left( \sqrt{\frac{n}{T}} \left\| \Sigma \right\| \right),
\]

with high probability where \( C > 0 \) is an absolute constant.

**Proof.** We begin by working with the isotropic version of the random vectors, \( \Sigma^{-1/2}Y_t \). For those, we adopt a similar strategy to Vershynin (2012) and proceed in three steps.

First we discretise the unit sphere using a net which allows us to approximate the spectral norm of the quantity of interest. We then apply the mixing concentration inequality of Merlevède et al. (2011) to control the spectral norm for every vector on the unit sphere. Next, we take the union bound over all such vectors to evaluate the spectral norm. At this point we have a concentration inequality for the isotropic vectors. In the fourth step, we translate that statement to the non-isotropic vectors \( Y_t \).

**Step 1: Approximation.** Apply Lemma 5.4 of Vershynin (2012) with a \( 1/4 \)-net \( \mathcal{N}_{1/4} \equiv \mathcal{N} \) to obtain

\[
\left\| \frac{1}{T} \Sigma^{-1/2}Y'Y\Sigma^{-1/2} - I_n \right\| \leq \frac{1}{1 - 1/2} \sup_{x \in \mathcal{N}} \left| x' \left( \frac{1}{T} \Sigma^{-1/2}Y'Y\Sigma^{-1/2} - I_n \right) x \right|
\]
\[
\begin{align*}
&= 2 \sup_{x \in \mathcal{N}} \left| \frac{1}{T} x' \Sigma^{-1/2} Y \Sigma^{-1/2} x - x' x \right| = 2 \sup_{x \in \mathcal{N}} \left| \frac{1}{T} \| Y \Sigma^{-1/2} x \|^2 - 1 \right|,
\end{align*}
\] and let
\[
\epsilon = C \sqrt{\frac{n}{T}} + \frac{r}{\sqrt{T}},
\]
for some \( r \geq 0 \) where \( C > 0 \) is a constant. We first wish to establish
\[
\sup_{x \in \mathcal{N}} \left| \frac{1}{T} \| Y \Sigma^{-1/2} x \|^2 - 1 \right| \leq \frac{\epsilon}{2},
\]
with high probability.

**Step 2: Concentration.** Fix a vector \( x \) with \( \| x \| = 1 \) and notice that \( \| Y \Sigma^{-1/2} x \|^2 \) may be written as a sum so that
\[
\frac{1}{T} \| Y \Sigma^{-1/2} x \|^2 - 1 = \frac{1}{T} \sum_{t=1}^{T} (Y_t \Sigma^{-1/2} x)^2 - 1 = \frac{1}{T} \sum_{t=1}^{T} Z_t,
\]
where \( Z_t = (Y_t \Sigma^{-1/2} x)^2 - 1 \). Notice that \( \mathbb{E} \left[ (Y_t \Sigma^{-1/2} x)^2 \right] = \| x \|^2 = 1 \) as \( \Sigma^{-1/2} Y_t \) are isotropic vectors, so \( Z_t \) are centred. Assumption 2 implies
\[
\mathbb{P} \left( \left| (Y_t \Sigma^{-1/2} x)^2 - 1 \right| > s \right) \leq c_3 e^{-(s/c_2)^{\gamma_2/2}},
\]
which is the required tail behaviour for the sequence \( \{Z_t\} \). (See Lemmas 7 and 8 of Gudmundsson, 2018a.) We may then apply Theorem 1 of Merlevède et al. (2011) to the sequence \( \{Z_t\} \). For any \( T \geq 4 \), there exist positive constants \( C_1, C_2, C_3 \) and \( C_4 \) that depend only on \( c_1, c_2, c_3, \gamma_1 \) and \( \gamma_2 \) such that for any \( \epsilon > 0 \)
\[
\mathbb{P} \left( \left| \frac{1}{T} \| Y \Sigma^{-1/2} x \|^2 - 1 \right| \geq \frac{\epsilon}{2} \right) = \mathbb{P} \left( \left| \sum_{t=1}^{T} Z_t \right| \geq \frac{T \epsilon}{2} \right) \leq \mathbb{P} \left( \sup_{r \leq T} \left| \sum_{t=1}^{r} Z_t \right| \geq \frac{T \epsilon}{2} \right) \leq T \exp \left( \frac{-(T\epsilon/2)^{\gamma}}{C_1} \right) + \exp \left( -\frac{(T \epsilon/2)^2}{C_2 (1 + TV)} \right) + \exp \left( -\frac{(T \epsilon/2)^2}{C_3 T} \exp \left( \frac{(T \epsilon/2)^{(\gamma_1 - \gamma)}}{C_4 (\log (T \epsilon/2)^\gamma)} \right) \right),
\]
where \( V \) is finite and \( \gamma \) is as in Assumption 3.

We begin by looking at the first term of (17). Substituting \( \epsilon = C \sqrt{\frac{n}{T}} + \frac{r}{\sqrt{T}} \) in, we obtain
\[
T \exp \left( \frac{-(T \epsilon/2)^{\gamma}}{C_1} \right) = \exp \left( \log(T) - \frac{T \left( C \sqrt{T} + \frac{r}{\sqrt{T}} \right)^\gamma}{2^{\gamma} C_1} \right) = \exp \left( \log(T) - \frac{1}{2^{\gamma} C_1} \left( C \sqrt{T n} + r \sqrt{T} \right)^\gamma \right) \leq \exp \left( \log(T) - \frac{1}{2^{\gamma} C_1} (C^2 T n + r^2 T)^{\gamma/2} \right),
\]

49
as $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Notice that
\[
\frac{(C^2 T n + T r^2)^{\gamma/2}}{C^2 n + r^2} = \frac{(T n)^{\gamma/2} (C^2 + r^2/n)^{\gamma/2}}{C^2 + r^2/n} = T^{\gamma/2} n^{\gamma/2 - 1} \left( C^2 + \frac{r^2}{n} \right)^{\gamma/2 - 1}.
\]
As $T = \Omega \left( n^{2/\gamma-1} \right)$, this ratio is not shrinking to zero. Furthermore, we may assume that $T = o(e^n)$ for all practical purposes, so we may write
\[
\exp \left( \log(T) - \frac{1}{2^\gamma C_1} (C^2 T n + r^2 T)^{\gamma/2} \right) \leq \exp \left( -C''(C^2 n + r^2) \right),
\]
for some appropriately small constant $C'' > 0$.

Plugging $\epsilon = C \sqrt{T} + \frac{r}{\sqrt{T}}$, into the second term of (17), we obtain
\[
\exp \left( -\frac{(T \epsilon/2)^2}{C_2 (1 + TV)} \right) = \exp \left( -\frac{1}{4 C_2 (1 + TV)} \left( C \sqrt{T n} + r \sqrt{T} \right)^2 \right)
\]
\[
\leq \exp \left( -\frac{1}{4 C_2 (V + 1/T)} (C^2 n + r^2) \right) \leq \exp \left( -C''(C^2 n + r^2) \right),
\]
for an appropriate constant $C'' > 0$, where we used the fact that $(a + b)^2 \geq a^2 + b^2$ if $a, b \geq 0$.

We begin by looking at the inner exponential term in the third term of (17)
\[
\exp \left( \frac{(T \epsilon/2)^{\gamma(1-\gamma)}}{C_4 (\log (T \epsilon/2))^{\gamma}} \right) = \exp \left( \frac{1}{2^\gamma (1-\gamma) C_4} \left( \frac{C \sqrt{T n} + r \sqrt{T}}{\log \left( C \sqrt{T n} + r \sqrt{T} - \log(2) \right)} \right)^{1-\gamma} \right).
\]
As $\gamma < 1$, we have $(1 - \gamma)x^{-\gamma} > x^{-1}$ for all $x$ large enough. This term is thus increasing in $T$, so we may bound the third term of (17) with
\[
\exp \left( -\frac{(T \epsilon/2)^2}{C_3 T} \exp \left( \frac{(T \epsilon/2)^{\gamma(1-\gamma)}}{C_4 (\log (T \epsilon/2))^{\gamma}} \right) \right) \leq \exp \left( -\frac{(T \epsilon/2)^2}{C_3 T} C^* \right)
\]
\[
= \exp \left( -\frac{C^*}{4 C_3} (C \sqrt{n} + r)^2 \right) \leq \exp \left( -\frac{C^*}{4 C_3} (C^2 n + r^2) \right) \leq \exp \left( -C''(C^2 n + r^2) \right),
\]
for appropriate constants $C^*, C'' > 0$.

Finally, let $c = \min\{C^*, C'', C'''\}$. Taking things together, we obtain from (17)
\[
\mathbb{P} \left( \frac{1}{T} \left\| Y \Sigma^{-1/2} x \right\|^2 - 1 \geq \frac{\epsilon}{2} \right) \leq 3 \exp \left( -c (C^2 n + r^2) \right),
\]
for the vector $x$ that we fixed.

**Step 3: Union bound.** Notice that we may choose the net $\mathcal{N}$ such that its covering number $\mathcal{N}(S^{n-1}, 1/4)$ is bounded by $(1 + 2/(1/4))^n = 9^n$ by Lemma 5.2 of Vershynin.
We then obtain
\[ P \left( \max_{x \in \mathcal{X}} \left| \frac{1}{T} \left( Y \Sigma^{-1/2} x \right)^2 - 1 \right| \geq \frac{\epsilon}{2} \right) = P \left( \bigcup_{x \in \mathcal{X}} \left( \left| \frac{1}{T} \left( Y \Sigma^{-1/2} x \right)^2 - 1 \right| \geq \frac{\epsilon}{2} \right) \right) \]
\[ \leq 3 \cdot 9^n \exp \left( - c (C^2 n + r^2) \right) \leq 3e^{-cr^2}, \]
where the second inequality follows for $C$ large enough. In other words, if $T = \Omega \left( n^{2/\gamma-1} \right)$ and $n$ is sufficiently large, we have for every $r \geq 0$
\[ P \left( \left\| \frac{1}{T} \Sigma^{-1/2} Y' Y \Sigma^{-1/2} - I_n \right\| \leq C \sqrt{\frac{n}{T} + \frac{r}{\sqrt{T}}} \right) \geq 1 - 3e^{-cr^2}, \]
where $c, C > 0$ are absolute constants.

**Step 4: Translation to the non-isotropic random vectors.** Notice that
\[ \left\| \Sigma^{1/2} \right\|^2 = \sup_{x : \|x\| = 1} \left\| \Sigma^{1/2} x \right\|^2 = \sup_{x : \|x\| = 1} \| x \Sigma x \| = \| \Sigma \|. \]
We then have
\[ \epsilon \| \Sigma \| = \epsilon \left\| \Sigma^{1/2} \right\|^2 \geq \left\| \Sigma^{1/2} \right\| \left\| \frac{1}{T} \Sigma^{-1/2} Y' Y \Sigma^{-1/2} - I_n \right\| \left\| \Sigma^{1/2} \right\| \]
\[ \geq \left\| \frac{1}{T} \Sigma^{1/2} \Sigma^{-1/2} Y' Y \Sigma^{-1/2} \Sigma^{1/2} - \Sigma \right\| = \left\| \frac{1}{T} Y' Y - \Sigma \right\|, \]
with probability at least $1 - 3e^{-cr^2}$. This shows that if $T = \Omega \left( n^{2/\gamma-1} \right)$ and $n$ is sufficiently large, we have for every $r \geq 0$
\[ P \left( \left\| \frac{1}{T} Y' Y - \Sigma \right\| \leq \left( C \sqrt{\frac{n}{T} + \frac{r}{\sqrt{T}}} \right) \| \Sigma \| \right) \geq 1 - 3e^{-cr^2}. \]

As the vectors $\Sigma^{-1/2} Y_t$ are isotropic, $c, C > 0$ are absolute constants and do not depend on $\| \Sigma \|$. Finally, take $r = C \sqrt{n}$ to obtain
\[ \left\| \hat{\Sigma} - \Sigma \right\| \leq 2C \sqrt{\frac{n}{T}} \| \Sigma \| \equiv C' \sqrt{\frac{n}{T}} \| \Sigma \|, \]
with probability at least $1 - 3e^{-cC^2 n} \geq 1 - 3e^{-n}$ for $C$ large enough. □