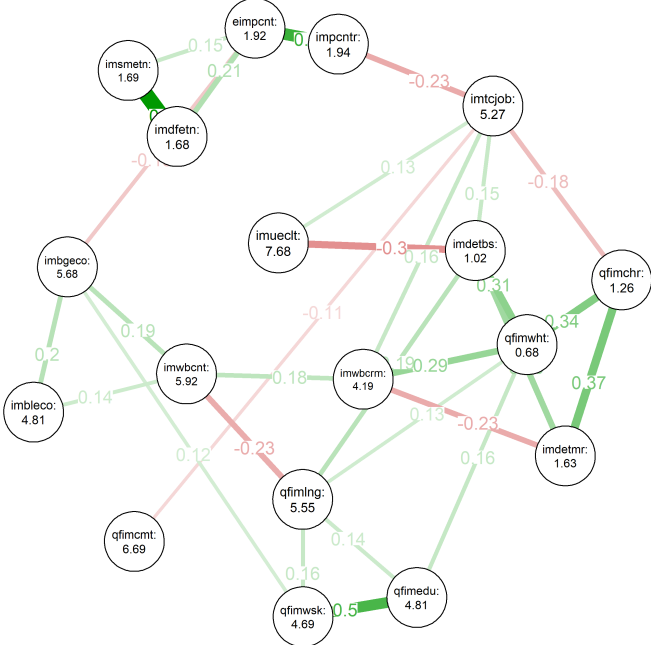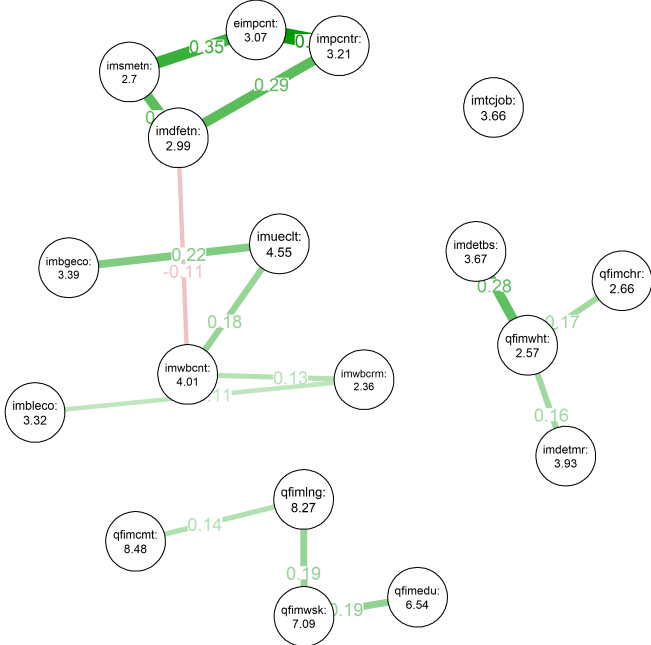# Identification and interpretation of heterogeneous sparse conditional independence structures using Gaussian Mixture Modelling and Gaussian Graphical Modelling

**Vladimir Hazeleger**



Immigration opinions of 94 GroenLinks voters

Immigration opinions of 62 PVV voters

| | |
|---|---|
| Student number: | 6597971 |
| Supervisor (UU): | Dr. Albert A. Salah |
| Second reader: | Dr. Emmeke Aarts |
| Supervisors (external): | Dr. Bob van der Vecht, Dr. Tom Powell |

# Contents

**Abstract**

In research of complex natural phenomena, a challenge arises when the data suggests the presence of subgroups from different distributions. This heterogeneity can be limited by identifying these subgroups. This is a challenge for most established research constructs, especially in clinical settings. For example, in research on major depressive disorder it is widely recognised that there is widespread heterogeneity within the diagnosed population, where different patients can have very different expressions of depressive symptoms. This poses challenges in diagnosis and treatment, for practitioners and researchers alike. Identifying subtypes from a theoretical standpoint requires considerable time and effort. Recently, attention has shifted to potential data-driven methods for the identification of subtypes. Unsupervised learning (or clustering) methods aim to categorise data into clusters. However, such methods require a proper representation of the data to distinguish meaningful clusters. A promising data representation is the conditional independence structure of the data. This structure directly shows conditional independence relations between variables of the data in the form of a network. Different conditional independence structures suggest different underlying data distributions. Therefore, finding heterogeneous conditional independence structures within the data can be indicative of heterogeneous subtypes. In addition, the resulting networks allow for more detailed interpretation through network analyses. To use unsupervised learning in unison with conditional independence structures, two models from the Gaussian model family are combined: the Gaussian Mixture Model for clustering, and the Gaussian Graphical Model for estimation of conditional independence structures. Gaussian Mixture Models perform clustering by fitting a Gaussian component per cluster to the data. Each component, defined by a means vector and a covariance matrix, represents a cluster. Data points are assigned a probability of belonging to each cluster. Gaussian Graphical Models estimate conditional independence structures by computing regularised estimates of partial correlations from the covariance matrix. This results in a sparse network structure, where variables that are not connected by an edge are conditionally independent of each other. By exploiting the direct relation between the covariance matrix and the conditional independence structure, the methods can be combined. This results in clusters that are distinguished by different conditional independence structures. Furthermore, these conditional independence structures are captured in networks that can be analysed and compared. The resulting methodology is referred to as the GGM/GMM method. To assess performance of the GGM/GMM method, it is first tested on simulated data with various group sizes and with various degrees of heterogeneity. Second, the method is applied to two real data sets: the Social Media Disorder data set investigating social media disorder among adolescents, and European Social Survey data measuring public opinion of immigration and refugees. Benchmarking results show that the method is able to correctly estimate conditional independence structures of subgroups, independent of the structural overlap between the groups. However, group size was found to be a limiting factor. Groups with estimated sizes under 400 have estimated conditional independence structures with severe structural errors and instability. For each of the real datasets (social media disorder & public opinion) two models were computed, one for identification of three subgroups, and one for identification of five subgroups. Selecting different models leads to substantively different representations of the subgroups. Finally, limitations of this work are discussed, as well as potential avenues for further research and development of the GGM/GMM method.

# Introduction

In an effort to understand natural phenomena, researchers distinguish them, categorise them, formally define them, and create theories about them. These phenomena are broken down further into parts, and through analysis of the interacting parts, we aim to understand the whole. Theory building is therefore an essential part of any research field. Theory frames the concept, and allows formulation of hypotheses that can be tested (Meredith, 1993). However, theory needs to be specific enough to formulate reasonable hypotheses. Therefore, theories are commonly formulated for very specific constructs. Consider the many distinct types of cancers, viruses, genetic abnormalities, diseases, psychopathologies, behavioural disorders, or public attitudes. All above examples have well-defined research constructs, securely founded in theory.

However, it is widely recognised that there is still heterogeneity within many such research constructs: within psychopathology (e.g. Nandi, Beard, & Galea, 2009; Nigg, Willcutt, Doyle, & Sonuga-Barke, 2005; Wardenaar & de Jonge, 2013), genetics (e.g. Kearns & Losick, 2005; McClellan & King, 2010), oncology (e.g. Fisher, Pusztai, & Swanton, 2013; Lawrence et al., 2013), virology (e.g. Bukh, Miller, & Purcell, 1995), consumer behaviour (e.g. Mueller & Rungie, 2009; Zenetti & Klapper, 2016) and attitude research (e.g. Ford & Lowles, 2016), for example.

Such heterogeneity can be challenging for practitioners and researchers alike: it limits the generality of treatment and makes accurate diagnosis more difficult. One way of handling heterogeneity while also building theory is by defining more specific *subtypes* of a construct (Doty & Glick, 1994). Subtypes further differentiate the construct by specifying how its parts relate to each other. Consider for example attention deficit hyperactivity disorder (ADHD). To set the diagnosis ADHD, mental health practitioners worldwide follow the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013). The DSM-5 considers two categories of symptoms of ADHD: inattention symptoms, and hyperactivity-impulsivity symptoms. Subsequently, the DSM-5 defines three subtypes of ADHD: predominantly inattentive subtype, predominantly hyperactive-impulsive subtype, and a combined subtype. Differentiating between these subtypes helps practitioners recognise a patient's problems, set an accurate diagnosis and provide care based on the patient's needs (Nigg, Blaskey, Huang-Pollock, & Rappley, 2002). In addition, ADHD theory is specified to account for these subtypes and focus research in this area.

## Subtypes

Useful subtypes have two important properties. First, a subtype should be necessary: it should have a clear utility. In a clinical setting, for example, the utility of subtypes of a disease is to improve treatment, by targeting subtypes differently. If different disease subtypes receive the same treatment, there is no benefit to distinguishing the subtypes, other than to further theory. Second, a subtype should be meaningful. Not any arbitrary difference suffices to be classified as a subtype. Any subtype should be meaningful, in the sense that the subtype has discriminatory power. It follows that the discovery of a new subtype is only meaningful if it is coherent: the distinction should be structural and reproducible.

Which subtypes are necessary and meaningful is the topic of ongoing debate. Consider again the DSM example. Up until now, the DSM has known five different iterations, changing some constructs significantly between iterations. For example, between the DSM-IV (American Psychiatric Association, 2000) and the DSM-5, 'autistic disorder', 'Asperger's disorder', 'childhood disintegrative disorder' and 'pervasive developmental disorder not otherwise specified' were combined into one single concept: 'autism spectrum disorder'. At the same time, the DSM-IV diagnosis 'panic disorder and agoraphobia', was separated into two different disorders in the DSM-5. These changes are made to reflect evolving theory and changes in consensus within the psychiatric community. Subtypes in the DSM are defined top-down by specialists in the field, based on years of empirical evidence and best practices.

In recent years, there has been a call for more structural, data-driven identification of subtypes in psychopathology to address the difficulty of identifying subtypes (e.g. Benjamins et al., 2017; Bora, Aydın, Saraç, Kadak, & Köse, 2017; Lombardo et al., 2016). Traditional, top-down approaches such as the approach of the DSM build subtypes from theory, starting with an hypothesis and collecting empirical data in support of the hypothesis. The hypothesis is then statistically tested and accepted or rejected based on the result. In contrast, data-driven (or bottom-up) approaches start by collecting data and then aim to learn patterns from this data. These patterns can subsequently be used to formulate new hypotheses and build theory (Oquendo et al., 2012).

# Unsupervised learning

More specifically, *unsupervised learning*, or clustering, is a data-driven method used to identify groups, or clusters, in the data. Unsupervised learning can be used to identify subtypes. For example, consider the work of Firoozbakht, Rezaeian, Porter, and Rueda (2014), who identify subtypes of breast cancer by learning the most informative biomarkers from genetic data.

Different clustering methods use different information to perform this grouping. For example, $k$-means clustering (Hartigan & Wong, 1979) groups data points based on their Euclidian distance from one another. Data points are placed in $p$-dimesional space, with $p$ being the number of variables in the data. After initialising the algorithm with $k$ random means, the algorithm iteratively assigns each point to the cluster with the nearest mean, then recalculates the means. This process is repeated until means no longer change. Gaussian Mixture Models (Reynolds, 2009), on the other hand, perform clustering based on the Gaussian distribution of the data. The model is initialised with $k$ Gaussian distributions, known as components. Then, for each point, probabilities of belonging to each component are computed, after which the distributions of the components are updated. The end result is slightly different: data points are assigned a *probability* of belonging to each of $k$ clusters (also known as *soft clustering*). For an illustration of both models, see Figure 1.1. In sum, what distinguishes the groups from one another is dependent on the clustering method and the data representation that is used.

However, unsupervised learning has a number of weaknesses. First, unsupervised learning methods are notoriously dependent on the data representation used (Bengio, Courville, & Vincent, 2013). For example, representing the variables of the data as scales or compound variables (e.g. through principal component analysis) will lead to different results than using the raw data. This is due to the fact that some information is lost when variables are combined or reduced. Consequently, this information cannot be used to distinguish clusters from each other. However, selecting a good data representation for the given clustering problem reduces noise and improves performance (Buhmann, 1995). Variable selection (i.e. selecting which variables are used for learning) ties in to the above. Variables that explain more variance than others may improve clustering performance, while variables with little or no explanatory power variables will reduce clustering performance by introducing noise. Similarly, introducing a variable with high explanatory power can render other variables redundant. This is the problem of spurious correlations (Simon, 1954). A famous example of spurious correlations is the observation that the number of ice cream sales is highly predictive of the number of drownings. However, if the variable 'temperature' were included in this prediction, ice cream sales would lose all predictive power for drownings. Understanding such spurious correlations often requires understanding of the context. As a consequence, learning algorithms are generally bad at selecting the relevant variables by themselves, as they do not naturally have an internal representation of how variables relate to each other, or to the learning problem. This can lead to a number of problems, such as algorithmic bias (Hajian, Bonchi, & Castillo, 2016). As a consequence, it falls upon the researcher to decide on relevant variables and data representation to feed into the algorithm.
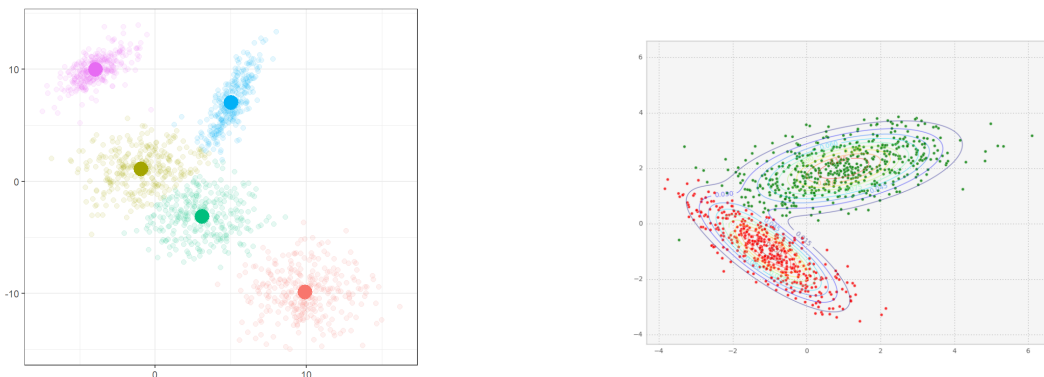


(a) Example of a k-means clustering model with $k = 5$ (Lim, 2018).

(b) Example of a Gaussian mixture model with $k = 2$ (Ramalho, 2014).

Figure 1.1: Examples of k-means clustering and Gaussian mixture modelling for two-dimensional data.

Second, unsupervised learning does not necessarily lead to interpretable results. For example, consider again the clustering method *k*-means clustering. To reiterate, this method groups data points based on their distance to each other. By definition, the resulting *k* clusters have the smallest possible within-cluster variance and the highest between-cluster variance for the data. Data points within the same cluster are close to each other, while the clusters themselves are far apart from each other. However, to determine whether clusters are empirically different, a statistical hypothesis test needs to be performed. In addition, the resulting cluster statistics do not provide much information as to *why* the clusters turned out to be different. In fact, this problem is the main focus of the field of explainable artificial intelligence (Holzinger, Biemann, Pattichis, & Kell, 2017). Specifically, Holzinger et al. (2017) define comprehensible AI as "systems that emit symbols enabling user-driven explanations of how a conclusion is reached". By this definition, the above example of k-means clustering is not comprehensible.

In sum, using unsupervised learning for the identification of subtypes requires deliberate consideration of the data representation used for learning, and the explainability of the results. Here, we argue that a choosing a suitable data representation could mitigate both weaknesses.

# Conditional independence structure

A promising data representation for identifying subtypes is the *conditional independence structure* (CIS; Baioletti, Busanello, & Vantaggi, 2009; Studeny, 2006; Wermuth & Lauritzen, 1990; Wille & Bühlmann, 2006). The conditional independence structure of the data represents the conditional dependencies and independencies between the variables. It can be visualised as a graph $G = (V, E)$ with node set $V$ and edge set $E$. The nodes represent the variables, and an edge connecting two nodes represents a dependence relation. The absence of an edge, then, represents independence between the variables, conditional on all other variables. The conditional independence structure is not an inherent property of the data. It needs to be estimated from the data using an appropriate method. In the following, we argue that different subtypes have different conditional independence structures.

Consider again the ADHD example. The CIS of ADHD shows how symptoms of ADHD are (in)dependent of each other. Recall that the DSM-5 defines three subtypes of ADHD: the predominantly inattentive subtype, the predominantly hyperactive-impulsive subtype, and the combined subtype. It is reasonable to expect the CIS of each subtype to be different, as the inattentive subtype is more dependent on the inattentive symptoms and the hyperactive-impulsive subtype is more dependent on the hyperactive-impulsive symptoms. Figure 1.2 visualises this expectation in graphs. Notice how both the predominantly inattentive and predominantly hyperactive-impulsive subtypes are densely connected *within* the corresponding symptoms, while the combined subtype shows more connections *between* the two categories of symptoms.

This graphical representation illustrates structural differences between the subtypes. Comparing the network structures of subtypes helps to interpret why and how subtypes differ. In addition, these networks can be analysed using methods from graph theory. Graph theoretical measures examine properties of the graph, such as the importance of nodes (e.g. centrality indices, Koschützki et al., 2005), or the relations between nodes (e.g. shortest path analyses, Dijkstra et al., 1959). In short, graphs of heterogeneous



(a) Dummy CIS of predominantly inattentive subtype.

(b) Dummy CIS of predominantly hyperactive-impulsive subtype.
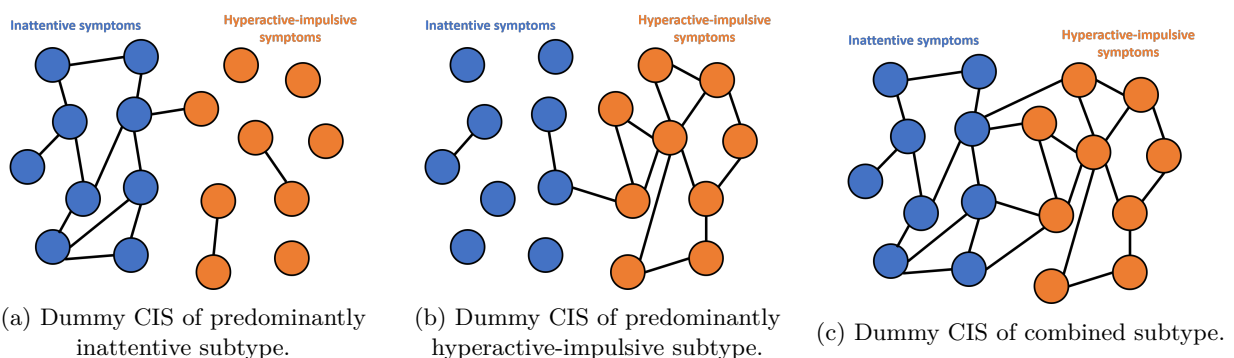
(c) Dummy CIS of combined subtype.

Figure 1.2: Expectation of differing (dummy) CISs for different subtypes of ADHD. Blue nodes represent inattentive symptoms, orange nodes represent hyperactive-impulsive symptoms. Edges represent conditional dependence between variables.

conditional independence structures can be used to represent and interpret different subtypes.

In sum, we aim to use unsupervised learning of heterogeneous conditional independence structures as a method for finding subtypes of a construct. The network representing the CIS then allows for analysis, interpretation and comparison of the learned subtypes. This overcomes a major weakness of unsupervised learning, namely the lack of explainability, and effectively combines identification and interpretation of subtypes.

# Clustering by conditional independence structure

Separate methods exist for the estimation of CISs and for clustering. However, the challenge lies in combining two methods, such that the conditional independence structure is used as a feature of clustering.

Hill and Mukherjee (2013) suggest combining Gaussian Mixture Modelling (GMM) for clustering and Gaussian Graphical Modelling (GGM) for estimation of conditional independence networks. Both methods are built on principles of the Gaussian distribution and therefore share components that make them compatible. See the Related Work section for a detailed description of the methods and see the Methods section for mathematical definitions of both methods.

The Gaussian Mixture Model is a probabilistic clustering method. It estimates $k$ different clusters by estimating a unique Gaussian distribution (i.e. component) for each cluster. Two parameters define a Gaussian component: the means vector and the covariance matrix (see Methods section for a mathematical description of the GMM). The model produces 'soft' cluster assignments: a probabilistic expression of cluster assignment. For estimation, it uses the expectation-maximisation (EM) algorithm (A. P. Dempster, Laird, & Rubin, 1977). The EM algorithm iteratively computes an expectation of the log-likelihood of cluster assignments using the current parameters, then updates the parameters by maximising the log-likelihood. The result is a (local) maximum likelihood estimate of the Gaussian components and soft cluster assignments for all data points.

The Gaussian Graphical Model is used to estimate conditional independence structures for variables that follow a Gaussian distribution. Specifically, it estimates partial correlations between variables. A partial correlation of exactly zero indicates independence between two variables, conditional on all other variables in the model. Partial correlations can be directly calculated from the off-diagonal elements of the precision matrix - the inverse of the covariance matrix. However, due to numerical precision, a direct inversion of the covariance matrix produces a precision matrix with no values exactly zero, and thus no conditional independence in the model. Therefore, GGMs are often estimated using regularisation: a technique to reduce the smallest parameters in a model to exactly zero. This results in a sparse conditional independence structure where only a few parameters are non-zero. Such models are easier to interpret and are often more accurate (Epskamp, Borsboom, & Fried, 2018).

The above methods are compatible due to their reliance on the covariance matrix. Within the Gaussian mixture model, we replace the default update of the variance-covariance matrix with a variance-covariance matrix estimated using the regularised Gaussian graphical model. For the resulting clusters, we draw networks representing their conditional independence structures.

By combining the Gaussian Mixture Model with the Gaussian Graphical Model, we develop a promising data-driven tool for the simultaneous identification and interpretation of subtypes in heterogeneous data. The resulting clusters represent different subtypes, that can be investigated and analysed through their network representation, which shows subtype-specific conditional independence structures. The new method is further referred to as the GMM/GGM method.

# Application of the methodology

While the focus of this thesis lies on the methodology, its value is best highlighted by applying the GMM/GGM method to specific domains. To this end, the GMM/GGM method is applied to two domains that have an interest in identifying and interpreting subtypes.

## Social media disorder

The first domain is within adolescent psychology. Specifically, we identify subtypes in data collected as part of the Digital Youth project, a 5-year longitudinal study on the role of social media and gaming in the lives of Dutch adolescents (Van Den Eijnden, Koning, Doornwaard, Van Gurp, & Ter Bogt, 2018). The aim

of the project is to advance theory on the disordered video game and social media use among adolescents, dubbed social media disorder (SMD). SMD symptoms were shown to have a negative effect on life satisfaction and perceived social competence. As SMD research is relatively novel, the structure of the related variables has not yet been investigated extensively. In addition, Van Den Eijnden et al. (2018) note as one of the limitations that the sample suffers from heterogeneity, which may have impacted the results. Furthermore, the SMD data is categorised into one of three categories representing levels of SMD: a normal group, a risky group, and a problematic group.

Application of the proposed method to the SMD data can address these limitations. Specifically, we focus on the question of heterogeneity. We use the method to generate a model with three subgroups, hypothesising that we find subgroups that correspond to the three levels of SMD. In addition, we generate another model, this time identifying five subgroups. This model serves as a comparison to the first model, to assess whether further clustering will yield different results. We hypothesise that one or more of the subgroups in the three-group model will be further broken up into two subgroups in the five-group model.

## Public opinion of immigration and refugees

The second domain is that of public opinion. Specifically, we identify subtypes in data from the European Social Survey, a bi-annual survey among European countries, polling public opinion on various societal and socio-economical topics, such as politics, media and social trust, health and inequality, justice and immigration (ESS round 7, 2014).

To limit the scope of the research, we focus on public opinion of immigration and refugees. This topic has seen considerable political and societal attention in the past decade, mainly due to societal crises such as the European migrant crisis (Poddar, 2016), the Rohingya refugee crisis in Bangladesh (Parnini, Othman, & Ghazali, 2013) and the migration crisis in Central America (Isacson, Meyer, & Morales, 2014). In fact, anti-immigration has been the focal point of recent political campaigns, most notably the Brexit campaign in the UK (Currie, 2016), and the 2016 presidential campaign of Donald Trump in the US (Martin, 2017), both of which have received major popular support.

Unraveling public opinion on this complex topic is no easy task. Public opinion is heterogeneous by definition. The endeavour of discovering profiles of public opinions has been labeled *attitudinal segmentation* (H. Dempster & Hargrave, 2017). For example, Purpose (2017) identified five different profiles of German public opinion regarding immigrants. However, H. Dempster and Hargrave (2017) review that most work on attitudinal segmentation finds three general profiles: a favourable group, an opposed group, and an ambivalent or conflicted middle group.

Application of the GGM/GMM method performs attitudinal segmentation on the large ESS data. Compared to previous work on attitudinal segmentation, the GGM/GMM method adds interpretability to the results, which may provide insight into existing ambiguity. Again, two models are generated for the ESS data: a three-group model and a five-group model. We hypothesise that the three-group model will be representative of the three profiles described above, namely favourable, opposed, and ambivalent/conflicted. In addition, we expect that the five-group model will further nuance the ambivalent/conflicted group.

# Problem definition

In this thesis, I focus on data-driven identification and interpretation of heterogeneous subtypes. To this end, I develop a clustering method that distinguishes clusters based on their conditional independence structure. Each cluster is seen as a subtype, and can be interpreted by analysing its conditional independence structure, revealing how variables interact. Building on the work of Hill and Mukherjee (2013), this method combines Gaussian Graphical Modelling and Gaussian Mixture Modelling, and introduces a new model selection criterion that is adapted for selection of cluster-specific sparse conditional independence structures. The GGM/GMM method is benchmarked and tested on artificial data, to assess method performance under varying circumstances. Finally, the method is used to generate models with three and five subgroups, for data in two domains: social media disorder and the public opinion of immigration. For both domains, results are expected to address existing questions of heterogeneity (*what are the different subtypes?*) and explainability (*why and how are subtypes different from each other?*).

# Related work

In this section, fundamental concepts and theory preceding the current investigation are outlined and reviewed. Two core concepts will be explored: Gaussian graphical modelling (GGM) and Gaussian mixture modelling (GMM). GGM is a method for estimation of conditional independence structures. GMM is a clustering method that fits Gaussian distributions to the data in order to distinguish clusters. Both methods rely on the variance of the data. By exploring both methods, I outline how they can be used in unison in order to perform simultaneous identification and interpretation of subtypes.

## Gaussian Graphical Modelling

*Gaussian graphical models* (GGM; Lauritzen, 1996) are also known as covariance selection models (A. P. Dempster, 1972), concentration graph models (Drton & Perlman, 2004), pairwise Markov random fields (Epskamp, Maris, Waldorp, & Borsboom, 2016) or regularised partial correlation networks (Epskamp & Fried, 2018). Despite varying nomenclature, the models have the same purpose: modeling conditional independencies between variables in the form of a graph. This results in a network where nodes represent variables and edges represent the partial correlation. The partial correlation represents the dependence relation between variables, controlled for all other variables in the model. Notably, a partial correlation of exactly zero implies conditional independence between variables. In the graph, this can be easily seen, as there is no edge between variables if their partial correlation is zero. This property of GGMs makes them an attractive method to inspect the structure of variables in a holistic manner.

The GGM and its graph need to be estimated from the data. Estimation, however, is not exact. Due to the high dimensionality of the data and the computational precision of the estimation method, a naive estimate of the partial correlation will almost always be non-zero. As a consequence, the graph will be fully connected: all nodes are connected to all other nodes, which implies that every variable is dependent on every other variable. However, it is highly likely this is not true in reality. The real structure is most likely sparse: it contains only a few edges. To eliminate false positives (a.k.a. spurious edges) and encourage sparsity, regularisation is applied to the estimation of the model. Regularisation techniques penalise the smallest estimates in order to reduce them to exactly zero, and remove them from the model. As a consequence, a regularised GGM will have a sparse structure. This makes interpretation easier and simultaneously reduces the amount of parameters in the model (i.e. the model complexity).

Over the past decades, GGMs started to gain traction in the fields of biometrics and computational biology (e.g. A. P. Dempster, 1972; Wong, Carter, & Kohn, 2003). The model proved powerful in handling high-dimensional biological data such as gene expression data (e.g. Schäfer & Strimmer, 2005; Toh & Horimoto, 2002). The method was quickly adopted by other fields in natural sciences, social sciences and economics, such as speech recognition (Bell & King, 2007), finance flows (Giudici & Spelta, 2016), cancer research (Zhao & Duan, 2019) and attitude research (Dalege et al., 2016). An enormous body of methodological literature on GGMs exists, covering estimation methods, regularisation methods, model selection methods and validation methods. The exact methods used for this thesis are described in the Methods section. The remainder of this section is structured as follows: first, it will provide a brief overview of various GGM methodologies, focusing on the unique characteristics of various model components. Second, it will review some of the more prominent applications of GGMs in recent works.

### Usage

Starting with an arbitrary data set, the pipeline for Gaussian graphical modelling roughly follows the following steps: (1) estimation of partial correlations between variables using (2) regularisation, followed by (3) a model selection procedure. For each of these steps, various techniques have been proposed, each applicable to a distinct problem. For an overview, see table 2.1.

#### Estimation

Gaussian graphical models rely on the covariance matrix of the data for estimation of partial correlations (Epskamp et al., 2018). To estimate partial correlations from the covariance matrix, two methods are proposed.

| Parameter estimated | Estimation method | Regularisation method | Model selection |
|---|---|---|---|
| Zero-order correlation | Correlation matrix (Costantini et al., 2015) | - | - |
| Partial correlation | Inverse covariance matrix (A. P. Dempster, 1972) | - | - |
| Regularised partial correlation | Inverse covariance matrix (A. P. Dempster, 1972) | Confidence intervals (Williams & Rast, 2018) | EBIC (Chen & Chen, 2008) |
| | | Graphical LASSO (Friedman et al., 2008) | |
| | | eLASSO (van Borkulo et al., 2014) | |
| | Node-wise regression (Meinshausen & Bühlmann, 2006) | LASSO (Tibshirani, 1996) | Cross-validation (Krämer et al., 2009) |
| | | Adaptive LASSO (Zou, 2006) | |

Table 2.1: Overview of methods used for estimating Gaussian Graphical Models. Dotted lines indicate interchangeability.

The first, proposed by A. P. Dempster (1972), uses an interesting property of the covariance matrix. The inverse of the covariance matrix, known as the precision matrix, can be used to obtain partial correlations. Specifically, an entry $(i, j)$ in the precision matrix is equal to exactly zero if and only if $i$ is conditionally independent of $j$, given all other variables. Non-zero entries $\omega_{ij} = (i, j)$ in the precision matrix can be used to calculate the value of the partial correlation $\rho_{ij}$ between $i$ and $j$ using the formula $\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$.

However, the above only holds in the case that all variables follow a multivariate Gaussian (i.e. normal) distribution and the precision matrix is positive definite. In most cases, Gaussianity is simply assumed and slight violation of the assumption is not problematic (Knief & Forstmeier, 2018). In the case that the data is binary or ordinal, the data are known not to be normally distributed and an adaptation has to be made.

For binary data, the Ising model can be applied (van Borkulo et al., 2014). The Ising model (Ising, 1925) was originally developed as a mathematical model of ferromagnetism that describes the spin of particles in a square grid known as a lattice graph. In the graph, nodes (representing particles), have a spin value of +1 or -1 and edges (representing interaction relations) indicate agreement, disagreement, or non-interaction between nodes. The model can then be used to compute the probability of different configurations based on the interactions. Despite apparent simplicity, the Ising model is a non-trivial description of pairwise interactions between binary random variables, that has since been adopted to fields outside theoretical physics. For binary data, the Ising model can perform the same functions as the GGM (van Borkulo et al., 2014).

For ordinal data, Epskamp (2016a) suggests using polychoric and polyserial correlations (Olsson, 1979; Olsson, Drasgow, & Dorans, 1982), which are compatible with the GGM as long as the resulting covariance matrix is positive definite. In case the covariance matrix is not positive definite, it can be coerced to become so (Bates & Maechler, 2019).

The second method to estimate partial correlations is through node-wise regressions, suggested by Meinshausen and Bühlmann (2006). Using (regularised) regression, this method estimates the set of direct neighbours of a node, for every node in the graph. To illustrate, let $\{A, B, C, D\}$ be the set of nodes. The neighbours of node $A$ are then estimated by regressing $A$ on $B, C$ and $D$, resulting in relations between nodes, e.g. $A \rightarrow B$. This process is repeated for $B, C$ and $D$. Finally, for each pair of relations $A \rightarrow B$ and $B \rightarrow A$, the average value is taken for the final model. Interestingly, this leads to the exact same estimates as the first method (Epskamp & Fried, 2018).

For this thesis the first method is employed, as it is slightly faster and applicable in the Gaussian Mixture Model described below. A more formal description of the estimation method can be found under the Methods section.

## Regularisation

The above methods compute an optimal fit of the GGM parameters to the provided data. However, a well-known problem with estimations of this kind is overfitting (Hawkins, 2004). An overfit model fits well on the sample used to estimate the model, but fits poorly on unseen observations. This problem scales with dimensionality: the more variables are included in the model, the more likely it is to be overfit.

To prevent overfitting, *regularisation techniques* can be applied to the estimation of the model. Regularisation techniques add a penalty term to the equation of parameter estimation. The degree of penalty scales with model complexity: the more parameters are non-zero, the higher the penalty term. Some regularisation techniques are able to penalise, or 'shrink', parameters to exactly zero, causing them to be excluded from the model entirely. This property of regularisation results in *sparse* estimates, a feature that is important in the estimation of GGMs (Epskamp et al., 2018). The following paragraphs outline a number of regularisation techniques used in estimating GGMs.

**LASSO.** A well-known regularisation technique in statistics and machine learning is the *least absolute shrinkage and selection operator* (LASSO, Tibshirani, 1996). Originally introduced for regression models, the LASSO penalises the sum of absolute values of all regression coefficients in the model, the so-called $l_1$ penalty. In other words, the LASSO penalises model complexity. The more and larger the regression coefficients, the higher the penalty that is applied to each coefficient. Furthermore, because the LASSO uses the absolute values, one of its properties is that coefficients can be reduced to exactly zero. To determine the degree of penalty, it employs a regularisation parameter $\lambda$. The higher the value of $\lambda$, the more coefficients are penalised, and more coefficients become exactly zero. Note that in the case of regression, LASSO regularisation simultaneously performs variable selection, as variables with a coefficient of zero are dropped from the model.

**Adaptive LASSO.** Multiple variations of LASSO regularisation have been suggested. First is the *adaptive LASSO*, proposed by Zou (2006). It is similar to the LASSO, but adds a weighting to the penalty term that is proportional to the coefficient value. Thus, large coefficients are punished more than small coefficients. This expands the consistency of the regularisation to scenarios where the regular LASSO would not be applicable. In estimating partial correlation networks, it outperforms the regular LASSO, more so when the true underlying structure is sparse (Costantini et al., 2015).

**glasso.** Second is the *graphical LASSO*, also known as simply *glasso*, introduced by Friedman et al. (2008). It was introduced specifically for estimating sparse graphs from the inverse covariance matrix. Specifically, it applies a penalty to the sum of absolute values of the precision matrix, the $\ell_1$ penalty. Similar to the regular LASSO, this leads to some estimates that are exactly zero. In this case, however, it is the values of entries of the precision matrix that are shrunk. Recall that any two variables are conditionally independent of each other given all other variables if and only if their entry in the precision matrix is exactly zero. In short, the glasso penalises model complexity and simultaneously labels some variable pairs as conditionally independent.

**eLASSO.** The third variation on the LASSO algorithm is the eLASSO, introduced by van Borkulo et al. (2014). This is an adaptation of the LASSO aimed specifically at binary data, for use with the Ising model. The main challenge here is that network models with discrete, binary data behave as discrete Markov Random Fields, which are known to be computationally intractable. To estimate such networks, van Borkulo et al. (2014) apply a $\ell_1$ (LASSO) penalty to node-wise *logistic* regressions, which apply to binary data. This method allows regularised estimation of Ising models for binary data.

**Confidence intervals.** Recently, a different method was proposed outside of the LASSO family. Williams and Rast (2018) propose a tried and trusted statistic as a method of regularisation: the confidence interval. They note that the highly used $\ell_1$ penalty using the glasso has a number of limitations. First, the accuracy of $\ell_1$ methods does not necessarily generalise to all settings, as for example highly correlated variables are not captured well. Second, they note that estimation error does not diminish with increasing sample size, unlike most Bayesian or frequentist approaches. Third, the authors highlight that statistical inference from a $\ell_1$ based model is not straight forward. Inclusion of some variable or parameter does not entail significance, and exclusion does not entail no effect: these inferences require formal hypothesis testing. In addition, the data-driven variable selection leads to model selection bias. Williams and Rast (2018) therefore propose a method that circumvents many of these limitations and has a closed form solution. The method involves the entries of the precision matrix, similar to the glasso approach. However, the entries are first standardised using Fisher-Z transformations. Subsequently, two-tailed $100(1-\alpha)\%$ confidence intervals of the estimates are defined, where $\alpha$ is manually set. The set of edges is then obtained from those estimates whose confidence

interval excludes zero. In practice this means that if $\alpha = 0.05$, there is a maximum 5% chance of false positives, and 95% of cases will have a true value in that interval range. They find that this latter percentage, the coverage probability, is identical to the specificity (or true negative rate) of the model, meaning that the specificity of the model can be manually selected. In contrast, they show that the glasso approach strongly declines in specificity as the underlying true networks become more densely connected. However, the glasso outperforms the CI approach in terms of sensitivity, or true positive rate. In sum, the CI method errs on the side of caution, aiming to be more rigid in uncovering true network structure, rather than aiming for discovery and predictive accuracy.

For this thesis, the regularisation method of choice is the glasso. This has two main reasons: first, the glasso works well in the *discovery* of sparse structures, which is the goal in the identification of multiple sparse networks through clustering. Second, the glasso is a fast algorithm that works well in conjunction with the Gaussian Mixture Model discussed below. A more detailed description of the glasso algorithm can be found under the Methods section.

## Model selection

To find the optimal value of the regularisation parameter, typically a large number of models is estimated for varying values of that parameter. The optimal model is then selected based on a model selection metric. Various such model selection metrics have been suggested for GGMs.

**Cross-validation.** A widely used model selection metric in statistics and machine learning is *cross-validation* (Krämer et al., 2009; Shao, 1993). For cross-validation, the dataset is divided into a training set and a test set. The model is learned from the training set, after which its predictive power is tested on the test set. This can be expanded to *k-fold cross-validation* (Mosteller & Tukey, 1968), where the data is randomly divided into $k$ sets of equal size, and each set is used as a test set once. This results in $k$ models that optimise the value of $\lambda$ for a training set of size $k-1$. The resulting $k$ models are then summarised, usually by taking the mean of the estimated values of $\lambda$ (Krämer et al., 2009).

$k$-fold cross-validation requires some consideration. First, the value of the hyperparameter $k$ is non-trivial. While $k = 10$ is common in machine learning applications (Anguita, Ghelardoni, Ghio, Oneto, & Ridella, 2012), the choice is somewhat arbitrary. Second, Gaussian graphical models are sensitive to changes in sample size (Foygel & Drton, 2010). A test set of size $n/k$, even for large $n$ and $k = 10$, may lead to unrepresentative test sets. This can cause bias in estimating $\lambda$.

**AIC.** The *Akaike Information Criterion* (AIC; Akaike, 1974) is a well-known information criterion for model selection. In contrast to cross-validation, information criteria weigh model complexity against goodness-of-fit in an effort to select the optimal model. The AIC is defined as

$$\text{AIC} = 2k - 2\log\left(\hat{\mathcal{L}}\right)$$

where $k$ denotes the number of parameters in the model and $\hat{\mathcal{L}}$ denotes the maximum likelihood estimate, a goodness-of-fit estimator. From this equation, it is easy to see that the AIC punishes model complexity and rewards goodness-of-fit: a lower value of the AIC implies a model with better trade-off between complexity and accuracy.

**BIC.** The *Bayesian Information Criterion* (BIC) is quite similar to the AIC. In fact, both information criteria can be used interchangably. The BIC is given by

$$\text{BIC} = k\log n - 2\log\left(\hat{\mathcal{L}}\right)$$

where $k$ and $\hat{\mathcal{L}}$ are the same as in the AIC, and $n$ denotes the sample size. The only difference is in the first term: for the BIC, the complexity penalty grows with sample size, where for the AIC, it does not. As a consequence, the BIC penalises model complexity more heavily and is thus better suited in settings where false positives are to be avoided.

**EBIC.** The *extended Bayesian Information Criterion* (EBIC) is a variation of the BIC aimed at selecting models from a very large candidate model space (Chen & Chen, 2008; Foygel & Drton, 2010). This is accomplished by adding another penalty term that targets the size of the model space, i.e. the number of candidate models with the same number of dimensions. The formula is given by

$$\text{EBIC}_\gamma = k\log n - 2\log\left(\hat{\mathcal{L}}\right) + 2\gamma\mathcal{S}_p$$

where $\mathcal{S}_p$ denotes the size of the model space with $p$ variables. For Gaussian graphical models, the size of the model space with equal dimensions is given by $p^{2k}$, where $p$ denotes the number of variables in the model and $k$ denotes the number of edges to be estimated (i.e. the dimensionality). The formula can be rewritten as

$$\text{EBIC}_\gamma = -2\log(\hat{\mathcal{L}}) + k\log(n) + 4\,\gamma\,p\log(k)$$

which is also noted by Epskamp and Fried (2018); Foygel and Drton (2010).

Chen and Chen (2008) further show that the EBIC further reduces false positives, thereby performing better at estimation of sparse structures. How much penalty is applied, is defined by the new parameter $\gamma$: this hyperparameter still needs to be set manually. For GGMs, the experimental value suggested by Epskamp et al. (2018) is $\gamma = 0.5$.

For this thesis, the model selection method of choice is the EBIC, for two main reasons. First, the EBIC is the most conservative, as it penalises model complexity most. This is beneficial as the goal is to find sparse structures. In addition, the aim is to find sparse structures for multiple unknown clusters in the data. This reduces the sample size of each cluster, which increases the risk of false positives. By using the EBIC, this risk can be minimised. Second, the EBIC can be easily extended to a mixture form, where it is used to select the optimal model for multiple groups simultaneously. This extension is elaborated further under the Methods section.

## Applications of GGMs

GGMs have been applied in various different fields, mainly because of their relative simplicity and easily interpretable graphical representation.

**Bioinformatics.** As GGMs excel at high-dimensional modelling, they have been extensively used in the field of bioinformatics, particularly in genomics. Dobra et al. (2004) used GGMs to explore and evaluate interactions and conditional independencies between gene expressions of breast cancer pathways. Wille et al. (2004) use GGMs to investigate two pathways of isoprenoid biosynthesis in thale cress and find candidate genes that connect the two pathways. Chu, Weiss, Carey, and Raby (2009) then use GGMs to specifically investigate a target gene of interest, map its neighbours, and use this to infer an integrated gene activation pattern in the pathogeny of asthma. Ma, Gong, and Bohnert (2007) more broadly investigated thale cress by modelling gene expressions of 6 760 genes, resulting in a network with 18 625 edges. They subsequently found coherent subnetworks representing metabolic functions, stress responses and regulatory functions. In a different application, Lee et al. (2019) used GGMs to assess the impact of external effects on complex metabolic processes, by capturing changes in interactions of metabolites. Zhao and Duan (2019) use the Cancer Genome Atlas to learn interactions of genes in cancer cells, and compare 15 specific types of cancer. They find a hypothesised common set of genes that is implicated in human carcinogenesis, and new gene interactions unique to cancer cells, paving the way for effective, targeted research.

In sum, GGMs are useful in bioinformatics to overcome the extremely high dimensionality of some biological data, while also providing insight into the structural relationships between variables in the data.

**Social sciences.** In social sciences, and in particular in psychology, the network methodology provided by GGMs has seen a steady increase in popularity. In lieu of the reproduction crisis, GGM methodology has been adopted as an alternative to then-dominant latent variable modelling. Specifically, the assumption is that human behaviour is an emergent property of causal interaction between (psychological and non-psychological) variables. The GGM, then, provides a means of discovering and investigating this complex, causal interplay of variables. Dubbed *psychological networks* by Epskamp et al. (2018), this method has seen numerous methodological and applied publications. For example, multiple tutorials exist for psychological networks (e.g. Costantini et al., 2015, 2019; Dalege, Borsboom, van Harreveld, & van der Maas, 2017; Epskamp et al., 2018; Epskamp & Fried, 2018; Hevey, 2018; Jones, Mair, & McNally, 2018), providing step by step instructions and examples for the application and interpretation of network models. Additionally, the psychological network methodology has been adapted to challenges common to psychology research, such as ordinal data (Epskamp, 2016a) and the relation to the reigning paradigm, latent variable modelling (Epskamp, Rhemtulla, & Borsboom, 2017).

Unsurprisingly, the method has been broadly applied to problems in psychology, especially in psychopathology (Borsboom & Cramer, 2013; Contreras, Nieto, Valiente, Espinosa, & Vazquez, 2019). Armour, Fried, Deserno, Tsai, and Pietrzak (2017) investigate the structure of post-traumatic stress disorder

(PTSD) symptoms in order to find highly comorbid symptoms and possible targets for intervention. Beard et al. (2016) conduct a similar investigation for symptoms of depression and anxiety to find the most central symptoms, while Fried, Epskamp, Nesse, Tuerlinckx, and Borsboom (2016) use networks to analyse the discrepancy between formally defined clinical symptoms of depression and empirically assessed symptoms of depression. Both studies plead for more focus on the central symptoms of depressive disorders: sad mood and worry. Forrest, Jones, Ortiz, and Smith (2018) and Smith et al. (2018) perform network analyses for eating disorders, finding that the network approach is able to highlight symptoms that may be core targets of clinical intervention.

Outside psychopathology, most applications of psychological network models can be found in personality and attitude research. Cramer et al. (2012) challenge the latent variable view of personality and posit that personality dimensions arise from the network structure of personality components. Costantini et al. (2015) provide a framework for applying network analysis to a common personality questionnaire, and Costantini and Perugini (2016) apply this notion to the personality dimension conscientiousness. In attitude research, Dalege et al. (2016) outline a novel theoretical framework for attitude research: the Causal Attitude Network (CAN) model. It is based on the same principle as psychological networks in general: variables interact directly with each other. For the CAN model, an attitude is a network of causally interacting evaluative reactions toward an attitude object. Dalege et al. (2017) provide an example of how the CAN model can be applied to practical settings.

# Gaussian Mixture Modelling

Gaussian mixture models (GMM) fall within the broad family of model-based clustering methods, aimed at unsupervised learning of cluster assignments. The main premise is that each cluster, or component, follows a unique Gaussian distribution, hence the full dataset is a mixture of Gaussians. The task is then to estimate the Gaussian components and calculate probabilities of belonging to a component for each case in the data. First, this section will briefly discuss methodological considerations in applications of GMMs. (However, a full methodological description applicable to the current research is given under the Methods section.) Second, typical applications of GMMs will be discussed.

## Usage

A concise summary of Gaussian Mixture Models is given by Reynolds (2009). GMMs are represented as a weighted sum of Gaussian components. For any dataset, let $n$ denote the rows and $p$ the columns (or variables) of that set. Let $K$ denote the number of components that will be modelled. Each component $k$ in the GMM can be seen as a Gaussian probability density function defined by its mean vector (size $p$) and covariance matrix (size $p \times p$).

### Design

The design of GMMs has two main considerations. First is the dimensionality of the covariance matrices. Constraining the covariance matrices to be tied, diagonal, or spherical reduces the number of parameters to
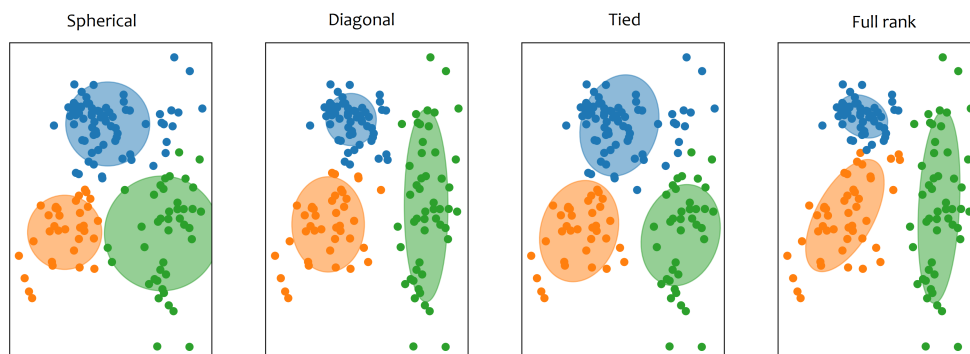


Figure 2.1: Illustration of Gaussian mixture model components for varying types of covariance matrices. Image retrieved from Müller (2018).

be estimated. However, this also changes the shapes of the components (see Figure 2.1) which in turn could impact the accuracy of the model.

Second is model selection: what number of components should be fitted? This parameter needs to be either manually set, or learned. Manually setting this parameter is highly dependent on the domain and the context. Several classes of methods have been suggested for learning the optimal number of components: maximum likelihood estimations (e.g. Keribin, 2000; Leroux, 1992; Roeder & Wasserman, 1997; Verbeek, Vlassis, & Kröse, 2003), distance measures (e.g. Chen & Kalbfleisch, 1996; James, Priebe, & Marchette, 2001; Woo & Sriram, 2006), and Bayesian approaches (e.g. Constantinopoulos, Titsias, & Likas, 2006; Corduneanu & Bishop, 2001; Steele & Raftery, 2010).

### Estimation

After the design is selected, the model needs to be estimated: for each of the components, the optimal mean vector and covariance matrix has to be found. A well-known, efficient algorithm for this is the *Expectation Maximisation* (EM) algorithm (A. P. Dempster et al., 1977). First, the model is randomly initialised, typically by assigning data points to components randomly. Each component now has a means vector and a covariance matrix. The EM algorithm then iteratively performs two steps:

(1) **Expectation** step: calculate prior probabilities of belonging to each component, based on values of means and covariances.

(2) **Maximisation** step: update value of means and covariances based on prior probabilities and corresponding maximum likelihood.

However, the EM algorithm has shortcomings. First, while searching for a maximum likelihood, the algorithm only finds *local* maxima. This means that the found maximum likelihood is not necessarily the global maximum likelihood. The algorithm is sensitive to initialisation: changing the initial values can potentially change the model outcome. A trivial solution to this problem is to restart the model multiple times with different initial values, and select the model that has the highest maximum likelihood. This reduces the chance of a 'bad' model significantly.

Second, the algorithm can run into singularities: the situation in which a covariance matrix is singular after the update, causing the likelihood function to spike to infinity. Singularity of a covariance matrix implies that the variance is exactly zero. Typically, this happens when a component is fit to a single point, causing the Gaussian to collapse to that point. Luckily, this issue can also be resolved by restarting the model multiple times. Any model initialisation that results in a singularity is then discarded.

### Regularisation

While less common, regularisation exists for GMMs as well. Regularisation, in this case, implies penalising the estimated parameters of the components, i.e. the means vectors and/or the covariance matrices. Zhou, Pan, and Shen (2009) apply the LASSO penalty (Tibshirani, 1996) to shrink means towards zero. Variables with means reduced to zero are deemed non-informative towards distinguishing Gaussian components and are subsequently removed from the model. Thus, applying a penalty to component means performs variable selection.

Regularisation of the covariance matrices serves the same purpose as for the graphical model: reducing model complexity by increasing sparsity. Interestingly, Zhou et al. (2009) use the *glasso* algorithm for this estimation.

## Applications of GMMs

Gaussian Mixture Models are a well-known and popular clustering method for high-dimensional problems.

**Computer vision.** GMMs are often applied in imaging studies for their discriminatory power. M.-H. Yang and Ahuja (1998) use GMMs for the automated detection of skin color in images, finding that a mixture of Gaussian components outperforms a single Gaussian density function for this task. Similarly, Zivkovic (2004) use GMMs for background subtraction in images. Their approach allows for real-time background subtraction from surveillance images, thereby reducing the image to intruding objects only. Image segmentation, the detection of regions in an image, has been investigated by Farnoush and Zar (2008), and by Kim and Kang (2007).

# Methods

This section will provide exact definitions and descriptions of the methods used in this thesis. First, an exact, mathematical description is given for Gaussian Graphical Models (GGM) and for Gaussian Mixture Models (GMM). Second, the pipeline for the GGM/GMM method is presented. Third, methods for interpretation of GGM networks are reviewed.

## Regularised Gaussian Graphical Models

A Gaussian Graphical Model is defined as an undirected graph $G = (V, E)$ that describes the conditional independence structure of a set of random variables. Let $X = (X_1, \ldots, X_p)$ denote the set of random variables, with a $p$-dimensional multivariate distribution around $N_p(\mu, \Sigma)$, with mean $\mu$ and covariance matrix $\Sigma$. Each of the $p$ variables $X_1, \ldots, X_p$ then corresponds with the vertices in the set $V$. The estimation of the network structure is then equivalent to the estimation of non-zero edges in edge set $E$.

An edge $(i, j) \notin E$ if and only if $X_i$ is conditionally independent of $X_j$ given all other variables. Stated differently, edge $(i, j)$ is not part of edge set $E$ if and only if there is zero partial correlation between $X_i$ and $X_j$ ($\rho_{ij} = 0$), conditional on all other variables. Non-zero partial correlations can be inferred directly from the inverse of the covariance matrix $\Sigma$. Let $\Omega = \Sigma^{-1}$ denote the inverse covariance matrix, also known as the precision matrix. Let $\omega_{ij}$ denote entry $(i, j)$ of $\Omega$. The relationship between partial correlations and $\Omega$ is given by the following equation:

$$\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

Hence, non-zero entries of $\Omega$ correspond to edges in edge set $E$. Formally:

$$\omega_{ij} \neq 0 \iff (i, j) \in E$$

Now, let $x_1, \ldots, x_n$ denote a random sample from the multivariate distribution $N_p(\mu, \Sigma)$. Then, sample mean is given by

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

and the sample covariance matrix is given by

$$S = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Note that obtaining the precision matrix requires inverting the *true* covariance matrix, but only the *sample* covariance matrix is available. Hence, the precision matrix needs to be estimated from the sample. Maximum likelihood estimation can be used to estimate the precision matrix $\Omega$ by maximising the following log-likelihood function:

$$\ell(\Omega) = \log \det \Omega - \operatorname{tr}(\Omega S)$$

The result of this maximisation is the maximum likelihood estimate $\hat{\Omega}$, which is an approximation of the true $\Omega$.

### Regularisation

Estimation of partial correlation estimates using the above equations rarely leads to estimates of $\Omega$ that are sparse (i.e. contains only few non-zero entries). In high-dimensional contexts with small sample sizes (i.e. $n < p$), the sample covariance matrix $S$ is singular, and can therefore not be used to estimate $\Omega$. Even when $n \geq p$, especially for large $p$, $\hat{\Omega}$ is often a poor estimator due to the fact that sampling variation leads to estimates that are never *exactly* zero, causing spurious edges, i.e. false positives. In order to limit spurious edges and encourage sparsity, the estimation of $\Omega$ can be regularised using the 'least absolute shrinkage and selection operator' (LASSO; Tibshirani, 1996). The LASSO places a penalty on the estimation, resulting in the following penalised log-likelihood:

$$\ell_p(\Omega) = \log \det \Omega - \operatorname{tr}(\Omega S) - \lambda ||\Omega||_1 \tag{3.1}$$

where the $\ell_1$ norm is given by

$$||\Omega||_1 = \sum_{i,j,i \neq j} |\omega_{i,j}|$$

16

and $\lambda$ denotes a non-negative tuning parameter controlling sparsity. To obtain the maximum likelihood estimate, the penalised log-likelihood is maximised over positive-definite matrices. This is a convex optimisation problem that has seen several suggested approaches (e.g. the 'maxdet' algorithm (Yuan & Lin, 2007); semi-definite interior point optimisation (d'Aspremont, Banerjee, & El Ghaoui, 2008); permutation-invariant estimation using Cholesky decomposition (Rothman, Bickel, Levina, Zhu, et al., 2008)). The method used here is the graphical LASSO (glasso; Friedman et al., 2008), which is a quick and widely used method.

The glasso algorithm iteratively finds values for the matrix $\hat{\Omega}$ such that

$$\hat{\Omega} = \arg\max_{\Omega \geq 0} \left( \log \det \Omega - \operatorname{tr}(\Omega S) - \lambda ||\Omega||_1 \right) \tag{3.2}$$

The algorithm uses the fact that in order to find $\hat{\Omega}$, we can invert an estimate of $\Sigma$, defined as $\hat{\Sigma}$. This is achieved using the following algorithm (Friedman et al., 2008):

1. Initiate $\hat{\Sigma} = S + \lambda I$, where $S$ denotes the covariance matrix of the sample, and $I$ denotes the identity matrix of $S$.

2. Let $i$ and $j$ represent the rows and columns of $\hat{\Sigma}$, respectively. For each $i = 1, 2, \ldots, p$ and each $j = 1, 2, \ldots, p$, solve the lasso problem, defined as:

$$\hat{\beta} = \min_{\beta} \left\{ \frac{1}{2} ||\hat{\Sigma}_{ii}^{1/2} \beta - \left( \hat{\Sigma}_{ii}^{-1/2} S_{ij} \right) ||^2 + \lambda ||\beta||_1 \right\}$$

where $\hat{\beta}$ is a $p-1$ vector solution. Fill in row $i$ and column $j$ with the values of $\hat{\Sigma}_{ii}\hat{\beta}$, ignoring the diagonal.

3. Repeat step 2 until convergence.

The resulting estimate $\hat{\Sigma}$ is then inverted to obtain $\hat{\Omega}$.

## Model selection

Typically, the optimal value of the tuning parameter $\lambda$ is not set manually, but determined in a model selection procedure. A large number of models is estimated under varying values of $\lambda$ in the range $[\lambda_{min}, \lambda_{max}]$. For Gaussian Graphical models, $\lambda_{max}$ is typically set to the largest absolute correlation, and $\lambda_{min}$ is typically chosen by multiplying $\lambda_{max}$ with some ratio $R$ (e.g. 0.01 or 0.1), giving $\lambda_{min} = R\lambda_{max}$ (Epskamp, 2016b). To select the optimal model from these values, cross-validation or an information criterion can be used. The Bayesian Information Criterion (BIC) is widely used in this context. An extended version of the BIC (EBIC; Chen & Chen, 2008) introduces an additional complexity penalty to the BIC:

$$\text{EBIC} = -2\ell + E\log(N) + 4\gamma E\log(V) \tag{3.3}$$

where $\ell$ denotes the log-likelihood, $E$ denotes the edge set and $V$ denotes the set of vertices. The EBIC uses the hyperparameter $\gamma$ between 0 and 1. When $\gamma = 0$, the EBIC reduces to the BIC. The higher the value of $\gamma$, the higher the preference for simpler models (i.e. less edges). Minimizing the EBIC with respect to the value of $\lambda$ (and the resulting log-likelihood) then selects the optimal model. The hyperparameter $\gamma$, however, needs to be set manually.

# Clustering using regularised Gaussian Graphical Models

In order to identify structurally heterogeneous subtypes of the investigated phenomenon, it is possible to use the conditional independence structure of the Gaussian Graphical Model. Hill and Mukherjee (2013) provide a method of estimating $K$ different clusters using Gaussian Mixture Modelling with glasso estimation of precision matrices. This method combines discovery and estimation of subtypes in the data, resulting in clusters with unique Gaussian Graphical Models. The method is described below.

# Gaussian Mixture Model

In order to cluster the data, Hill and Mukherjee (2013) extend the above estimation to a multivariate Gaussian Mixture Model, where cluster assignments are considered the latent variable. Every Gaussian mixture, in essence, represents a cluster and a distinct Gaussian Graphical Model. The Gaussian mixture distribution is defined by

$$f(x_i; \Theta) = \sum_{k=1}^{K} \pi_k N_{p,k}(x_i | \mu_k, \Sigma_k)$$

with mixing proportions $\pi_k$ satisfying $0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$, and where $N_{p,k}$ is the multivariate Gaussian density with mixture-specific mean $\mu_k$ and covariance matrix $\Sigma_k$. Important to note is that the covariance matrices are of full rank. Since the aim is to estimate the off-diagonal elements of the precision matrix by inverting the covariance matrix, using a spherical covariance matrix will result in a singularity. Using a diagonal covariance matrix will result in components that are more similar to each other, decreasing the discriminatory power of the model. $\Theta$ represents the set of all unknown parameters for all mixtures $\{(\pi_k, \mu_k, \Sigma_k) : k = 1, \dots, K\}$.

# Regularisation

The penalised log-likelihood function (3.1) can be extended to a Gaussian mixture form:

$$\ell_p(\Omega) = \sum_{i-1}^{n} \log \left( \sum_{k=1}^{K} \pi_k N_{p,k}(x_i | \mu_k, \Sigma_k) \right) - \frac{n}{2} p_\lambda(\Theta) \tag{3.4}$$

with penalty term

$$p_\lambda(\Theta) = \lambda \sum_{k=1}^{K} \pi_k ||\Omega^k||_1$$

Note that the penalty term for each cluster is weighted by its mixing proportion. Previous work has shown that including the mixing proportion in the penalty term yields better results in general (Hill & Mukherjee, 2013; Zhou et al., 2009). As such, this penalty term is used.

In contrast to (3.1), equation (3.4) is not convex, and thus its maximisation is a non-trivial task. Hill and Mukherjee (2013) use the Expectation-Maximisation (EM) algorithm (A. P. Dempster et al., 1977) to obtain maximum likelihood estimates. First follows a brief explanation of the EM algorithm. Then, the EM algorithm is defined for the maximisation of (3.4).

The EM algorithm uses a set $X$ of observed data, a set $Z$ of latent variables, the vector of unknown parameters $\Theta$, and the likelihood function $L(\Theta; X, Z) = p(X, Z | \Theta)$. Directly maximising the marginal likelihood of the observed data is normally intractable. EM, then, iteratively applies two steps to find an optimal solution:

*Expectation step*: define expected value of the log likelihood function of $\Theta$:

$$Q(\Theta | \Theta^{(t)}) = \mathbb{E}_{Z|X, \Theta^{(t)}} [\log L(\Theta; X, Z)]$$

*Maximisation step*: find parameters that maximise $Q(\Theta | \Theta^{(t)})$:

$$\Theta^{(t+1)} = \arg \max Q(\Theta | \Theta^{(t)})$$

After each maximisation step, the newfound parameters $\Theta^{(t+1)}$ are used in the next expectation step. This process is repeated until convergence at a (local) maximum.

For the estimation of the penalised log-likelihood function of a Gaussian Mixture Model, the problem is defined as follows.

Let $z_i$ be a latent variable satisfying $z_i = k$ if observation $x_i$ belongs to cluster $k$. Then it follows that $P(z_i = k) = \pi_k$ and $p(x_i|z_i = k) = N_{p,k}(x_i|\mu_k, \Sigma_k)$. The complete data can then be denoted as $\{x_i, z_i\}_{i=1}^n$. The penalised log-likelihood for the complete data is then:

$$\ell_{p,c}(\Omega) = \sum_{i-1}^n \log(\pi_{z_i}) + \log(N_{p,z_i}(x_i|\mu_{z_i}, \Sigma_{z_i})) - \frac{n}{2}p_\lambda(\Theta) \tag{3.5}$$

The expectation step then uses the current parameter estimates $\Theta^{(t)}$ to compute

$$\begin{aligned}
Q(\Theta|\Theta^{(t)}) &= \mathbb{E}\left[\ell_{p,c}(\Theta)|\{x_i\}_{i=1}^n, \Theta^{(t)}\right] \\
&= \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(t)}\left[\log(\pi_k) + \log(f_k(x_i|\mu_k, \Sigma_k))\right] - \frac{n}{2}p_\lambda(\Theta)
\end{aligned} \tag{3.6}$$

where $\tau_{ik}^{(t)}$ is the posterior probability of observation $x_i$ belonging to cluster $k$, and can be seen as a 'soft' cluster assignment:

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t)} N_{p,k}(x_i^{(t)}|\mu_k^{(t)}, \sum_k^{(t)})}{\sum_{j=1}^K \pi_j^{(t)} N_{p,j}(x_i^{(t)}|\mu_j^{(t)}, \sum_j^{(t)})} \tag{3.7}$$

In the maximisation step, $Q\left(\Theta|\Theta^{(t)}\right)$ is maximised to provide estimates of the parameters $\Theta^{(t+1)} = \{\pi_k^{(t+1)}, \mu_k^{(t+1)}, \Sigma_k^{(t+1)}\}$. The standard updates for each of these parameters are as follows:

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)}}{n} \tag{3.8}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)} x_i}{\sum_{i=1}^n \tau_{ik}^{(t)}} \tag{3.9}$$

The update for $\Sigma_k^{(t+1)}$ can be rewritten as the inverse of the update for $\Omega_k$, which is:

$$\begin{aligned}
\Omega_k^{(t+1)} &= \arg\max\left[\sum_{i=1}^n \tau_{ik}^{(t)}\left(\log\det\Omega_k - \operatorname{tr}(\Omega_k S_k^{(t)})\right) - n\lambda\left(\pi_k^{t+1}\right)||\Omega^k||_1\right] \\
&= \arg\max\left[\log\det\Omega_k - \operatorname{tr}(\Omega_k S_k^{(t)}) - \lambda||\Omega^k||_1\right]
\end{aligned} \tag{3.10}$$

where the standard update of $S_k^{(t+1)}$ is:

$$S_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)}\left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^n \tau_{ik}^{(t)}} \tag{3.11}$$

The expectation and maximisation steps are repeated until either (1) a maximum number of iterations is reached (default is 100), (2) a minimum cluster size is reached (default is 4), or (3) the relative change of the penalised log-likelihood is below a certain threshold (default is $10^{-4}$).

## Model selection

Finally, model selection is addressed by comparing the cross-validation and BIC methods. The EBIC has been shown to outperform the BIC in the discovery of sparse estimates (Chen & Chen, 2008). Normally, the EBIC applies to the selection of sparse estimates a single Gaussian. However, the GMM consists of $k$ Gaussians, each needing sparse estimates. Therefore, we extend the literature by adapting the EBIC to the mixture model, in order to make it usable for model selection in the GMM.

The EBIC as described in equation (3.3) is tailored to a single Gaussian Graphical Model. The EBIC can be extended to the Gaussian Mixture Model by considering that the edge set $E$ equals the degrees of freedom

of the model. The degrees of freedom for the penalized Gaussian Mixture Model is given by the following equation (Hill & Mukherjee, 2013; Yuan & Lin, 2007):

$$\text{df}_\lambda = K(P+1) - 1 + \sum_{k=1}^{K} E_k$$

where $p$ is the number of variables and $E_k$ is the edge set of cluster $k$ under $\lambda$, equivalent to the number of non-zero entries in the upper triangle of $\hat{\Omega}_k$, formally given by

$$E_k = \# \left\{ (i,j) : i \leq j, (\hat{\omega}_{k\lambda})_{ij} \neq 0 \right\}$$

The EBIC for the penalised Gaussian Mixture Model is then:

$$\text{EBIC} = -2\ell(\hat{\Theta}_\lambda) + \text{df}_\lambda \log(n) + 4 \, \gamma \, \text{df}_\lambda \, \log(p) \tag{3.12}$$

The resulting Gaussian Mixture Model contains, in each mixture $k$, the penalised precision matrix $\hat{\Omega}_k$ that can be used to draw the network of the corresponding Gaussian Graphical Model.

# Pipeline

Integrating the above methods results in the following pipeline:

1. Generate 100 log-scaled values of the penalisation parameter $\lambda$ between $\lambda_{min}$ and $\lambda_{max}$.

2. For each value of $\lambda$, do the following:

   For $R$ restarts, do the following:

   i. Initialise $K$ components by randomly assigning data points to each component. Compute means $\mu_k^{(0)}$ and variance-covariance matrices $S_k^{(0)}$ for each component. Use $S_k^{(0)}$ as input for the `glasso` algorithm to obtain $\Omega_k^{(0)}$ and invert the result to obtain $\Sigma_k^{(0)}$.

   ii. Repeat the following EM steps until convergence:

      A. **Expectation step:** for each data point, compute posterior probabilities $\tau_{ik}^{(t)}$ using $\mu_k^{(t)}$ and $\Sigma_k^{(t)}$, as in Equation (3.7).

      B. **Maximisation step:** for each component, compute $\mu_k^{(t+1)}$ and $S_k^{(t+1)}$ using (3.9) and (3.11), respectively. Use $S_k^{(t+1)}$ as input for the `glasso` algorithm to obtain $\Omega_k^{(t+1)}$ (Eq. (3.10)) and invert the result to obtain $\Sigma_k^{(t+1)}$.

      C. Increment $t$.

   iii. Compute the EBIC score for this model using equation (3.12).

3. Return the model with the lowest EBIC score.

4. For each group in the best model, draw networks and compute network features.

# Intepretation and comparison of regularised Gaussian Graphical Models

Once a Gaussian Graphical Model has been computed and drawn as a network, it can be interpreted using several network features. If multiple GGMs are computed (e.g. for varying conditions, groups, or clusters), these features can be used to compare the networks.

**Visualisation.** A GGM can be interpreted through *visualisation of the GGM*. The graph visually represents the conditional independence structure of the variables, which can aid in the generation of new hypotheses about variable relations and interaction. Additional methods can be applied to create an informative graph. For example, the package `qgraph` (Epskamp et al., 2018) in the statistical language R applies several. First, by default, it determines the layout of the graph using the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), creating a force-directed layout, causing strongly connected variables to be closer together and vice versa. Second, it allows for a threshold value, where any edges with weights lower than this value will not be drawn. Third, it allows for the grouping of variables, if any variable structure is known or expected a priori. By using (a combination of) these settings, a visual inspection of the graph can provide valuable information about the studied construct.

**Centrality.** Second, the *centrality* of the nodes in the graph can be investigated. Centrality measures are graph-theoretic measures used to identify the most important or impactful nodes in the network. The four relevant centrality measures are *strength centrality*, *betweenness centrality*, *closeness centrality* and *expected influence.*.

*Strength* centrality simply sums the absolute edge weights of the edges tied to a certain node $v$:

$$C_S(v) = \sum^{E} |(e_{ij} \in E : i = v \lor j = v)|$$

*Betweenness* centrality counts how often node $v$ is part of a shortest path between two random nodes. It signals nodes that function as bridges between other nodes in the network:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{st}$ is the total number of shortest paths from node $s$ to node $t$ and $\sigma_{st}(v)$ is the number of these shortest paths that go through $v$.

*Closeness* centrality quantifies how close a node is, on average, to all other nodes. It is the average length of the shortest path between the node itself and all other nodes:

$$C_C(v) = \frac{1}{\sum_s d(s, v)}$$

where $d(s, v)$ denotes the distance between node $s$ and node $v$.

*Expected influence* is a relatively new centrality measure proposed by Robinaugh, Millner, and McNally (2016). It is similar to strength, but takes into account positive and negative values of edge weights by *not* taking the absolute value:

$$C_{EI}(v) = \sum^{E} (e_{ij} \in E : i = v \lor j = v)$$

**Stability.** Third, the *stability* of the estimated edges can be interpreted as a measure of the rigidity of relations between variables. Edge weight stability can be estimated by using bootstrapping techniques. The R package `bootnet` does exactly this (Epskamp et al., 2018). By resampling with replacement and re-estimating the edge weights for each resampling, an edge weight distribution for each edge is calculated, with corresponding mean, standard deviance, and confidence interval. Edges that are sensitive to this resampling will have wide distributions and can be considered unstable, and vice versa.

# Results

## Benchmarking

To test whether the method is able to accurately identify conditional independence structures of subgroups, the method is tested on artificial data. Artificial data is generated as follows. First, for each subgroup, a $p \times p$ precision matrix is generated, containing the wanted conditional independence structure. Second, these precision matrices are converted to a covariance matrix. Third, multivariate data is randomly sampled from these covariance matrices, resulting in an artificial subsample for each subgroup. Finally, the artificial subsamples are combined into one sample.

Such artificial datasets are generated for varying values of three conditions: degree of structural overlap, balanced sample sizes and unbalanced sample sizes. The exact list of conditions is presented in Table 4.1.

Performance of the model is assessed for every condition. For each condition, an optimal model is computed with $k = 3$ and $\gamma = 0.5$. Edges estimated to be lower than 0.1 are excluded from the resulting networks. Precision of the estimated structures is assessed by comparing the estimated structures of the subgroups with the true structure of the subgroups. This comparison includes estimation accuracy (i.e. whether the estimated network contains no spurious edges and all true edges) and edge weight accuracy (i.e. how closely the true edge weights are approximated by the estimated edge weights).

This section is divided as follows. First, a baseline condition is presented as an example. For this baseline condition, an extensive evaluation of the results is performed to provide an overview of how benchmarking tests are assessed. Second, summarised results are presented for varying values of the three conditions. Results of centrality indices and centrality stability are excluded from these summaries, as centrality values are dependent on the structure of the estimated networks. However, full results of network structure, network stability, centrality indices and centrality stability are available upon request.

| Conditions | | |
|---|---|---|
| Degree of structural overlap | Varying balanced group sizes | Varying unbalanced group sizes |
| 0/4 overlap | $N = 500$ | $N_k = \{1000, 1000, 500\}$ |
| 2/5 overlap | $N = 400$ | $N_k = \{1000, 500, 100\}$ |
| 4/6 overlap | $N = 300$ | $N_k = \{500, 500, 100\}$ |
| 6/7 overlap | $N = 200$ | |
| | $N = 100$ | |

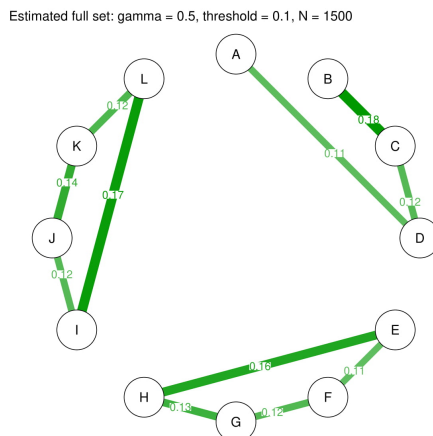Table 4.1: Overview of conditions tested in the benchmarking tests.



Figure 4.1: Estimated CIS of the full artificial dataset generated using the CISs shown in Figure 4.2. Nodes represent variables, edge weights represent partial correlations between variables.

# Baseline condition

In this section, we create an artificial dataset that serves as a baseline case that is used for testing the method. We use this baseline case as an example of the application of the GGM/GMM method. We compare network structure, network stability, centrality scores and centrality stability of the estimated networks with those of the true networks.

An artificial dataset was generated with $p = 12$ variables, $k = 3$ subgroups, each with a subsample size of $N = 500$. The conditional independence structures used to generate this dataset are presented in Figure 4.2. In this condition, each subgroup has a variable structure of four interdependent variables, connected cyclically by edges with random weights in the range $[0.3, 0.4]$. The structures of the subgroups do not have any variables in common. This is the simplest case, where the subgroups have no overlap in terms of conditional independence structure.

The resulting dataset has sample size $N = 1500$. Estimating the conditional independence structure of this aggregate dataset yields the graph seen in Figure 4.1. Visually inspecting this graph, it resembles the structures of the three subgroups, combined into one graph. However, the edge $A - B$ (belonging to group 1) was estimated to be lower than 0.1, and has been excluded from the graph.
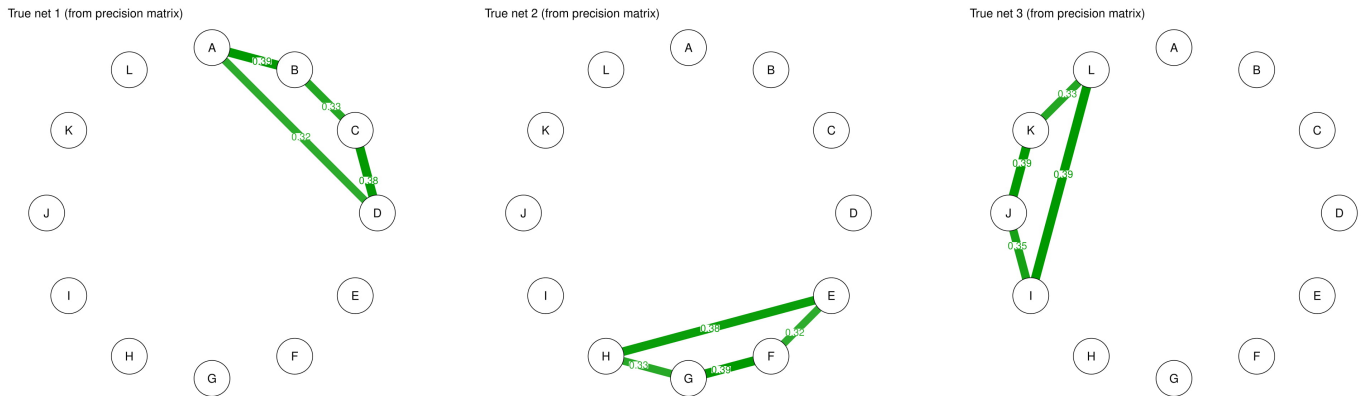


Figure 4.2: True CISs used to generate the artificial sample. Nodes represent variables, edge weights represent partial correlations between variables.
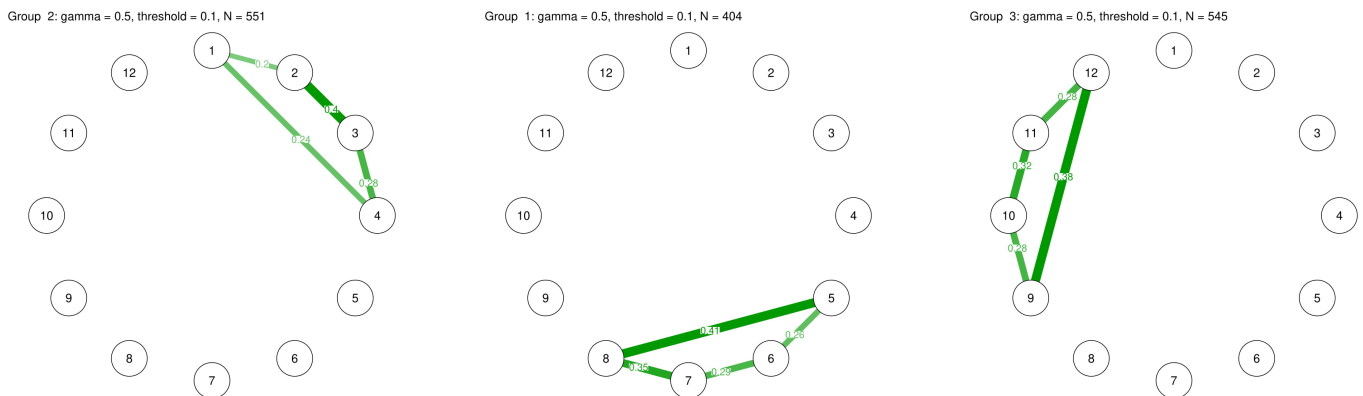


Figure 4.3: CISs for $k = 3$ subgroups estimated using the proposed GMM/GGM method. Nodes represent variables, edge weights represent partial correlations between variables.

**Structural estimation.** From the artificial sample, a family of models was computed with $k = 3$, distinguishing three subgroups in the sample. The estimated conditional independence structures are shown in Figure 4.3. Comparing the networks in Figure 4.3 with those in Figure 4.2 reveals that the method is able to accurately estimate the structures of the subgroups. Note that the edge weights of the estimated networks are lower than those in the true networks, due to regularisation.

**Centrality plots.** Figure 4.4 shows the standardised centrality plots of the true networks and the estimated networks. A visual comparison reveals that the strength and expected influence centralities are almost identical between the true and the estimated networks, save for differences as a result of smaller edge weights in the estimated networks. Closeness and betweenness centralities show slight differences between the true and estimated networks. For example, in the betweenness plot for group 3, variable J/V10 has a high positive value (2.0) in the true network, but a slightly negative value ($-0.5$) in the estimated network. Such discrepancies may be the result of the fact that betweenness and closeness are less stable in very sparse networks (Epskamp, 2016a). As a consequence, these measures should be interpreted with care.

**Edge weight stability analysis.** Figure 4.5 shows the edge weight stability analyses for the true and estimated networks. Visual comparison reveals that the estimated networks are more likely to have spurious edges, indicated by the presence of more grey area, which represents confidence intervals for an edge weight value. In addition, estimated edges are sometimes less stable (e.g. group 3, edge V9-V12), indicated by the larger confidence interval around the edge.

**Centrality stability analysis.** Figure 4.6 shows the plots of the stability of the centrality indices for the true and estimated networks. Visual inspection of the plots shows that overall, the centrality scores seem less stable for the estimated networks, compared to the true networks. This is especially true for the betweenness centrality. Nonetheless, the centralities of the true networks also seem unstable, as indicated by the large area around the lines indicating the confidence intervals.
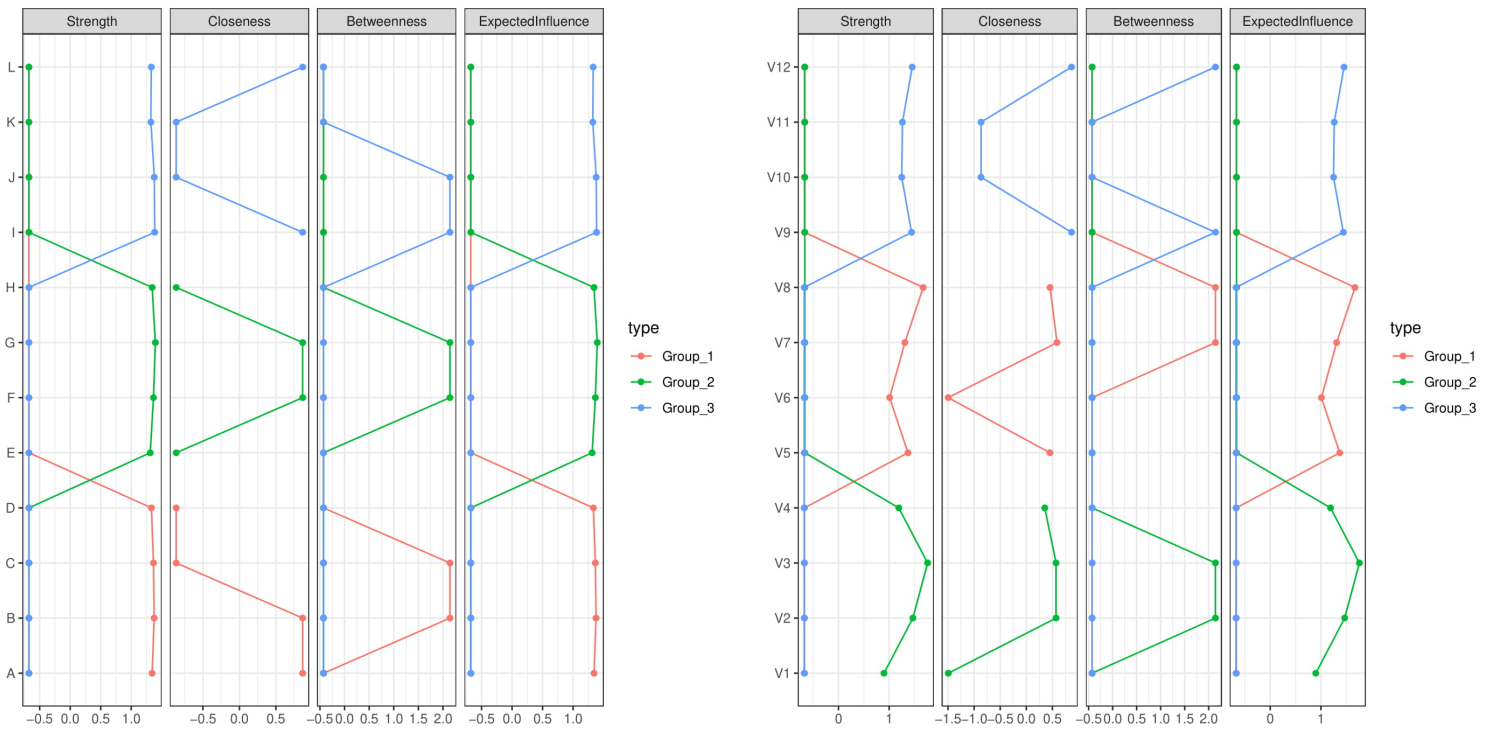


Figure 4.4: Standardised centrality plots for graphs of the true networks (left) and estimated networks (right) of the artificial data with k = 3, p = 12, equal N = 500, and no structural overlap. The x-axis shows standardised centrality scores. The y-axis shows the nodes. Groups are indicated by colours.
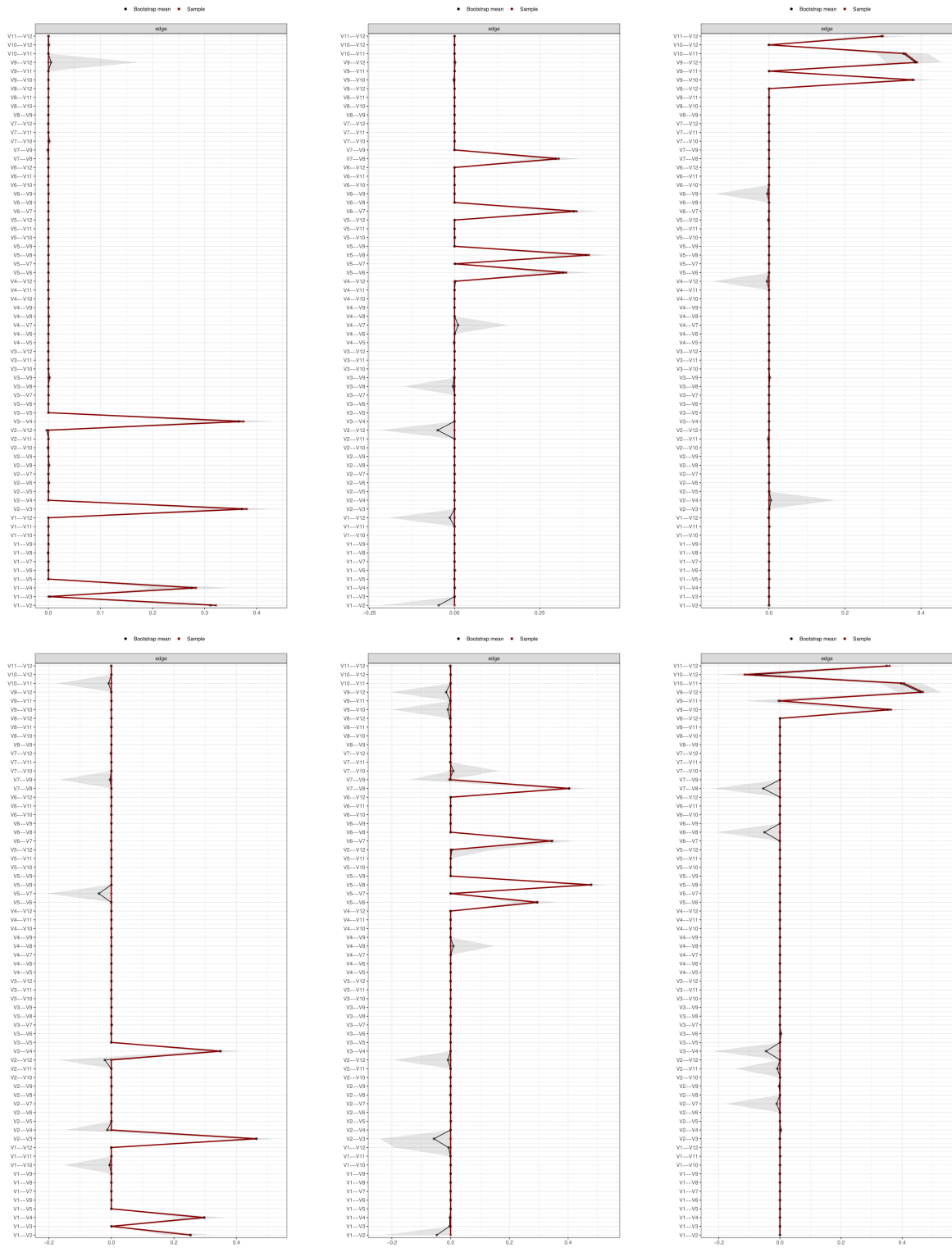
24

Figure 4.5: Stability analyses of true networks (top row) and estimated networks (bottom row) of the artificial data with k = 3, p = 12, equal N = 500, and no structural overlap. Red line indicates estimated network weights, black line indicates bootstrap means, grey areas indicate 95% confidence interval for bootstrap means. From left to right: group 1, group 2, group 3.
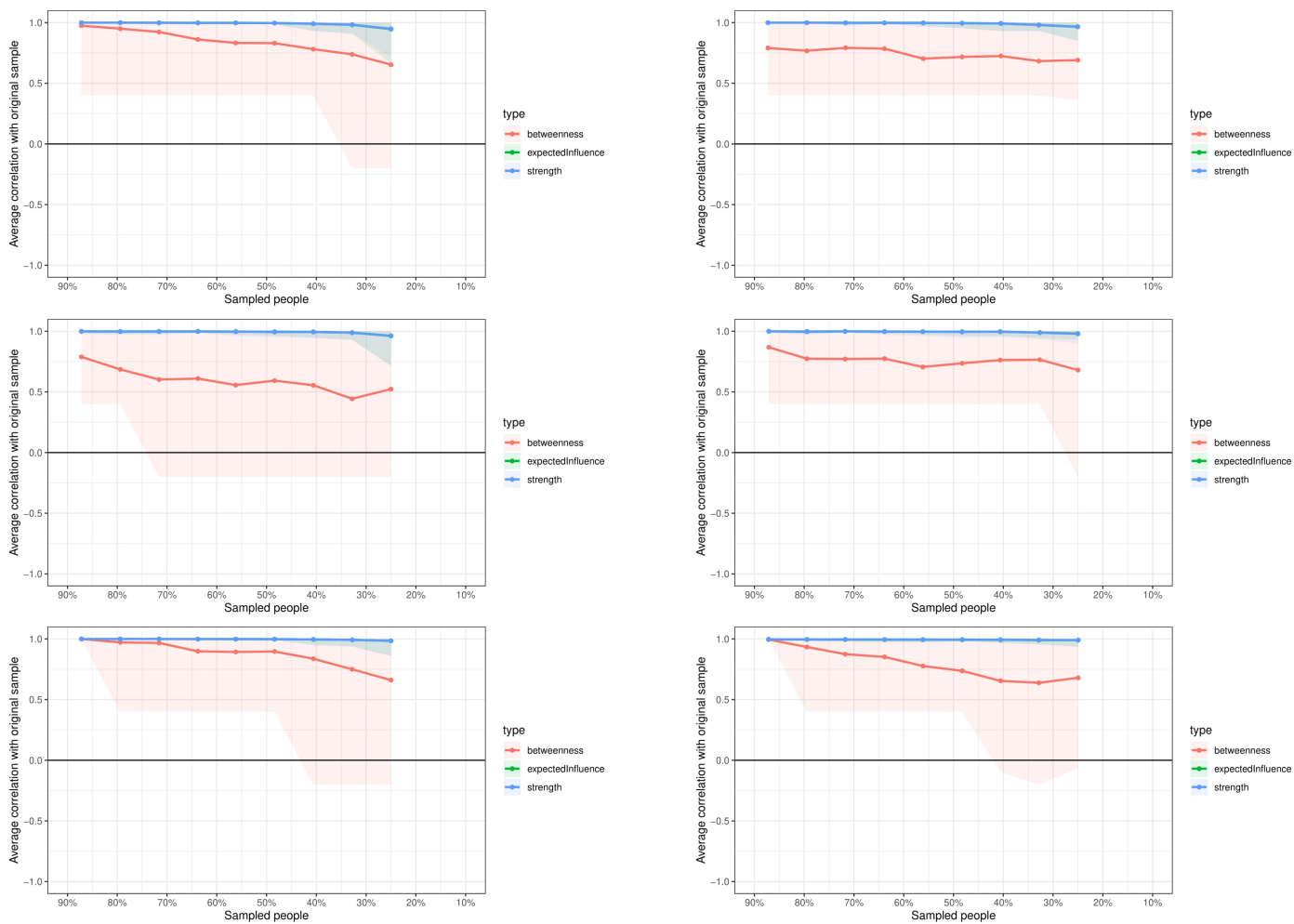
Figure 4.6: Centrality stability plots for graphs of the true networks (left column) and estimated networks (right column) of the artificial data with k = 3, p = 12, equal N = 500, and no structural overlap. x-axis indicates % of retained cases, y-axis indicates average correlation with the 100% sample, colours indicate centrality metrics, areas indicate 95% confidence intervals around average correlations. From top to bottom: group 1, group 2, group 3.

# Variations of baseline condition: degree of structural overlap

To estimate model performance in cases where subgroups have variables in common, the method was tested with subgroups that have overlapping structures. The structure of each subgroup in the baseline condition was extended by adding one or more variables to the structure clockwise. This was done for one, two and three variables, respectively. This results in subgroup structures that have one, two or three variables in common with each other subgroup structure, resulting in four conditions in total: 0/4 overlap (i.e. the baseline condition), 2/5 overlap, 4/6 overlap and 6/7 overlap. Each subgroup retains a sample size of $N = 500$. Figure 4.7 provides the true structures of all groups for each condition.

For each of the conditions, a family of models with $k = 3$ was computed, and the model with the lowest EBIC score was selected as the optimal model. In addition, a single network for the full sample is computed. The resulting estimated networks for each condition are shown in Figure 4.8. Results are summarised in Figure 4.9.

Results show that all but one group network are correctly estimated by the method. Only the network for group 2 in the 6/7 overlap condition produced one spurious edge. In addition, the networks for the full samples are often unable to capture the full structures of all groups in one graph. For these networks, the strongest edges are those that are shared by two groups, i.e. overlapping edges. Edge weights are consistently estimated to be lower than the true edge weights, ranging from $\pm 70\%$ to $\pm 85\%$ of the true value.
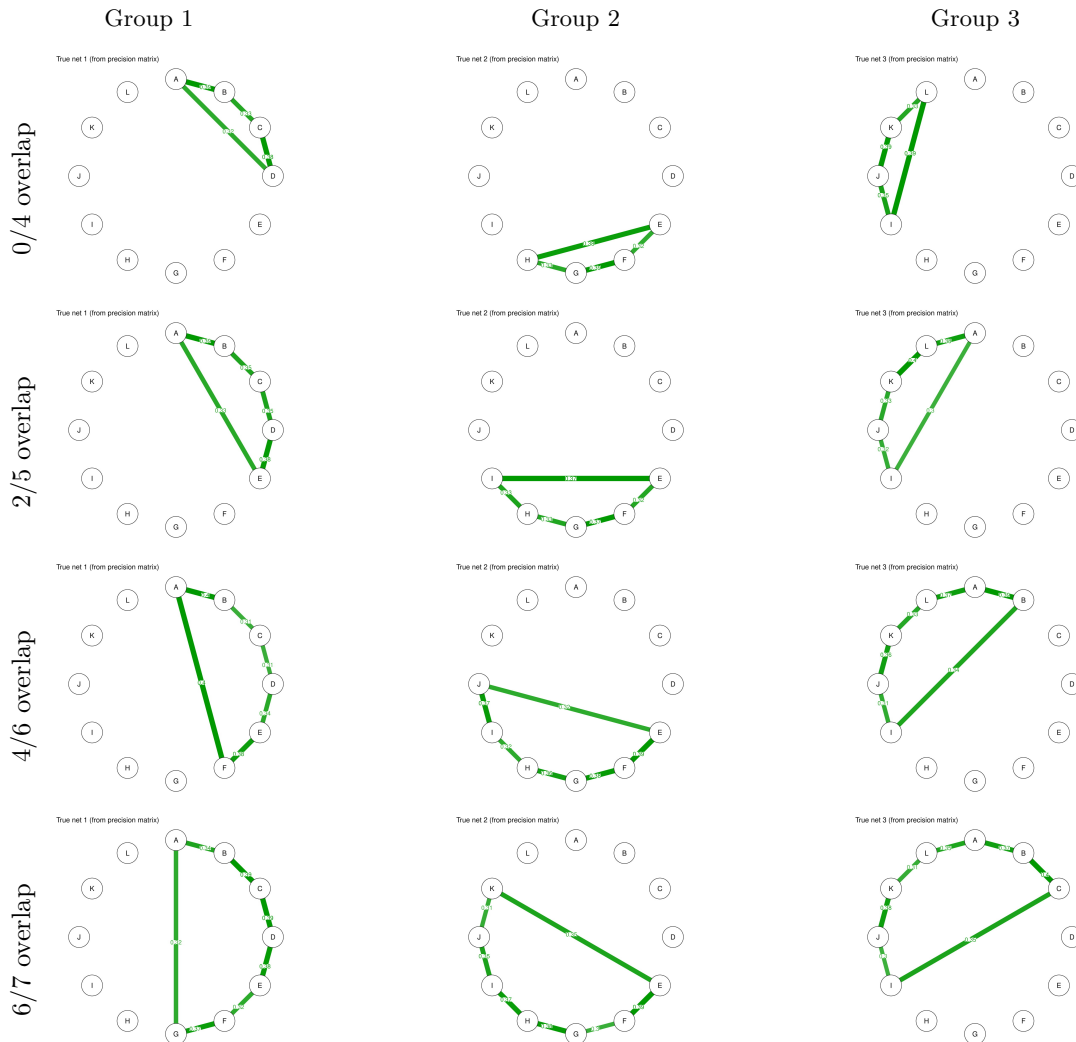


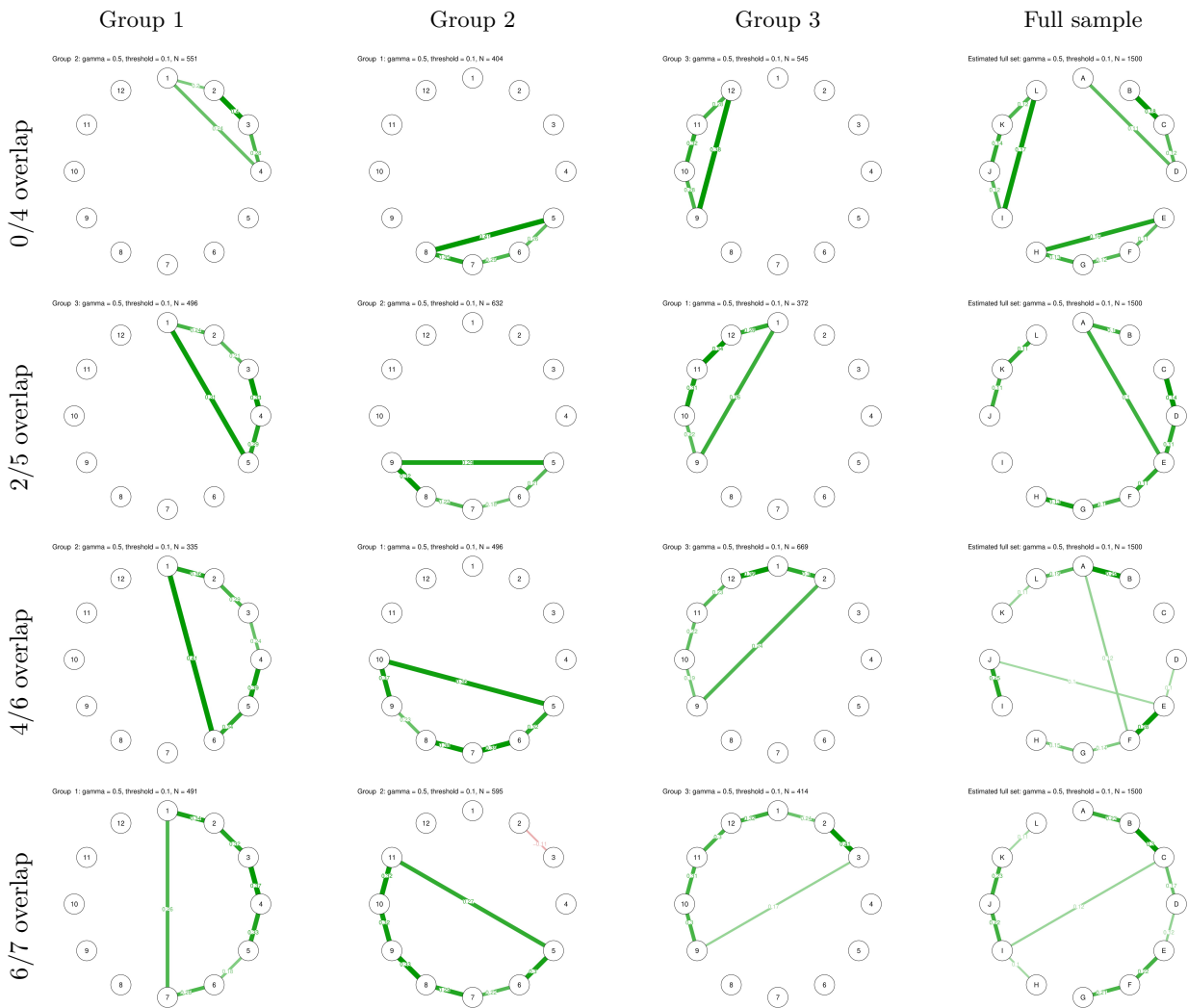Figure 4.7: True structures for subgroups of every structural overlap condition.

Figure 4.8: Estimated structures for subgroups and full sample of every structural overlap condition.
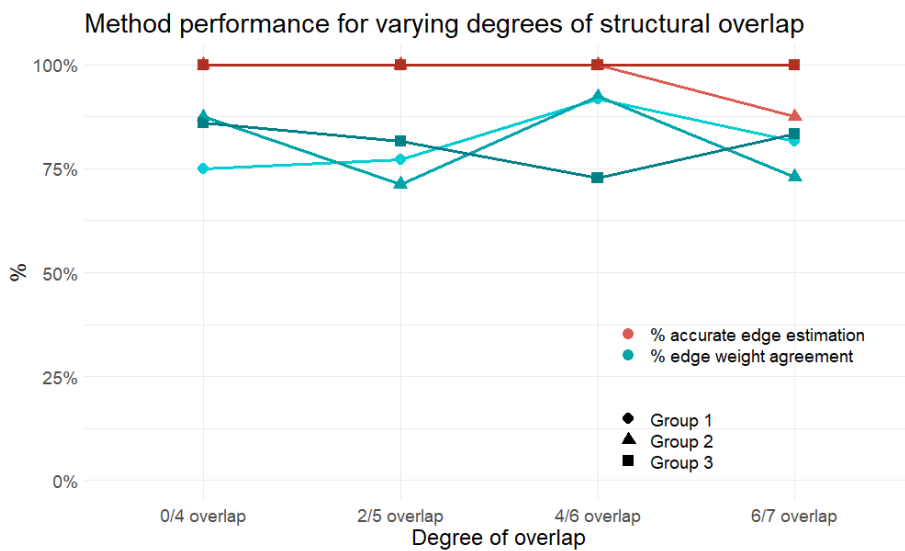


Figure 4.9: Summary of estimation performance for varying degrees of structural overlap. Red lines indicate % of edges correctly estimated, blue lines indicate % to which estimated edge weights approach true edge weights. x-axis represents conditions, y-axis represents percentages.
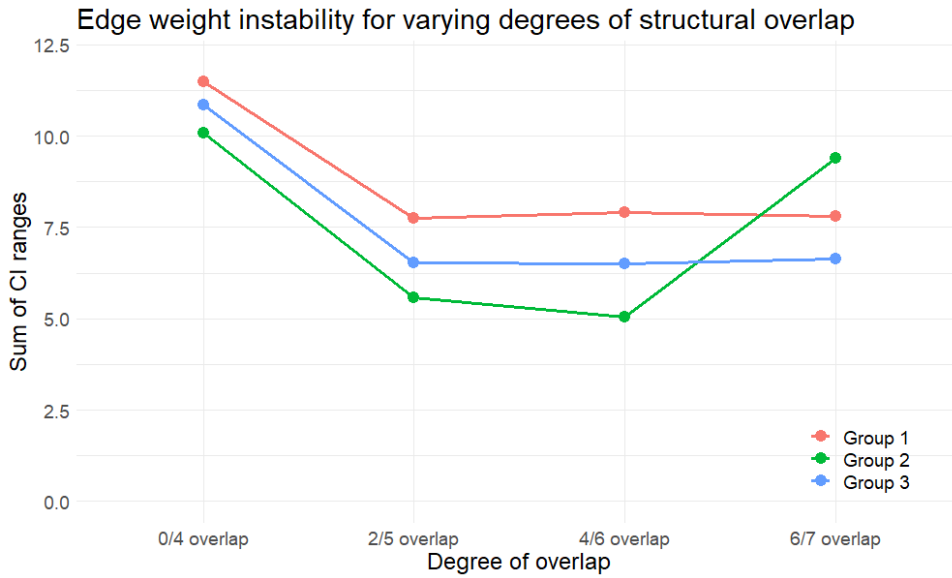
Figure 4.10: Summary of edge weight instability for varying degrees of structural overlap. Network edge weight stability was calculated as the sum of confidence interval ranges for all edges of the network. Colours indicate groups. x-axis represents conditions, y-axis represents sum of CI ranges.

Figure 4.10 shows an overview of edge weight instability for all conditions. Two meaningful observations can be made. First, edge weight instability is higher for the 0/4 overlap condition (i.e. the baseline condition) than for the other conditions. Second, there is a spike in instability of the network for group 2 in the 6/7 overlap condition. This corresponds with the spurious edge that was estimated for this network.

## Variations of baseline condition: balanced sample sizes

The method was also tested for decreasing balanced sample sizes. In addition, the conditional independence structure chosen for testing balanced sample sizes was that of the 4/6 overlap condition, as some structural overlap is expected for most real data. Furthermore, for $N = 500$, this structure yields stable results with no estimation errors. Starting from $N = 500$, the sample size was decreased with 100, 200, 300 and 400, respectively, resulting in five conditions: $N = 500$, $N = 400$, $N = 300$, $N = 200$, $N = 100$. For each of the conditions, a family of models with $k = 3$ was computed, and the model with the lowest EBIC score
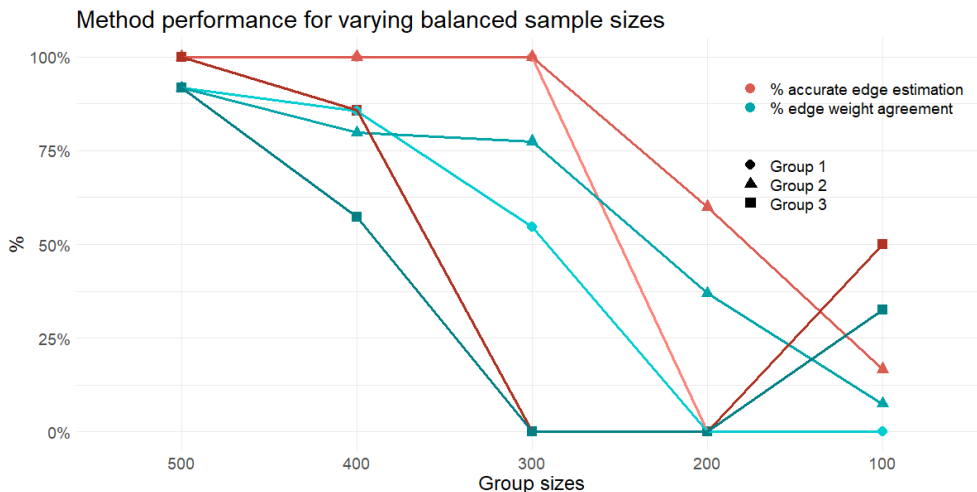


Figure 4.11: Summary of estimation performance for varying balanced sample sizes. Red lines indicate % of edges correctly estimated, blue lines indicate % to which estimated edge weights approach true edge weights. x-axis represents conditions, y-axis represents percentages.
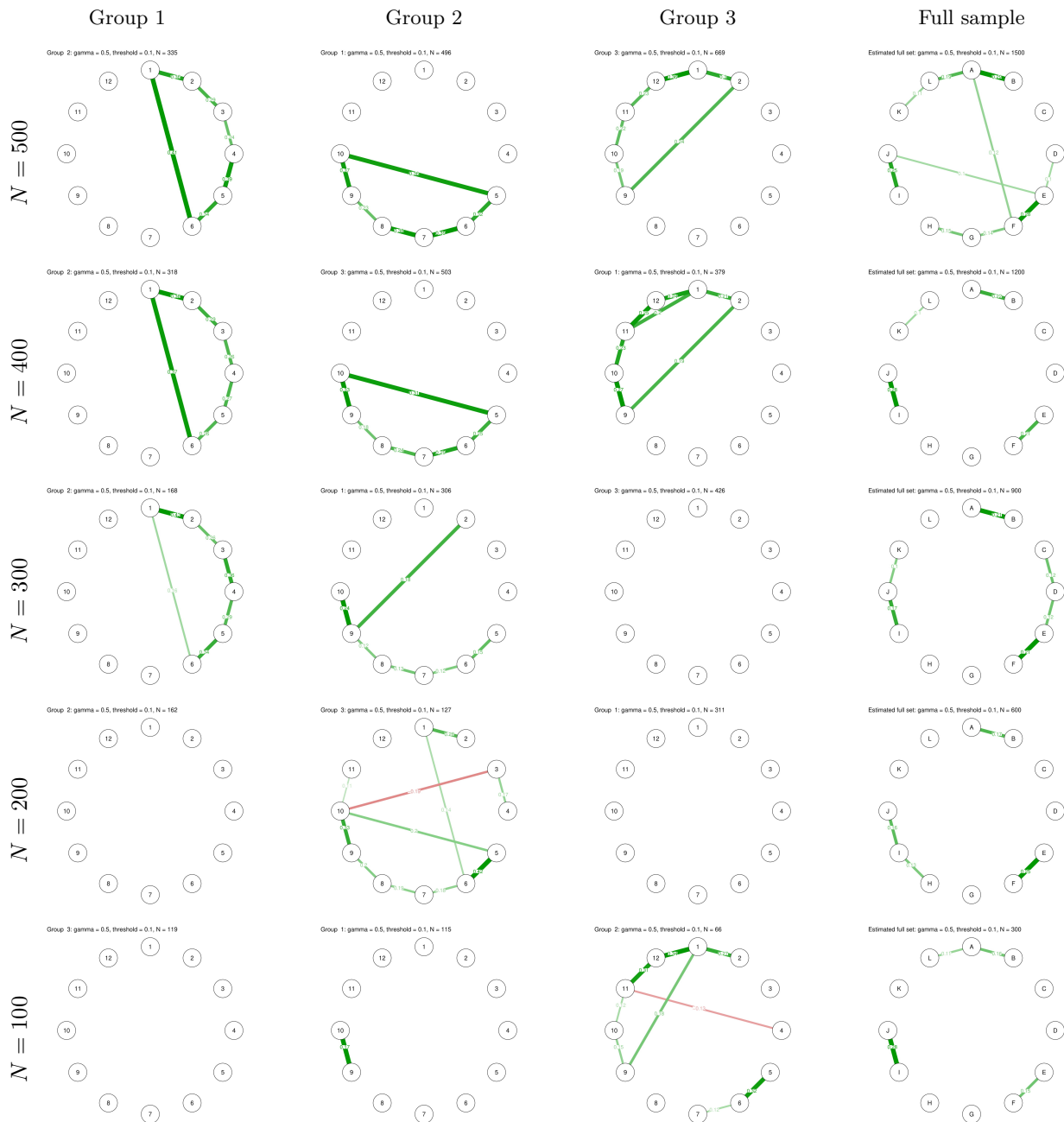
Figure 4.12: Estimated structures for subgroups and full sample of every balanced sample size condition.

was selected as the optimal model. In addition, a single network for the full sample is computed for each condition. The resulting estimated networks are shown in Figure 4.12. Results are summarised in Figure 4.11.

Results show that all conditions with sample sizes lower than 500 have at least one estimation error in one of the groups. For conditions with a sample size of 300 or less, one or more networks are estimated empty (i.e. without any edges). With decreasing sample sizes, accuracy of edges and edge weights strongly decreases. Inspecting networks with erroneous edges, results show that in some cases spurious edges in one group are part of the true structure of another group (e.g. $N = 300$, group 2; $N = 200$, group 2). However, for the two conditions with the lowest sample sizes, erroneous edges are estimated that are not part of any group's true structure. The full sample networks also become more sparse with decreasing sample size. However, for every group size, the full sample network shows three cliques of connected variables that correspond to the three groups.
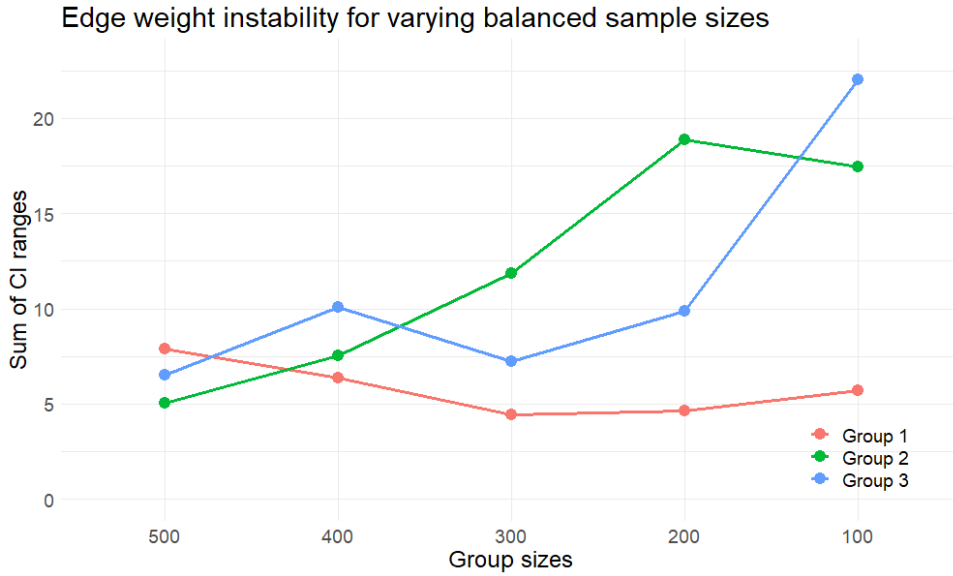
Figure 4.13: Summary of edge weight instability for decreasing balanced sample sizes. Network edge weight stability was calculated as the sum of confidence interval ranges for all edges of the network.

Edge weights are still estimated lower than the true edge weights. With decreasing sample size, the edge weights seem to deviate more from the true value. In general, edge weight agreements follow the same pattern as edge estimation accuracy. In other words, if an edge is accurately estimated, the estimated edge weight is likely within $\pm 25\%$ of the true edge weight.

Figure 4.13 shows an overview of edge weight instability for all conditions. Generally, instability seems to increase with decreasing sample size. However, instability for group 1 seems to stay at the same level for all conditions, despite the fact that the group 1 network is empty for $N = 200$ and $N = 100$. The empty networks for group 3 ($N = 300$ and $N = 200$) also show low instability.

## Variations of baseline condition: balanced sample sizes

Finally, the method was tested for different combinations of unbalanced sample sizes. Again, the conditional independence structure chosen for testing was that of the 4/6 overlap condition. Group sizes were set to be $N = 1000$ (large group), $N = 500$ (medium group) or $N = 100$ (small group). Using these sizes, three cases were investigated. First, the case where one group is significantly smaller, but all groups have large
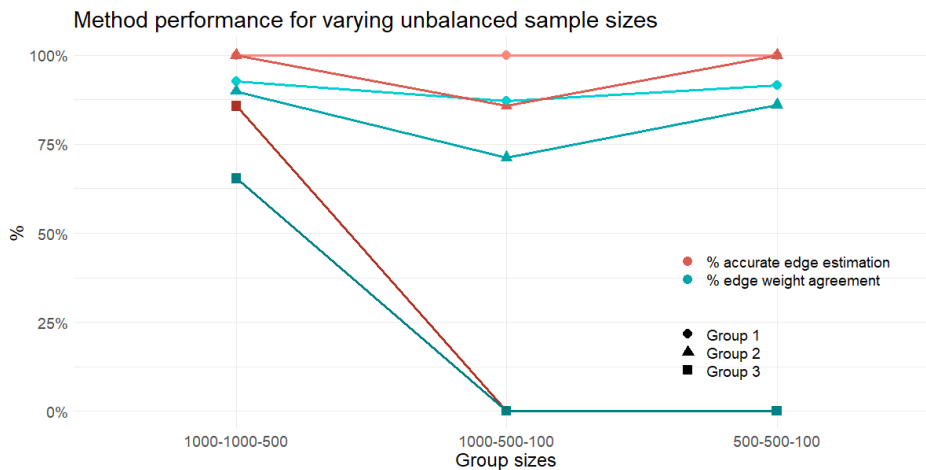


Figure 4.14: Summary of estimation performance for varying unbalanced sample sizes. Red lines indicate % of edges correctly estimated, blue lines indicate % to which estimated edge weights approach true edge weights. x-axis represents conditions, y-axis represents percentages.
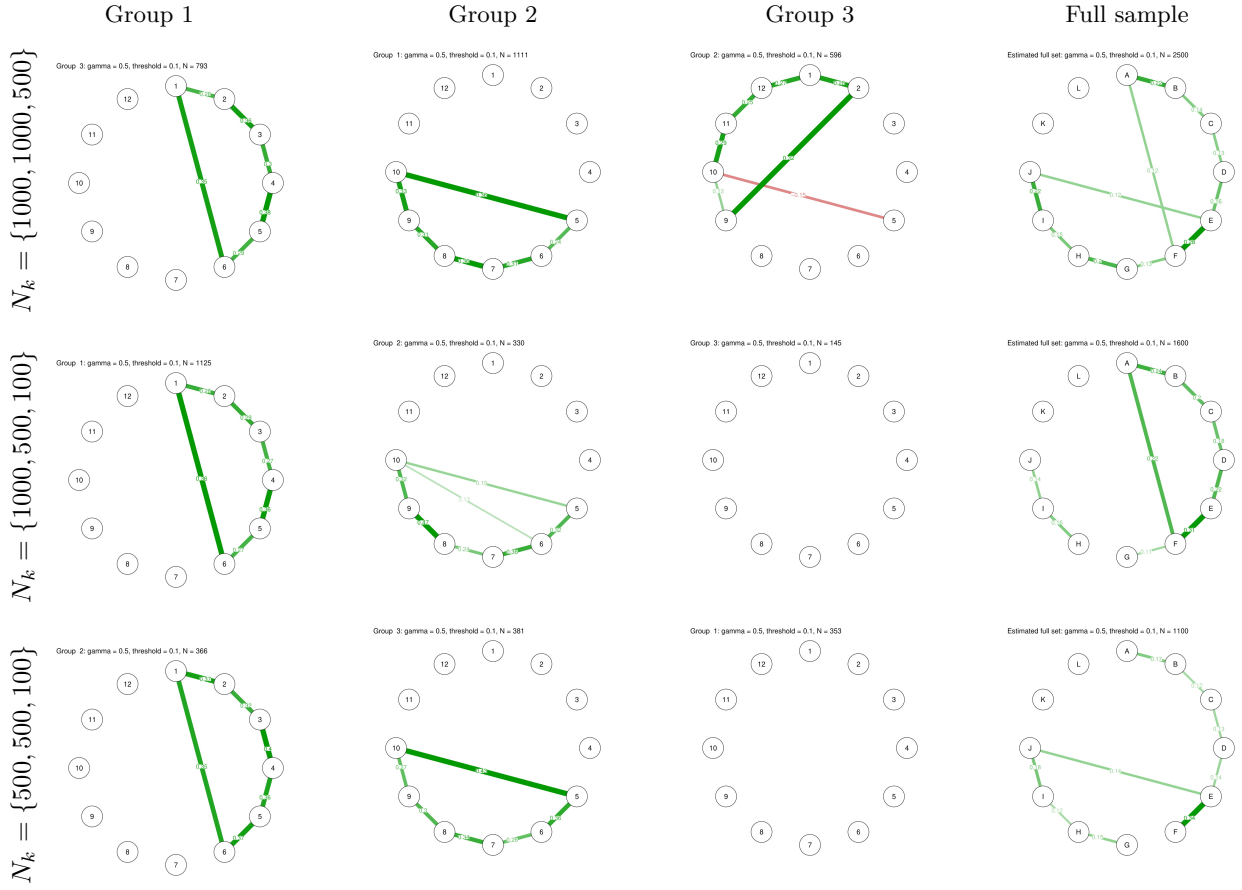
Figure 4.15: Estimated structures for subgroups and full sample of every balanced sample size condition.

enough sample sizes for stable results. Second, the case where there is one large majority group, one small minority group, and one middle sized group. Third, the case where on group is again significantly smaller, but has a problematically low sample size. These cases translate to the following combinations of group sizes: $N_k = \{1000, 1000, 500\}$, $N_k = \{1000, 500, 100\}$ and $N_k = \{500, 500, 100\}$. Again, full sample networks are computed for each condition. The resulting estimated networks are shown in Figure 4.15. Results are summarised in Figure 4.14.

Results show that all networks with $N = 1000$ are estimated correctly. Networks with $N = 500$ are estimated with errors when there is another, larger group in the same condition (i.e. $N_k = \{1000, 1000, 500\}$, group 3 and $N_k = \{1000, 500, 100\}$, group 2), but not when $N = 500$ is the largest sample size (i.e. $N_k = \{500, 500, 100\}$, groups 1 and 2). Networks with $N = 100$ are always estimated to be empty. For the full sample networks, the network structure resembles the largest groups of that condition. For example, in the $N_k = \{1000, 1000, 500\}$ condition, the structures of groups 1 and 2 are both fully present in the network of the full sample. This effect is less pronounced, but still present in the $N_k = \{500, 500, 100\}$ condition.

Edge weights follow the estimation accuracy: when estimation accuracy is high, the edge weights approach the true edge weights closely. However, lower group sizes lead to lower estimation accuracy, and a less close approximation of the true edge weights.
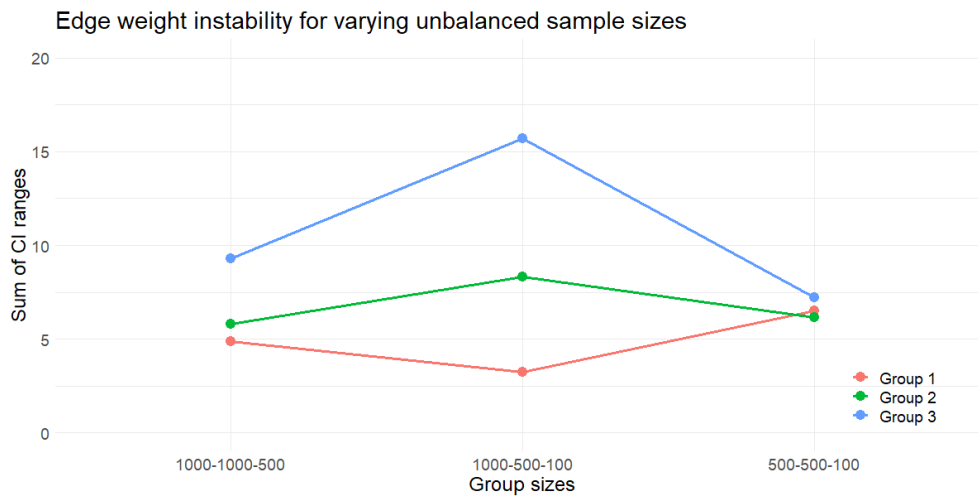
Figure 4.16: Summary of edge weight instability for varying combinations of unbalanced sample sizes. Network edge weight stability was calculated as the sum of confidence interval ranges for all edges of the network.

In terms of instability, the networks with the largest group sizes have the lowest edge weight instability. However, the instability of smaller groups seems to be relative to the larger groups, as evidenced by the edge weight stability for the $N_k = \{1000, 500, 100\}$ condition. In this condition, group 3 (empty network with $N = 100$) has a very high instability. In contrast, group 3 in the $N_k = \{500, 500, 100\}$ condition (also empty and with $N = 100$), has a very low instability.

# Social Media Disorder

## Data description

The data used here was collected as part of the Digital Youth project, a 5-year longitudinal study on the role of social media and gaming in the lives of Dutch adolescents. The study is based on self-report measures from high-school students in several Dutch cities. Data was collected annually in February and March through an online survey administered at school. Three waves of measurements were administered in 2016, 2017 and 2018, respectively. Differences in sample sizes between waves are largely due to entire schools joining or declining participation in the study. As a consequence, the distribution of demographic variables may vary between waves.

The resulting set contains responses by 4716 children and adolescents in the age range 10 to 18. Demographics for each wave are shown in Table 4.2. Each wave consisted of a questionnaire measuring 19 variables, presented in Table 4.3. Out of the 4716 participants, only 644 (13.7%) participated in all three waves.

Since the current application does not take into account time-series or directionality, waves were treated as separate, cross-sectional datasets. In the pre-processing stage, all incomplete cases (i.e. cases with missing values) were removed from each wave. The wave 2 dataset was selected for further analysis as it had the most complete cases out of all three waves.

In estimating the networks, the descriptive variables sex, age and education level were excluded from the estimations, to avoid results that discriminate subgroups based on descriptives. Furthermore, the categorical SMD score was excluded from the networks, as the numerical SMD score was already included. Figure 4.17 shows the estimated network for the full wave 2 SMD data.

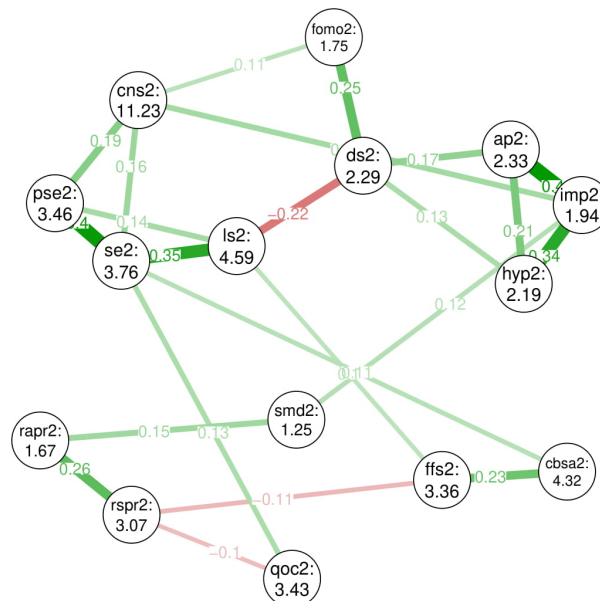| Wave | Complete cases | Gender | | Age | | | |
|------|----------------|--------|--------|-----|-----|------|----------|
| | | Male (%) | Female (%) | Min | Max | Mean | Std. dev. |
| 1 | 1640 | 897 (54.7%) | 743 (45.3%) | 10 | 16 | 13.3 | 0.916 |
| 2 | 2376 | 1224 (51.5%) | 1152 (48.5%) | 11 | 17 | 13.9 | 1.21 |
| 3 | 1890 | 953 (50.4%) | 937 (49.6%) | 11 | 18 | 14.4 | 1.50 |

Table 4.2: Demographics of SMD wave data.



Figure 4.17: Estimated CIS of the full SMD dataset. Nodes represent variables, edges represent partial correlations between variables.

| Variable | Abbreviation | Description |
|---|---|---|
| Sex | sex | Gender of the participant. |
| Age | age | Age of the participant. |
| Education level | edulvl | Education level of the participant. Contains five levels corresponding to Dutch education levels (i.e. VMBO, VMBO/HAVO, HAVO, HAVO/VWO, and VWO). |
| Social media disorder (raw score) | smd | Sum score of 9 binary items of the Social Media Disorder scale (Van Den Eijnden, Lemmens, & Valkenburg, 2016), measuring the severity of social media disorder (American Psychiatric Association, 2013). |
| Social media disorder (level) | smd_gr | Ordinal transformation of social media disorder score containing three (increasingly severe) levels, labeled *normal user*, *risky user* and *problematic user*, respectively. |
| Attention problems | ap | Part of the Dutch ADHD-questionnaire (Scholte & Van der Ploeg, 1999). Mean score of 9 items measuring symptoms of attentive problems, where higher values indicate more severe symptoms. |
| Impulsivity | imp | Part of the Dutch ADHD-questionnaire (Scholte & Van der Ploeg, 1999). Mean score of 6 items measuring symptoms of impulsivity, where higher values indicate more severe symptoms. |
| Hyperactivity | hyp | Part of the Dutch ADHD-questionnaire (Scholte & Van der Ploeg, 1999). Mean score of 6 items measuring symptoms of hyperactivity, where higher scores indicate more severe symptoms. |
| Depressive symptoms | ds | Mean score of 6 items of the Dutch version of the Depressive Mood Inventory (Kandel & Davies, 1982), measuring symptoms of depression, where higher scores indicate more severe symptoms. |
| Life satisfaction | ls | Mean score of 7 items of the Dutch version of the Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), where higher scores indicate higher life satisfaction. |
| Self esteem | se | Mean score of 5 out of 10 items of the Dutch translation of the Rosenberg Self-Esteem Scale (RSE; Rosenberg, Schooler, & Schoenbach, 1989), where higher scores indicate a higher degree of self-esteem. |
| Physical self esteem | pse | Mean score of 5 items measuring physical self-esteem: one's esteem of their appearance. Higher scores indicate a higher degree of physical self-esteem. |
| Narcissism | cns | Sum score of 10 items of the Dutch version of the Childhood Narcissism Scale (CNS; Thomaes, Stegge, Bushman, Olthof, & Denissen, 2008) measuring narcissistic qualities, where higher scores indicate a higher degree of narcissism. |
| Fear of missing out | fomo | Mean score of 5 items of the Fear of Missing Out Scale (Przybylski, Murayama, DeHaan, & Gladwell, 2013), measuring the degree to which one is fearful of missing out on positive experiences with their social circle. Higher scores indicate a higher degree of fear of missing out. |

| Variable | Abbreviation | Description |
|---|---|---|
| Perceived social competence | `cbsa` | Mean score of 5 items of the Dutch version of the Self-perception Profiles for Adolescents questionnaire (Competentie Belevingsschaal voor Adolescenten, CBSA; Treffers et al., 2002), measuring the degree to which one feels competent in social interactions. Higher scores indicate higher perceived social competence. |
| Intensity of meeting friends | `ffs` | Mean score of 4 items measuring the frequency and intensity with which the participant meets with friends, where higher scores indicate a higher intensity. |
| Restrictive parental rules | `rspr` | Mean score of 5 items adapted from the parenting practices questionnaire (van Den Eijnden, Spijkerman, Vermulst, van Rooij, & Engels, 2010), measuring the degree to which parents apply restrictive rules to their children's internet and game use. Higher scores indicate more use of restrictive rules. |
| Reactive parental rules | `rapr` | Mean score of 4 items adapted from the parenting practices questionnaire (van Den Eijnden et al., 2010), measuring the degree to which parents apply reactive rules to their children's internet and game use. Higher scores indicate more use of reactive rules. |
| Quality of communication | `qoc` | Mean score of 3 items asking participants whether they feel comfortable, understood, and taken seriously when talking about internet or game use with their parents. Higher scores indicate a higher quality of communication. |

Table 4.3: Variables in the SMD data.

# Three-group model

From the wave 2 SMD data, a family of models was computed with $k = 3$, distinguishing three subgroups in the data. The optimal $k = 3$ model was selected with $\lambda = 0.01933$, with EBIC = 89747. Figure 4.18 shows the estimated networks for the $k = 3$ model. For all networks, variable names and mean values are drawn in the nodes. Figure 4.19 shows the edge weight stability plots for the networks in Figure 4.18.

**Edge weight stability.** From the edge weight stability plots in Figure 4.19, a number of edges stand out as potentially spurious, since their confidence interval includes zero, while the edge is non-zero. Table 4.4 presents unstable edges for all groups. Any interpretation of these edges should be made with additional caution.

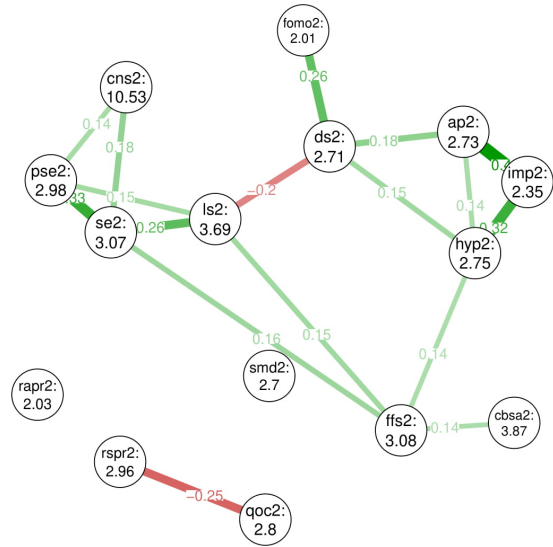| Groups | | |
|---|---|---|
| 1 | 2 | 3 |
| `ls2-qoc2` | `ds2-hyp2` | `se2-qoc2` |
| `fomo2-se2` | `cns2-pse2` | `smd2-ls2` |
| `ffs2-rspr2` | `cbsa2-ffs2` | `ds2-fomo2` |
| `ap2-ds2` | `ffs2-ls2` | `cns2-se2` |

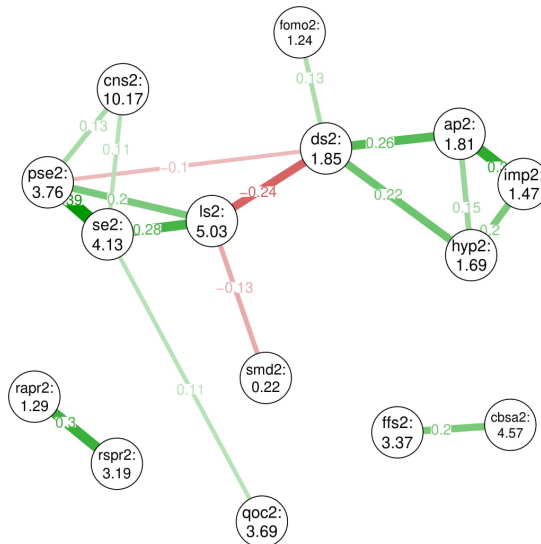Table 4.4: Unstable edges for networks in Figure 4.18.
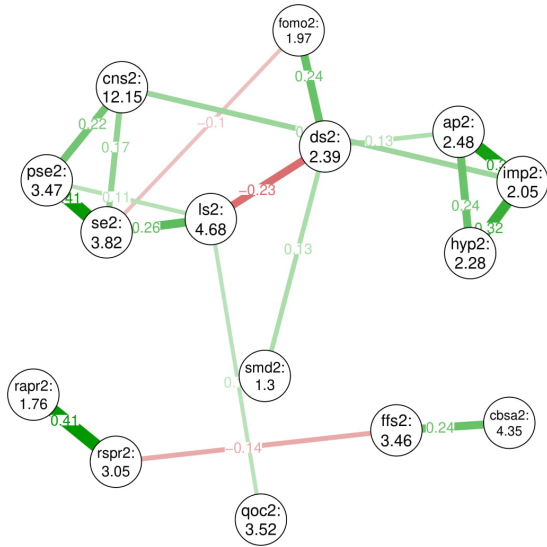
Figure 4.18: CISs for three subgroups estimated in the wave 2 SMD data using the proposed GMM/GGM method. Nodes represent variables, edge weights represent partial correlations between variables.

**Network structures.** Group 1 is the largest group in this model. Its network shows two disconnected structures. The strongest edges are `rapr2-rspr2`, `pse2-se2`, `ap2-imp2`, and `ap2-hyp2`. Group 1 is the only group that connects the variables `rapr2` and `rspr2` to any of the other variables. Specifically, there is a negatively weighted edge between `rspr2` and `ffs2`.

Group 2 is the smallest group in the model. Its network shows a more disconnected structure. The variables `rapr2` and `smd2` are disconnected, and therefore conditionally independent of all other variables. The strongest edges are `ap2-imp2`, `se2-pse2` and `hyp2-imp2`. Compared to the networks for groups 1 and 3, group 2 has more edges connected to the variable `ffs2`. In addition, group 2 does not have an edge `rapr2-rspr2`, unlike groups 1 and 3. Furthermore, group 2 has a unique, strong edge between `rspr2` and `qoc2`. Finally, group 2 has the highest mean values for `ds2` and `fomo2`, and the lowest mean values for `se2`, `pse2`, `ls2` and `cbsa2`.
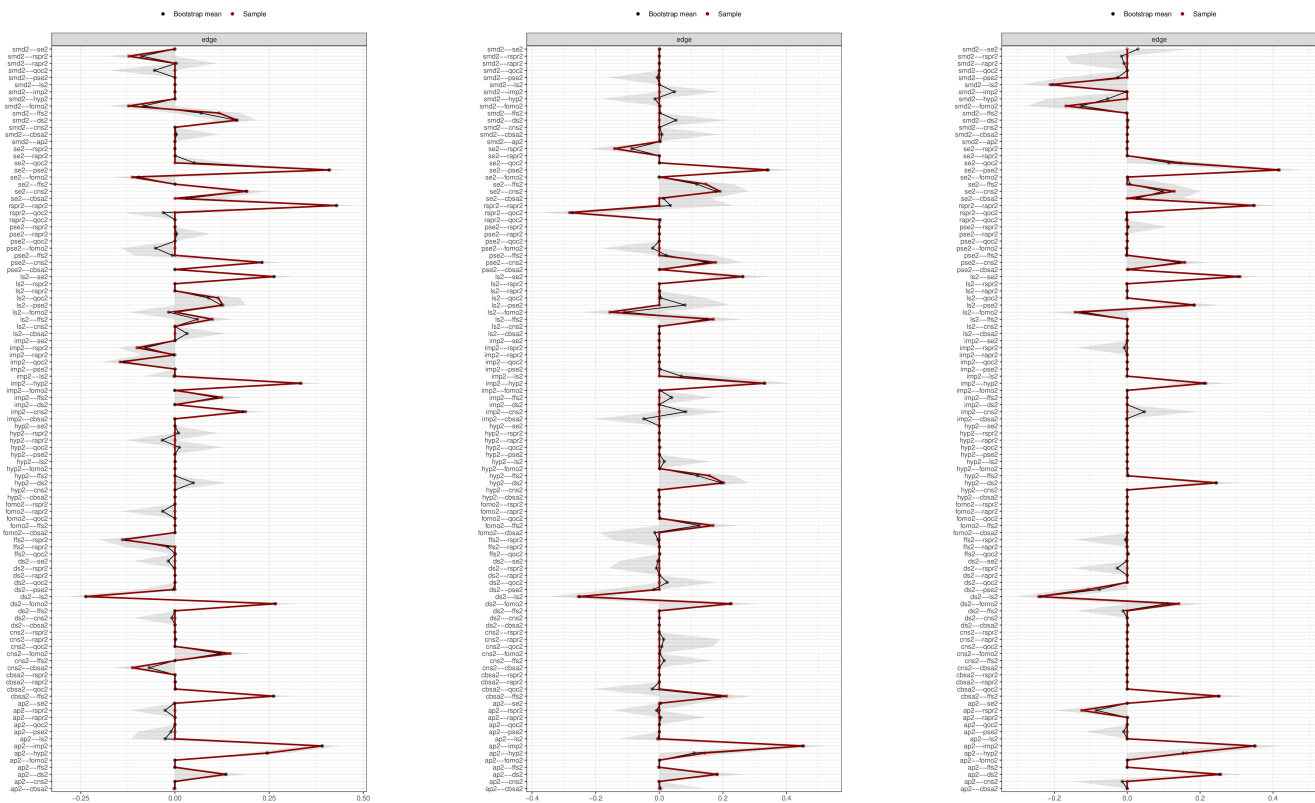
Figure 4.19: Stability analysis plots of the networks in Figure 4.18. From left to right: group 1, group 2, group 3.

Group 3 shows three cliques of variables, the two smallest being `rapr2-rspr2` and `ffs2-cbsa2`. The strongest edges are `se2-pse2`, `ap2-imp2` and `rapr2-rspr2`. Compared to groups 1 and 2, group 3 has most similarities with group 1. However, in group 3 the variable `smd2` is connected with a negatively weighted edge to the variable `ls2`. In addition, group 3 has the lowest mean values on `ds2`, `smd2`, `ap2`, `hyp2` and `imp2`, and the highest mean values on `ls2`, `se2`, `pse2` and `cbsa2`.

Comparing the three networks to each other, we see that their networks have a strong structure in common: `pse2-se2-ls2-ds2-ap2-imp2-hyp2`. In addition, the variables `ap2`, `hyp2` and `imp2` (i.e. ADHD variables) are connected with positively weighted edges for all groups, and so are the variables `pse2`, `se2` and `ls2` (i.e. quality of life variables). Interestingly, the variables `rapr2` and `rspr2` (i.e. parental rules) are connected for groups 1 and 3, but not for group 2.

**Centrality analysis.** Figure 4.20 shows plots of the standardised centrality for the subgroups in the three-group model. Figure 4.21 shows the stability of the centrality measures for each group. For all groups, strength and expected influence centrality are the most stable, never going below the recommended critical value of 0.5 (Epskamp et al., 2018). Betweenness centrality is stable for group 3, and only marginally unstable for groups 1 and 2, dropping below the critical value only at 30% of the original sample. Closeness centrality is absent for groups 2 and 3, but is stable for group 1. Therefore, strength and expected influence will be interpreted for all groups, betweenness centrality will be interpreted for all groups, and closeness centrality will be interpreted for group 1.

The three most central nodes for each group for each centrality measure are shown in Table 4.5. Based on strength and expected influence centrality, the most central nodes for all groups are `se2` (self-esteem), `pse2` (physical self-esteem), `ls2` (life satisfaction), `imp2` (impulsivity), and `ap2` (attentive problems). Betweenness and closeness centralities vary strongly per group, due to the different connectivity of the networks. For group 1, the nodes with the highest betweenness correspond to the nodes with the highest strength. For group 3, the nodes with the highest betweenness are `imp2` (impulsivity), `ffs2` (intensity of meeting friends) and `cns2` (narcissism). Incidentally, these are nodes that connect strong nodes to each other (e.g. `se2-cns2-imp2`), or that connect strong nodes to nodes that would otherwise be disconnected (e.g. `imp2-ffs2-rspr2`).
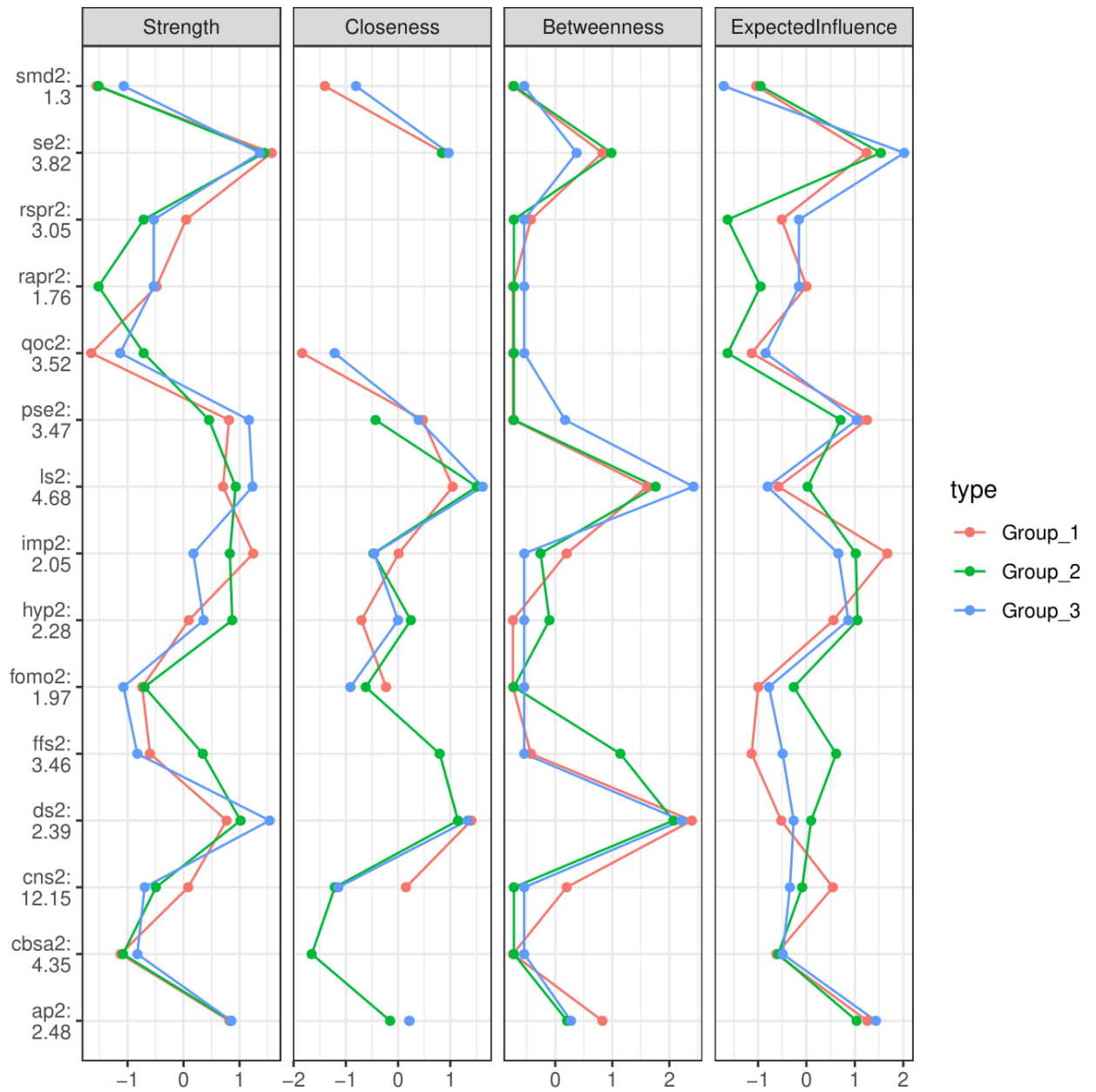
Figure 4.20: Standardised centrality plots for subgroups in the three-group model of the SMD data.

| Group | Strength | Expected Influence | Betweenness | Closeness |
|---|---|---|---|---|
| | se2 | se2 | ds2 | ds2 |
| 1 | imp2 | imp2 | ls2 | ls2 |
| | pse2 | pse2 | se2 | se2 |
| | se2 | se2 | ds2 | |
| 2 | ds2 | hyp2 | ls2 | NA |
| | ls2 | imp2 | ffs2 | |
| | ds2 | se2 | ls2 | |
| 3 | se2 | ap2 | ds2 | NA |
| | ls2 | hyp2 | se2 | |

Table 4.5: Three most central nodes per group per centrality measure, for networks of subgroups in the three-group model.
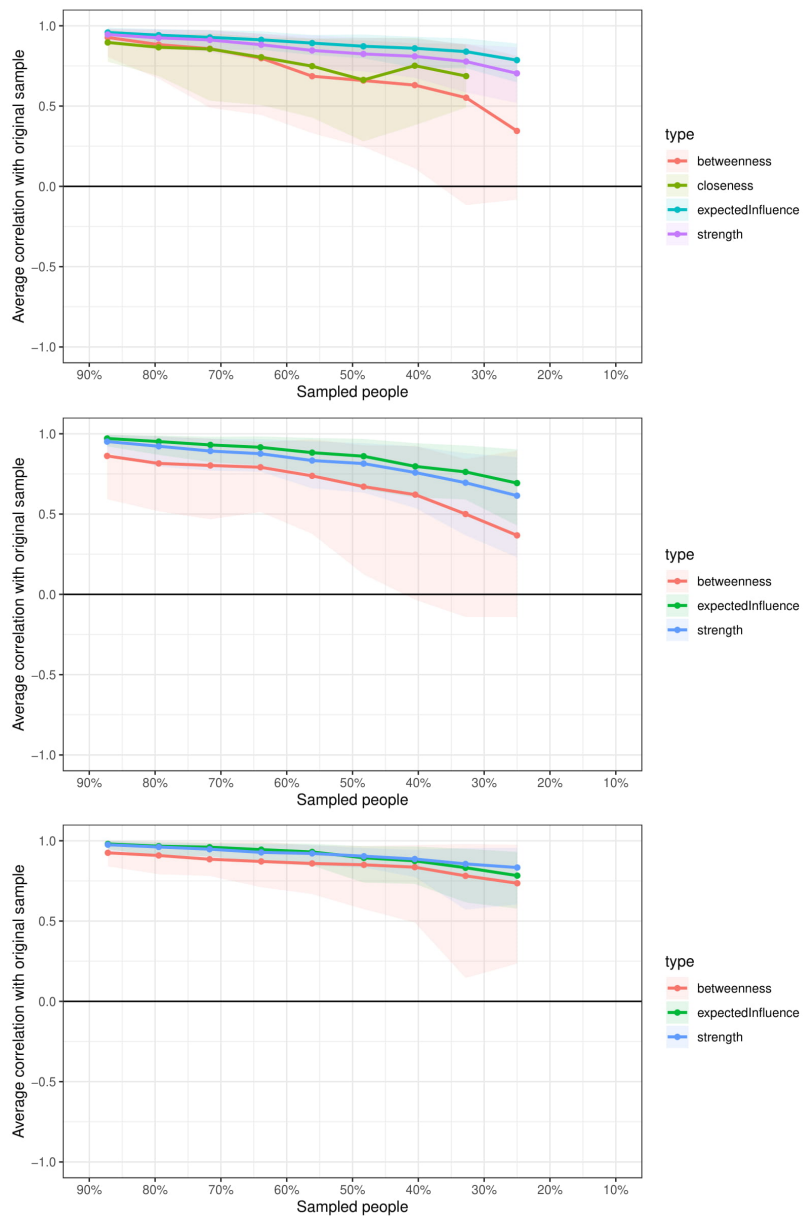


Figure 4.21: Centrality stability plots for subgroups in the three-group model of the SMD data. From top to bottom: group 1, group 2, group 3.

**Summary.** Intuitively, the network structure of group 1 suggests an *'average'* subgroup. This is the largest group, containing nearly half of the entire sample. Unique properties of this network include the negative partial correlation between restrictive parental rules and intensity of meeting friends, as well as a negative partial correlation between self-esteem and fear of missing out. The most central nodes are self-esteem and depressive symptoms. At face value, the structure of group 1 does not stand out compared to groups 2 and 3.

Intuitively, the network structure of group 2 suggests an *'insecure'* subgroup. Compared to groups 1 and 3, this group shows many connections stemming from intensity of meeting friends, which has positive correlations with life satisfaction, self-esteem and perceived social competence. In addition, fear of missing out has the highest mean value for this group and has the strongest correlation to depressive symptoms out of all groups. Furthermore, unlike groups 1 and 3, group 2 does not have a connection between reactive and restrictive parental rules. Instead, restrictive parental rules are negatively correlated to quality of communication: the more restrictive parental rules are used, the worse the quality of communication is, and vice versa. Finally, this group has the highest mean values for ADHD symptoms, depressive symptoms, fear of missing out and social media disorder score and the lowest mean values for quality of life measures (life satisfaction, self-esteem & physical self-esteem). In sum, the network for group 2 indicates a different role for restrictive parental rules and social interactions with peers.

Intuitively, the network structure of group 3 suggests a *'happy & healthy'* subgroup, most notably defined by low mean scores for clinical measurements (all ADHD variables, narcissism, depressive symptoms, and social media disorder score), and high mean values for quality of life measurements (life satisfaction, self-esteem & physical self-esteem). The structure is similar to group 1, where depressive symptoms and life satisfaction are very central in the network. However, unlike group 1, there are no negatively weighted edges between restrictive parental rules and intensity of meeting friends and between self-esteem and fear of missing out, implying that those variables are conditionally independent of each other for this group.

# Five-group model

From the wave 2 SMD data, another family of models was computed with $k = 5$, now distinguishing five subgroups in the data. The optimal $k = 5$ model was then selected with $\lambda = 0.00663$ and EBIC = 88793. For group 5, the method returned a group assignment that set the values of the variable `smd2` to zero for all group members. As a consequence, this variables does not have any variance and cannot be used to calculate partial correlations. Consequently, these variables were removed from the group 5 network. The respective edge weight stability plot, centrality plot and centrality stability plot were adapted to the revised network. Any comparisons of group 5 with other groups should be made with additional caution. Figure 4.22 shows the estimated networks for the $k = 5$ model. For all networks, variable names and mean values are drawn in the nodes. Figure 4.23 shows the edge weight stability plots for the networks in Figure 4.22.

**Edge weight stability.** Edge weight stability plots for the $k = 5$ model are shown in Figure 4.23. Edges for which the confidence interval included zero are summarised in Table 4.6. These edges should be interpreted with additional caution, as they are the least stable. Coincidentally, most unstable edges have small edge weights, i.e. $< 0.2$. The reverse is not necessarily true.

| | | Groups | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| ap2-smd2 | ap2-hyp2 | ap2-ds2 | ap2-ds2 | cns2-pse2 |
| cns2-pse2 | ap2-ds2 | cns2-se2 | ap2-hyp2 | cns2-se2 |
| cns2-ls2 | cbsa2-ffs22 | cbsa2-cns2 | cbsa2-ffs2 | cbsa2-ls2 |
| ds2-fomo2 | cbsa2-se2 | ds2-ls2 | cns2-smd2 | ds2-se2 |
| imp2-hyp2 | cns2-pse2 | imp2-cns2 | cns2-imp2 | imp2-rspr2 |
| rapr2-rspr2 | ds2-fomo2 | fomo2-qoc2 | ds2-se2 | |
| rspr2-qoc2 | | | ls2-ffs2 | |
| | | | rapr2-smd2 | |

Table 4.6: Unstable edges for networks in Figure 4.22
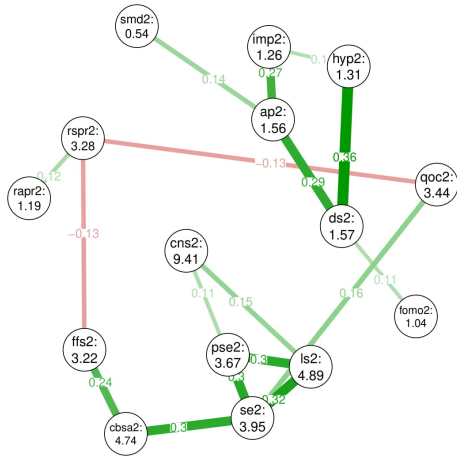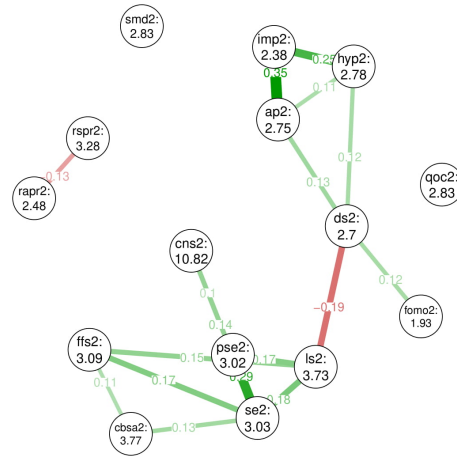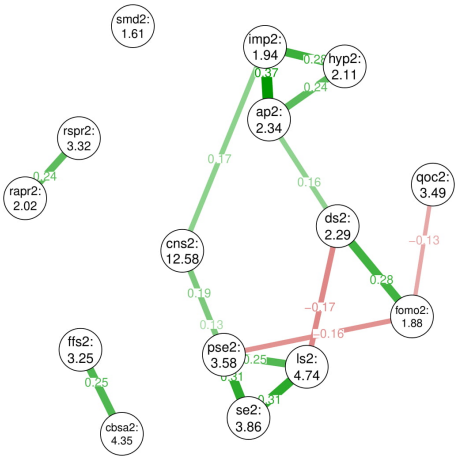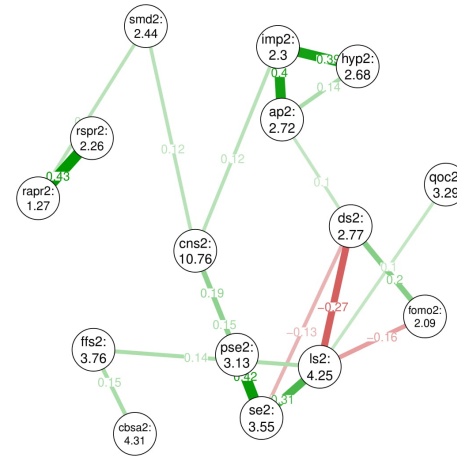
Figure 4.22: CISs for five subgroups estimated in the wave 2 SMD data using the proposed GMM/GGM method. Nodes represent variables, edge weights represent partial correlations between variables.

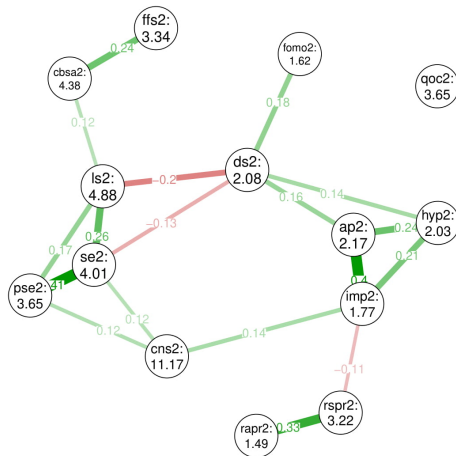**Network structures.** Group 1 is the smallest group of the five-group model. Its network shows two disconnected structures that together include all variables. The strongest edges are `hyp2-ds2`, `se2-ls2` and `cbsa2-se2`. Compared with the other groups, group 1 has two unique, negatively weighted edges, namely `rspr2-qoc2` and `rspr2-ffs2`. In addition, there is no negatively weighted edge `ds2-ls2`, which is present in all other networks. Furthermore, compared to groups 2, 3 and 4, group 1 has the lowest mean values on `ds2`, `smd2`, `ap2`, `hyp2` and `imp2`, and the highest mean values on `ls2`, `se2`, `pse2` and `cbsa2`.

For group 1, the absent edge between depressive symptoms and life satisfaction virtually disconnects two parts of the network: clinical measurements on one hand (all ADHD measures, depressive symptoms, social media disorder score, and fear of missing out), and the remaining variables on the other hand. It also implies that life satisfaction is independent of depressive symptoms for this group.

Group 2 shows a more sparse structure, with the variables `smd2` and `qoc2` disconnected from the rest of the network. The strongest edges are `imp2-ap2`, `se2-pse2`, **imp2**-`hyp2` and `ls2-ds2`. Compared to the other groups, group 2 is the only group that has a negative edge `rapr2-rspr2`. In addition, the mean value of `qoc` is the lowest out of all groups.

Group 2 shows high connectivity around the variable intensity of meeting friends. Furthermore, reactive and restrictive parental rules have a unique negative correlation in this group. Social media disorder has the highest mean score in this group, but is disconnected, implying conditional independence from all other variables.

Group 3 shows a network structure where, again, `smd2` is disconnected from the network. In addition, `rapr2-rspr2` and `ffs2-cbsa2` form two small, disconnected cliques. The strongest edges are `imp2-ap2`, `se2`-`pse2`, `imp2-hyp2` and `ds2-fomo2`. Compared to the other groups, group 3 shows the relative importance of the variable `fomo2`, which is strongly connected in the network.

Group 4 has a network structure that connects all variables to a single structure. The strongest edges in the network are `rapr2-rspr2`, `pse2-se2`, and `ap2-imp2`. Compared to the other groups, group 4 has the highest mean values on `ds2` and `fomo2`, and the second lowest mean values on `ls2`, `pse2` and `se2`.

Group 5 is the largest group in the five-group model. It has a different structure than groups 1 through 4, due to the exclusion of the variable `smd2` which was equal to zero for all group members. The network structure connects all variables except `qoc2`. The strongest edges in the network are `se2-pse2`, `ap2-imp2`, and `rapr2-rspr2`. Compared to the other groups, group 5 has the highest mean values on `ls2`, `se2` and `pse2` and the lowest mean values on `ds2`, `ap2`, `imp2` and `hyp2`.
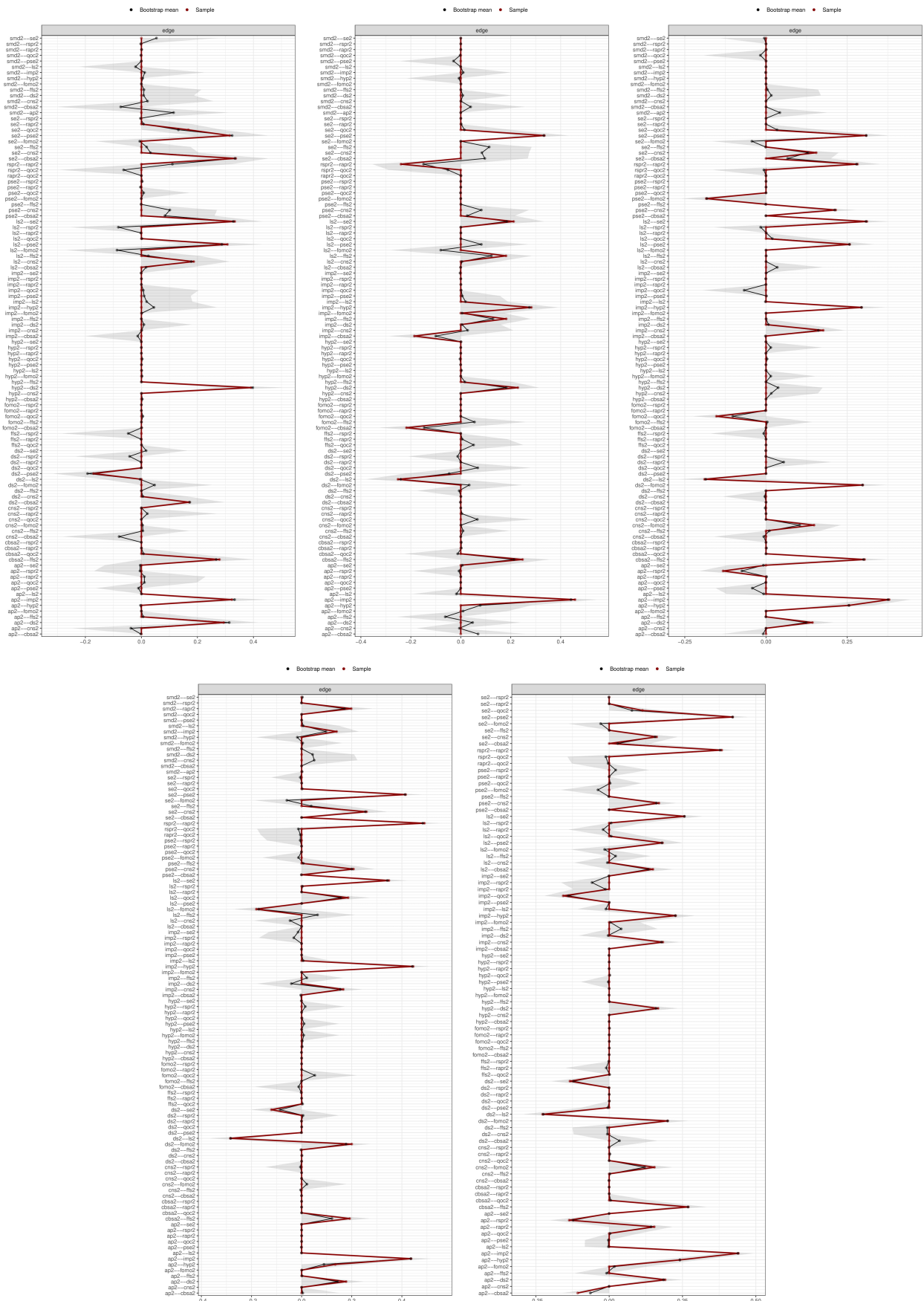
Figure 4.23: Stability analysis plots of the networks in Figure 4.18. From left to right and top to bottom: group 1, group 2, group 3, group 4, group 5.
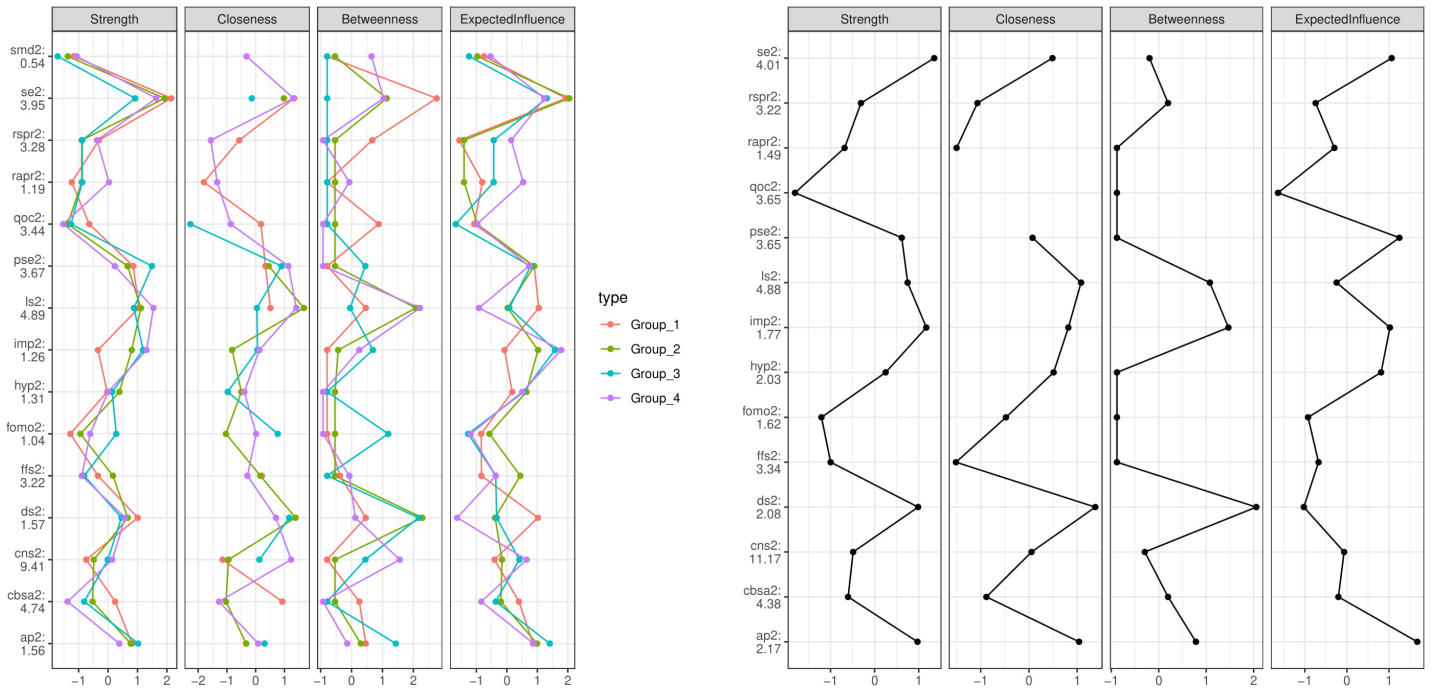
Figure 4.24: Standardised centrality plots for networks of subgroups in the five-group model of the SMD data. Left panel: groups 1-4. Right panel: group 5.

**Centrality analysis.** Figure 4.25 shows the stability of the centrality measures for the groups in the five-group model. Figure 4.24 shows plots of the standardised centrality for each group. Strength and expected influence centrality are stable for most groups, only reaching values below 0.5 in group 2. The stability of betweenness centrality differs strongly for each group. Whereas it is clearly unstable for groups 2, 3 and 4, it shows a much more stable pattern for groups 1 and 5. Closeness centrality was not applicable to any of the groups, due to the unstable nature of their networks. Therefore, strength and expected influence will be interpreted for all groups, betweenness centrality will be interpreted for groups 1 and 5, and closeness centrality will not be taken into account.

The three most central nodes for each group, for each centrality measure are shown in Table 4.7. Based on strength and expected influence centrality, some of the most central nodes for all groups are `se2` (self-esteem), `ls2` (life-satisfaction), `pse2` (physical self-esteem) and `ds2` (depressive symptoms). Betweenness centrality for groups 1 and 5 reflects the differences in structure: group 1 uniquely connects the variable `rspr2` (restrictive parental rules) to the variables `ffs2` (intensity of meeting friends) and `qoc2` (quality of communication). For group 5 on the other hand, the variable `ds2` (depressive symptoms) provides a bridging function between variables on the right- and left-hand side of the graph.

**Summary.** Intuitively, group 1 resembles the 'happy & healthy' subgroup from the three-group model, shown by the low mean values of depressive symptoms, social media disorder and ADHD symptoms. However, the same similarities apply to group 5. A distinguishing property of this group, compared to group 5, are the negative correlations connected to restrictive parental rules. In addition, life satisfaction and depressive symptoms are conditionally independent for this group. This suggest that group 1 may be a subgroup of the 'happy & healthy' group, characterised by the above properties.

Intuitively, group 2 resembles the 'insecure' group from the three-group model, evidenced by low mean values on quality of life measures, high mean values of social media disorder, depressive symptoms and ADHD symptoms, and the central position of depressive symptoms in the network. However, the same similarities apply to group 4. In addition, similar to the 'insecure' group, but unlike group 4, group 2 has strong connectivity around intensity of meeting friends, and a low mean value for perceived social competence. This suggests that group 2 is a subgroup of the 'insecure' group, characterised by social competence.

Intuitively, group 3 resembles the 'average' group from the three-group model in terms of means and

| Group | Strength | Expected Influence | Betweenness |
|:-:|:-:|:-:|:-:|
| 1 | se2<br>ls2<br>ds2 | se2<br>ls2<br>ds2 | se2<br>qoc2<br>rspr2 |
| 2 | se2<br>ls2<br>imp2 | se2<br>imp2<br>ap2 | NA |
| 3 | pse2<br>imp2<br>ap2 | imp2<br>ap2<br>se2 | NA |
| 4 | ls2<br>se2<br>imp2 | imp2<br>se2<br>ap2 | NA |
| 5 | se2<br>imp2<br>ds2 | ap2<br>pse2<br>se2 | ds2<br>imp2<br>ls2 |

Table 4.7: Three most central nodes per group per centrality measure, for networks of subgroups in the five-group model.
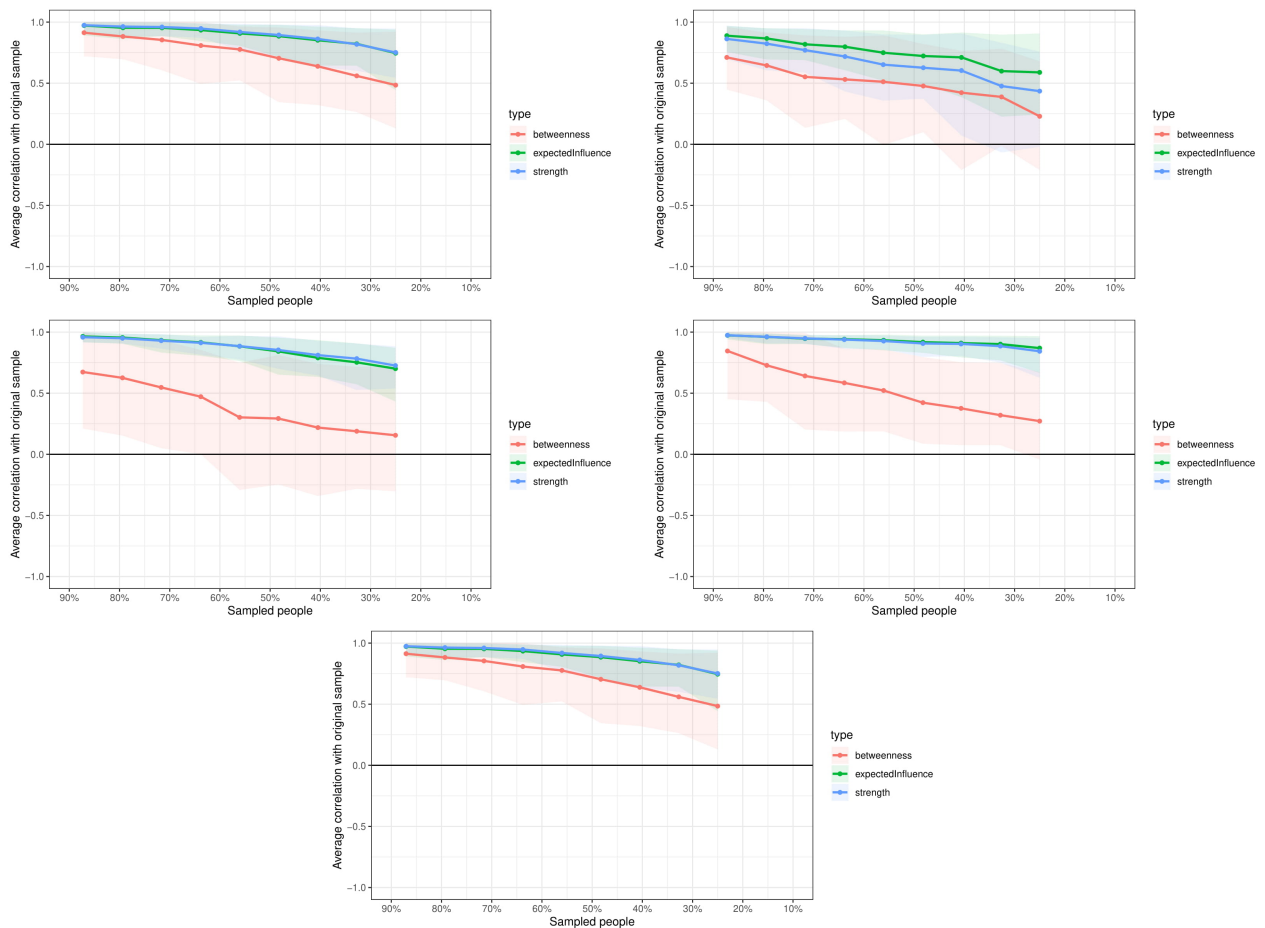


Figure 4.25: Centrality stability plots for subgroups in the three-group model of the SMD data. From left to right, from top to bottom: group 1, group 2, group 3, group 4, group 5.

structure. However, it has two unique edges connected to fear of missing out.

Intuitively, group 4 resembles the 'insecure' group from the three-group model, similar to group 2. However, compared to group 2, it has less connectivity around intensity of meeting friends, and more edges connected to life satisfaction, suggesting that group 4 is also a subgroup of the 'insecure' group, but characterised by life satisfaction.

Intuitively, group 5 also resembles the 'happy & healthy' group, mainly due to the fact that social media disorder score was equal to zero for this group. In addition, quality of life measures have high mean values. However, it's structure does not have any unique properties that significantly distinguish it from other groups.

# Public opinion of immigration and refugees

## Data description

To investigate potential subtypes of public opinion of immigration and refugees, data was used from round 7 of the European Social Survey (ESS round 7, 2014). The ESS is a bi-annual survey among European countries, polling public opinion on various societal and socio-economical topics, such as politics, media and social trust, health and inequality, justice and immigration. The topics polled differ from round to round. The topic of immigration was polled in round 1 (2002) and round 7 (2014). The latter was selected because it is more recent.

Round 7 of the ESS contained 41 items polling opinions about immigration, immigrants and refugees. From these items, 18 items were retained for further analysis. The main reasons for exclusion were if a variable was nominal (i.e. variable has only a few levels with no clear ordering), if the question asked did not directly relate to one's opinion about immigrants or refugees, or if the variable represented an administrative value (e.g. group number). Table 4.9 presents descriptions of the 18 retained items.

In the pre-processing stage, items selected for inclusion were retained. Subsequently, all incomplete cases were removed from the dataset. The resulting dataset contained a total of 31 385 participants from 21 countries. Age and gender emographics can be found in Table 4.8. Figure 4.26 shows the estimated network for the full ESS data.

| Gender | | | Age | | | |
|---|---|---|---|---|---|---|
| Male (%) | Female (%) | Missing | Min | Max | Mean | Std. dev. |
| 15183 (48.4%) | 16191 (51.6%) | 11 (0.00%) | 14 | 104 | 48.4 | 18.3 |

Table 4.8: Age and gender demographics for final ESS dataset.



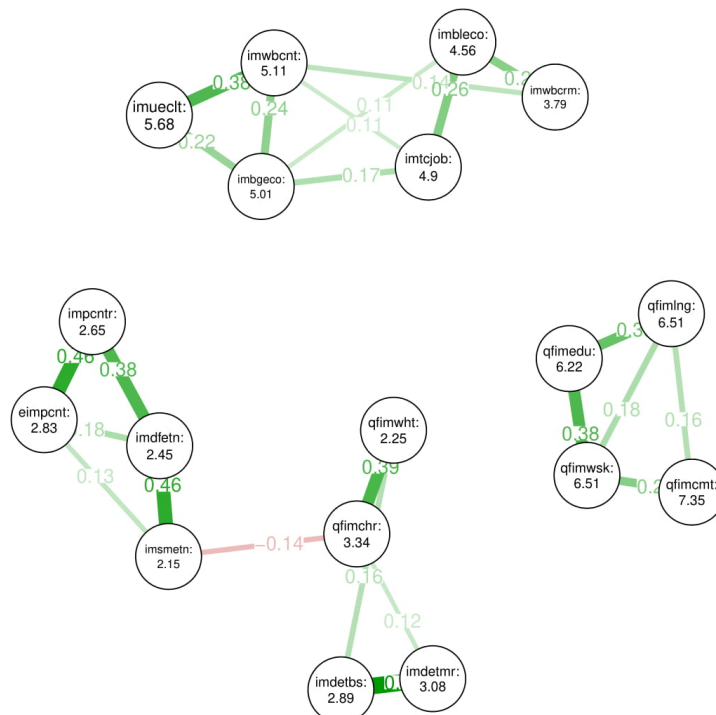Network of full ESS data: gamma = 0.5, threshold = 0.1, N = 31385

Figure 4.26: Estimated CIS of the full ESS dataset. Nodes represent variables, edge weights represent partial correlations between variables.

| Item | Abbreviation | Measurement |
|---|---|---|
| Allow many/few immigrants of same race/ethnic group as majority | `imsmetn` | 4-point Likert scale |
| Allow many/few immigrants of different race/ethnic group from majority | `imdfetn` | 4-point Likert scale |
| Allow many/few immigrants from poorer countries outside Europe | `impcntr` | 4-point Likert scale |
| Allow many/few immigrants from poorer countries inside Europe | `eimpcnt` | 4-point Likert scale |
| Qualification for immigration: good educational qualifications | `qfimedu` | 11-point Likert scale |
| Qualification for immigration: speak country's official language | `qfimlng` | 11-point Likert scale |
| Qualification for immigration: Christian background | `qfimchr` | 11-point Likert scale |
| Qualification for immigration: be white | `qfimwht` | 11-point Likert scale |
| Qualification for immigration: work skills needed in country | `qfimwsk` | 11-point Likert scale |
| Qualification for immigration: committed to way of life in country | `qfimcmt` | 11-point Likert scale |
| Taxes and services: immigrants take out more than they put in or less | `imbleco` | 11-point Likert scale |
| Immigration bad or good for country's economy | `imtbgeco` | 11-point Likert scale |
| Immigrants take jobs away in country or create new jobs | `imtcjob` | 11-point Likert scale |
| Immigrants make country's crime problems worse or better | `imwbcrm` | 11-point Likert scale |
| Immigrants make country worse or better place to live | `imwbcnt` | 11-point Likert scale |
| Immigrant different race/ethnic group majority: your boss | `imdetbs` | 11-point Likert scale |
| Immigrant different race/ethnic group majority: married close relative | `imdetmr` | 11-point Likert scale |
| Country's cultural life undermined or enriched by immigrants | `imueclt` | 11-point Likert scale |

Table 4.9: Variables in the ESS data.

# Three-group model

From the ESS data, a family of models was computed with $k = 3$, distinguishing three subgroups in the data. The optimal $k = 3$ model was selected with $\lambda = 0.00843$ and EBIC $= 2019887$.

For group 3, the method returned a group assignment that set the values of the variables `qfimwht`, `imdetbs`, and `imdetmr` to zero for all group members. As a consequence, these variables do not have any variance and cannot be used to calculate partial correlations. Consequently, these variables were removed from the group 3 network. The respective edge weight stability plot, centrality plot and centrality stability plot were adapted to the revised network. Any comparisons of group 3 with other groups should be made with additional caution.

Figure 4.27 shows the estimated networks for the $k = 3$ model. For all networks, variable names and mean values are drawn in the nodes. Figure 4.28 shows the edge weight stability plots for the networks in Figure 4.27.
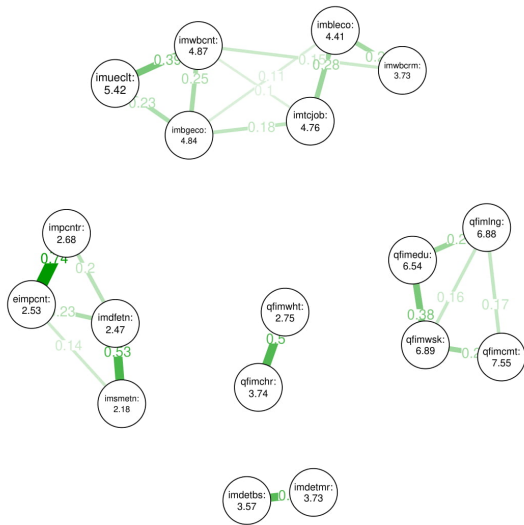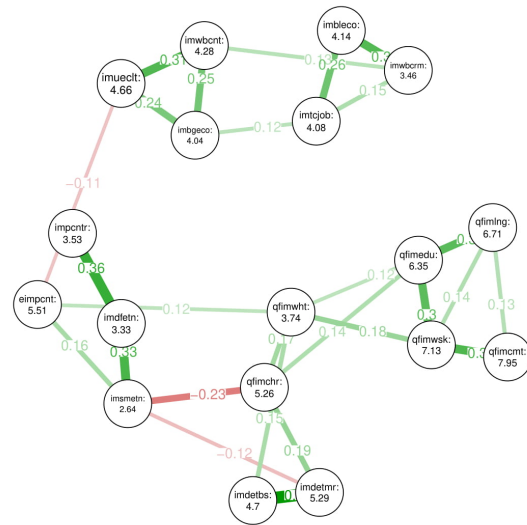


Figure 4.27: CISs for three subgroups estimated in the ESS data using the proposed GMM/GGM method. Nodes represent variables, edge weights represent partial correlations between variables.

Figure 4.28: Stability analysis plots of the networks in Figure 4.18. From left to right: group 1, group 2, group 3.

**Edge weight stability.** Edge weight stability plots are presented in Figure 4.28. For all groups, no edges have CIs overlapping with zero, indicating that all edges are stable.

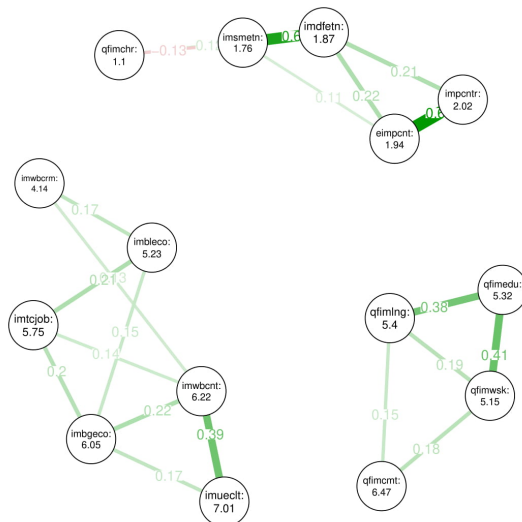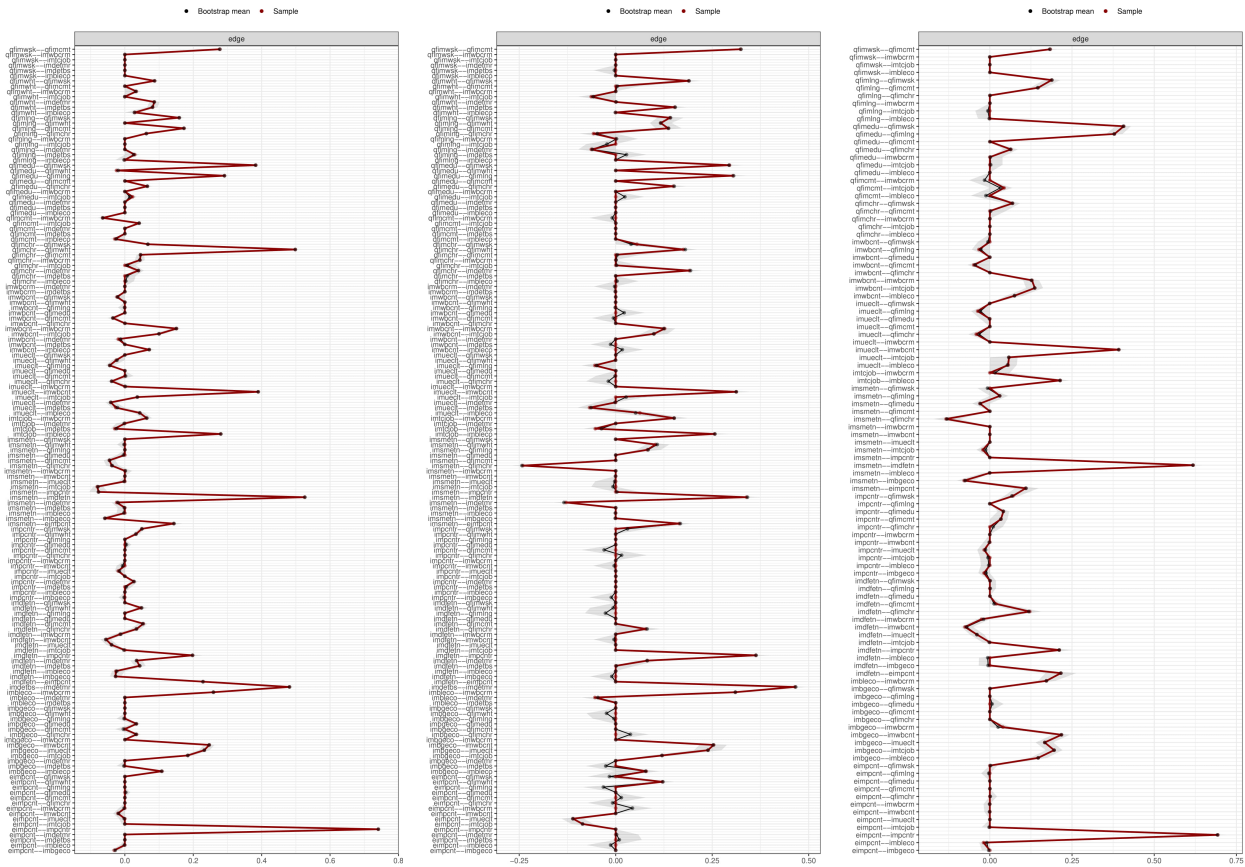**Network structures.** Group 1 has the largest group size. Its network shows five separate cliques of variables of differing sizes. All edges in the network are positively weighted. The strongest edges are `eimpcnt-impcntr`, `imdfetn-imsmetn`, `qfimchr-qfimwht`, and `imdetbs-imdetmr`. Compared to group 2, the group 1 network is more sparse and has a more fractured structure. In addition, group 1 has a small number of very strong edge weights.

The network for group 2 shows a single structure that connects all variables. The strongest edges are `imdetbs-imdetmr`, `impcntr-imdfetn` and `imsmetn-imdfetn`. Compared to group 1, group 2 has no edge `eimpcnt-impcntr`, even though this was by far the strongest edge of group 1. In addition, a relatively strong, negatively weighted edge `imsmetn-qfimchr` was introduced. While all nodes are connected, the top clique is connected to the rest of the network through the negatively weighted edge `imueclt-eimpcnt`.

For all group members of group 3, the values of variables `eimpcnt`, `imdetbs` and `imdetmr` are exactly zero. As a consequence, these variables are excluded from the conditional independence structure, as they have no variance and thus no partial correlations with any of the other variables. This is especially noteworthy because the excluded variables have very strong connections in groups 1 and 2. The network for group 3 shows three cliques of variables. The strongest edges in the remaining network are `imdfetn-imsmetn`, `eimpcnt-impcntr`, `qfimedu-qfimwsk` and `qfimedu-qfimlng`. Compared to groups 1 and 2 in terms of mean values, group 3 has the lowest mean values on `qfimchr`, `qfimlng`, `qfimedu`, `qfimwsk`, `qfimcmt`, `imsmetn`, `imdfetn`, `impcntr`, and `eimpcnt`. Also, group 3 has the highest mean values on `imueclt`, `imwbcnt`, `imbgeco`, `imbleco`, `imwbcrm`, and `imtcjob`.

Comparing the groups structurally, they share all of the edges between the variables `qfimedu`, `qfimlng`, `qfimwsk`, `qfimcmt`, and most of the edges between the variables `imueclt`, `imwbcnt`, `imbgeco`, `imtcjob`, `imbleco` and `imwbcrm`. Interestingly, both of these structures are also visible in the network for the full data (Figure 4.26).

51

| Group | Strength | Expected Influence | Betweenness | Closeness |
|-------|----------|--------------------|-------------|-----------|
|       | eimpcnt  | eimpcnt            | imwbcnt     | imbgeco   |
| 1     | imdfetn  | imdfetn            | imbgeco     | imwbcnt   |
|       | impcntr  | impcntr            | imdfetn     | imtcjob   |
|       | qfimwsk  | qfimwsk            | eimpcnt     | eimpcnt   |
| 2     | imsmetn  | qfimwht            | imueclt     | imsmetn   |
|       | imdetmr  | qfimedu            | imsmetn     | qfimwht   |
|       | imdfetn  | imdfetn            | imdfetn     | imwbcnt   |
| 3     | eimpcnt  | eimpcnt            | imwbcnt     | imbgeco   |
|       | imwbcnt  | imwbcnt            | imbgeco     | impcntr   |

Table 4.10: Three most central nodes per group per centrality measure, for networks of subgroups in the three-group model of the ESS data.

**Centrality analysis.** Figure 4.29 shows plots of the standardised centrality for the subgroups in the three-group model. Figure 4.30 shows the stability of the centrality measures for each group. For all groups, strength and expected influence centrality are the most stable. In addition, none of the centrality indices go below the recommended critical value of 0.5. Therefore, all centrality indices will be interpreted for all groups.

The three most central nodes for each group for each centrality measure are shown in Table 4.10. In terms of strength and expected influence centrality, the most central nodes for groups 1 and 3 are eimpcnt (allow immigrants from poorer countries in Europe) and imdfetn (allow many/few immigrants from different ethnic group as majority). For group 2, this is the variables qfimwsk (qualification for immigration: work skills needed in country) and imsmetn (allow many/few immigrants from same ethnic group as majority). For betweenness centrality, the most central nodes for groups 1 and 3 are imwbcnt (immigrants make country worse or better place to live), imdfetn (allow many/few immigrants from different ethnic group as majority) and imbgeco (immigrants bad or good for country's economy). For group 2, however, the variables with the highest betweenness centrality are eimpcnt (allow immigrants from poorer countries in Europe) and imueclt
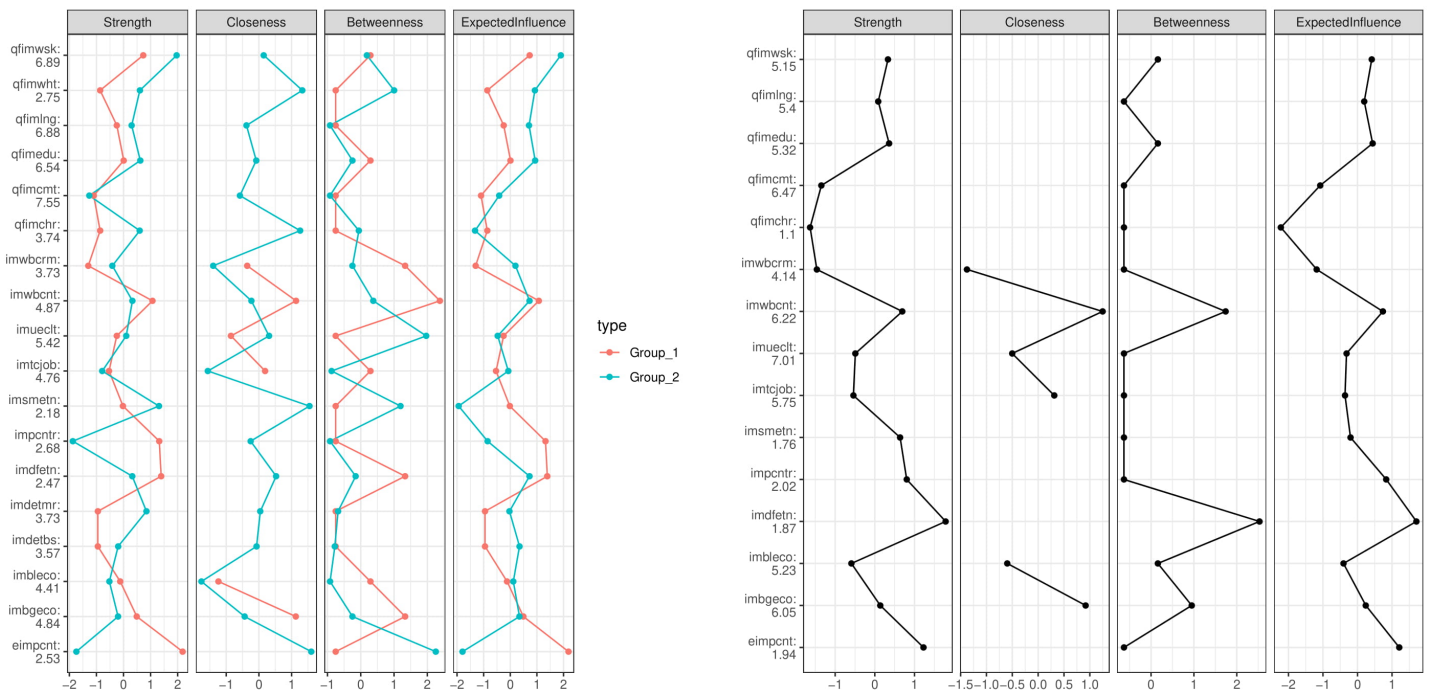


Figure 4.29: Standardised centrality plots for subgroups in the three-group model of the ESS data. Left panel: groups 1 & 2. Right panel: group 3.

(country's cultural life undermined or enriched by immigrants), due to the bridging function of the edge `eimpcnt-imueclt`. In terms of closeness centrality, groups 1 and 3 again share the same most central nodes, `imbgeco` (immigration bad or good for country's economy) and `imwbcnt` (immigrants make country worse or better place to live) being the most closely connected nodes in the networks. Group 2, however, shows a higher closeness centrality for the variables `eimpcnt` (allow immigrants from poorer countries in Europe) and `qfimwht` (qualification for immigration: be white).
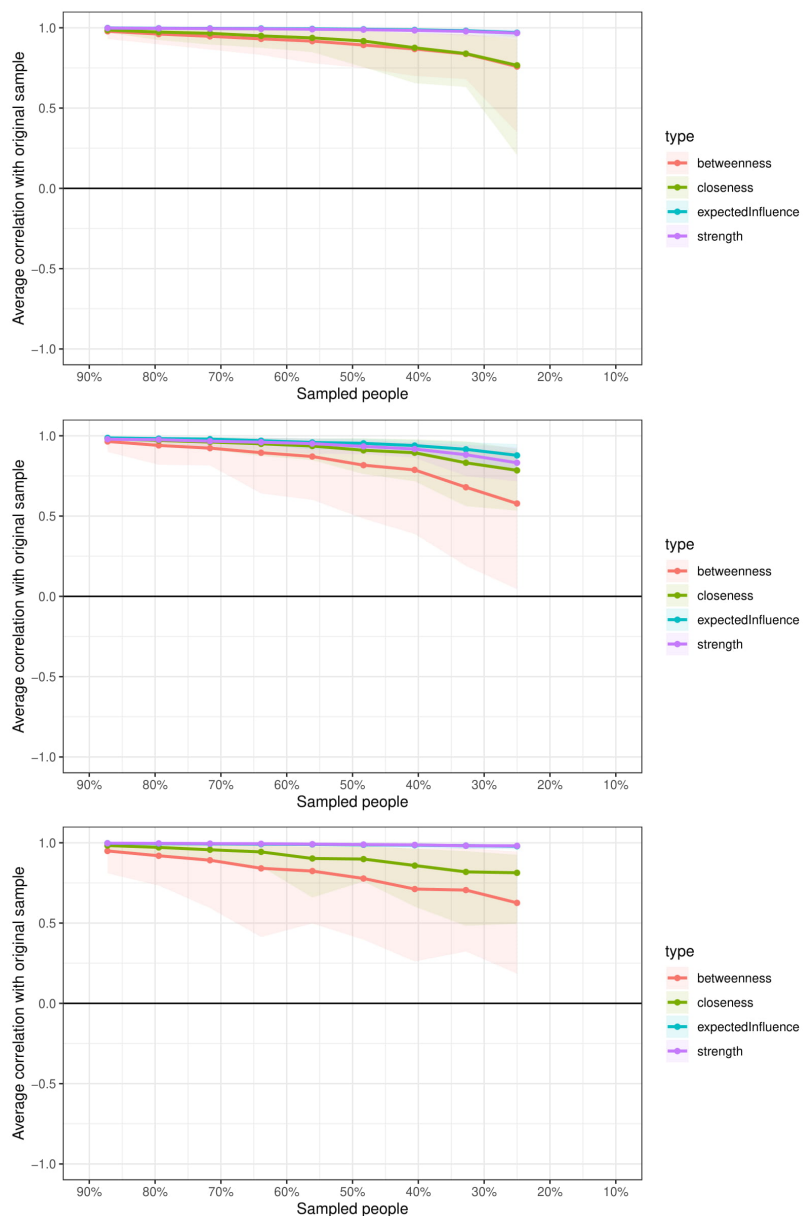


Figure 4.30: Centrality stability plots for subgroups in the three-group model of the ESS data. From top to bottom: group 1, group 2, group 3.

**Summary.** Intuitively, the network structure of group 1 suggests an *'indifferent'* subgroup. It is the largest group. Furthermore, the very high positive edge weight between `eimpcnt` and `impcntr` suggests that his subgroup views immigrants from outside or from within Europe the same. In addition, how much this subgroup values an immigrant being white (`qfimwht`) or Christian (`qfimchr`) has no impact on any of the other variables. However, compared to group 3, they are not unbothered by immigrants being their boss (`imdetbs`) or marrying a close relative (`imdetmr`). Furthermore, they value immigrants' education (`qfimedu`), native language proficiency (`qfimlng`), work skills (`qfimwsk`) and commitment to native values (`qfimcmt`) on average about as much as group 2, and much higher than group 3. How important this is found, is moderated by how this subgroup views the impact of immigrants on jobs, the economy, and country welfare, as evidenced by the high betweenness and closeness centralities of `imbgeco`, `imtcjob` and `imwbcnt`, respectively. In general, the network properties of this subgroup do not suggest either positive or negative attitudes towards immigrants.

Intuitively, the network structure of group 2 suggests an *'opposed'* subgroup. This is evidenced most by the differences in central nodes, compared with the other groups. For group 2, the variables `qfimwht` (qualification for immigration: be white) and `qfimwsk` (qualification for immigration: work skills needed in country) are strongly connected within the network, whereas this is not the case for groups 1 and 3. In addition, this network shows high importance of immigrants being Christian (`qfimchr`). This variable is connected closely to the rest of the network and has the highest mean value out of all groups. The strong negative partial correlation between `qfimchr` and `imsmetn` (allow immigrants of the same ethnicity) suggests that for this group, immigrants of the same ethnicity being welcomed depends on their being Christian. Furthermore, this group has the highest mean values on the items measuring how much they would mind having an immigrant as their boss (`imdetbs`) or married to a close relative (`imdetmr`). Finally, compared to groups 1 and 3, group 2 is missing the edge `eimpcnt`-`impcntr` (allow immigrants from poorer countries in Europe & allow immigrants from poorer countries outside Europe). Given the high edge weights in groups 1 and 3, this is a notable difference. Statistically, it means that there is no common variance explained between these variables, even though both variables inquire an opinion about immigrants from poor countries. The subtle, but crucial difference is whether these immigrants are from inside or outside Europe. Interestingly, the mean value for `eimpcnt` is higher, indicating less willingness to allow immigrants from poor countries within Europe.

Intuitively, the network structure of group 3 suggests a *'favourable'* subgroup. For this group, all 7423 group members gave the lowest possible value (0) when asked how much they would mind having an immigrant be their boss (`imdetbs`), be married to a close relative (`imdetmr`), or be non-white (`qfimwht`). This alone is a clear definition of this group, which is reflected in the graphical representation. In addition, similar to group 1, this subgroup views immigrants from within and outside Europe equally, but is more willing to allow immigrants compared to group 1. This subgroup has the lowest mean values for all items evaluating how many immigrants should be barred (i.e. immigrants with the same ethnicity (`imsmetn`), or with different ethnicity (`imdfetn`), and items evaluating the bar to qualify for immigration (i.e. education (`qfimedu`), work skills (`qfimwsk`), native language (`qfimlng`), commitment to national way of life (`qfimcmt`), their being white (`qfimwht`) and their being Christian (`qfimchr`)). Finally, out of all groups, this subgroup has the highest mean values on items evaluating whether immigrants make the country worse or better (`imwbcnt`), how much immigrants enrich culture (`imueclt`), and how good immigrants are for the economy (`imbgeco`). In general, this group has a favourable attitude towards immigrants. Structurally, three variables were removed. The remaining network has three cliques for the different categories of questions: qualification for immigration, how many or few immigrants should be allowed, and perceived added value of immigrants.

# Five-group model

From the ESS data, another family of models was computed with $k = 5$, now distinguishing five subgroups in the data. The optimal $k = 5$ model was selected with $\lambda = 0.0111$ and EBIC = 1978750. For the five-group model, the method again set group estimations in such a way, that some variables in groups 2 and 3 had no variance. Specifically, group 2 is missing the variables `qfimwht`, `imdetbs` and `imdetmr`, for which all values were zero. Group 3 is missing the variable `eimpcnt`, for which all values were equal to 3.

Figure 4.22 shows the estimated networks for the $k = 5$ model. For all networks, variable names and mean values are drawn in the nodes. Figure 4.23 shows the edge weight stability plots for the networks in Figure 4.22.

**Edge weight stability.** Edge weight stability plots are presented in Figure 4.32. Again, for all groups, no edges have CIs overlapping with zero, indicating that all edges are stable.

**Network structures.** The network structure for group 1 has three cliques, with all but one edge (`qfimchr-imsmetn`) having positive weights. The strongest edges are `impcntr-eimpcnt`, `imwbcnt-imueclt` and `imdetbs-imdetmr`.

Group 2 shows a more sparse network structure, with three cliques and one variable (`qfimchr`) completely disconnected from the rest of the network. The strongest edges are `eimpcnt-impcntr` and `imdfetn-imsmetn`. Compared to all other groups, it has the lowest mean value of `qfimchr`, `qfimedu`, `qfimlng`, `qfimcmt` and `qfimwsk`.

The network structure for group 3 has two disconnected structures, with only one negatively weighted edge (`imdetbs-imtcjob`). The strongest edges are `imdetbs-imdetmr`, `imdfetn-impcntr`, and `qfimwht-qfimchr`.

Group 4 shows a network structure with two disconnected structures and slightly more edges. The strongest edges are `impcntr-eimpcnt`, `imdetbs-imdetmr`, `imsmetn-imdfetn` and the negatively weighted edge `qfimchr-imsmetn`. Compared to other groups, group 4 has the highest mean value of `qfimchr`.

Group 5 shows a more fractured network structure, with four disconnected cliques. All edges are positive. The strongest edges are `imdetbs-imdetmr`, `impcntr-eimpcnt`, `impcntr-imdfetn` and `eimpcnt-imdfetn`. In addition, group 5 is the largest group in terms of sample size.

**Centrality analysis.** Figure 4.33 shows plots of the standardised centrality for the subgroups in the five-group model. Figure 4.34 shows the stability of the centrality measures for each group. For all groups, strength and expected influence centrality are the most stable, never taking values below 0.5. Betweenness centrality is stable for groups 1, 2 and 5, but not for groups 3 and 4. Closeness centrality is stable for all groups except group 3. Therefore, strength and expected influence centralities will be interpreted for all groups, while betweenness is only interpreted for groups 1, 2 and 5, and closeness is interpreted for all groups except group 3.

The three most central nodes for each group for each centrality measure are shown in Table 4.11. Based on strength and expected influence centrality, in general the most central nodes are `impcntr` (allow immigrants from poor countries outside Europe), `eimpcnt` (allow immigrants from poor countries inside Europe), `imdfetn` (allow immigrants with a different ethnicity), `imsmetn` (allow immigrants with the same ethnicity), `qfimwsk` (qualification for immigration: work skills) and `imwbcnt` (immigrants make country better/worse place to live). In terms of betweenness centrality, groups 1, 2 and 5 each have different nodes in the top three. For group 1, `qfimchr` (qualification for immigration: be Christian) and `qfimwht` (qualification for immigration: be white) lead the top three. For group 2, it's `imbgeco` (immigrants bad/good for the economy) and `imbleco` (taxes and services: immigrants take out more/less than they put in). For group 5, the variables `imwbcnt` (immigrants make country better/worse place to live) and `imwbcrm` (immigrants make crime situation worse/better) score the highest on betweenness centrality. On closeness centrality, groups 2 and 5 share the top three: `imwbcnt` (immigrants make country better/worse place to live), `imbgeco` (immigrants bad/good for the economy) and `imtcjob` (immigrants take/create jobs). Groups 1 and 4 share the varibles `qfimchr` (qualification for immigration: be Christian) and `imdfetn` (allow immigrants with a different ethnicity) in the top three of closeness centrality.
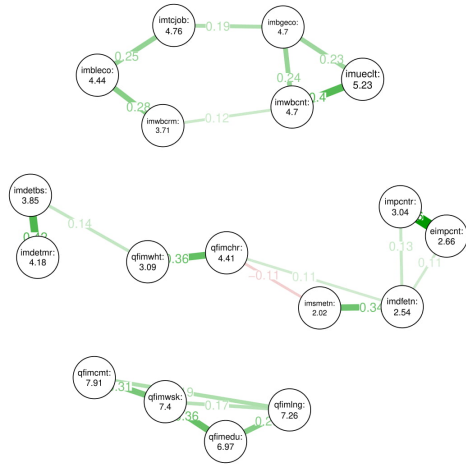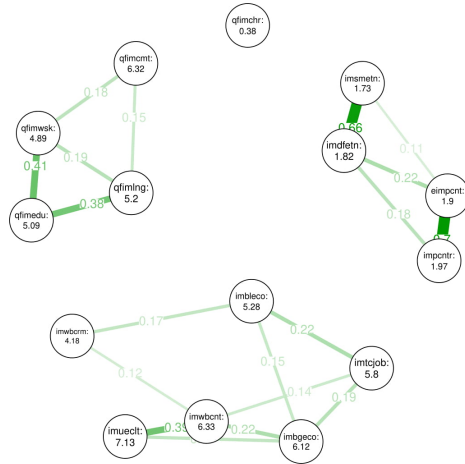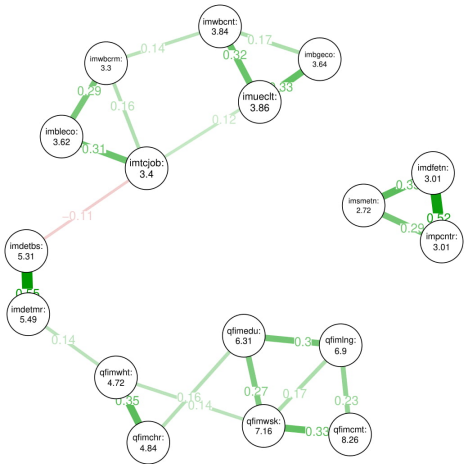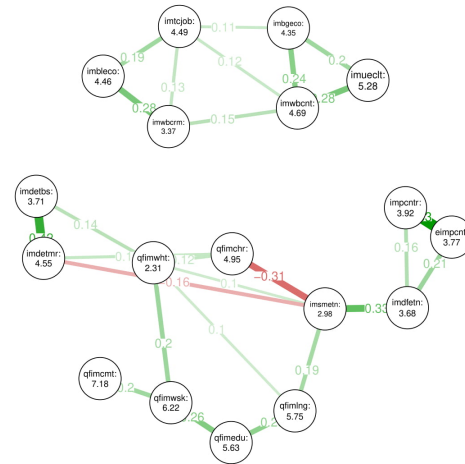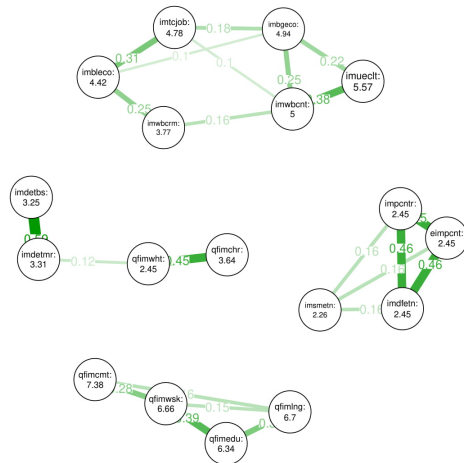
Figure 4.31: CISs for five subgroups estimated in the ESS data using the proposed GMM/GGM method. Nodes represent variables, edge weights represent partial correlations between variables.
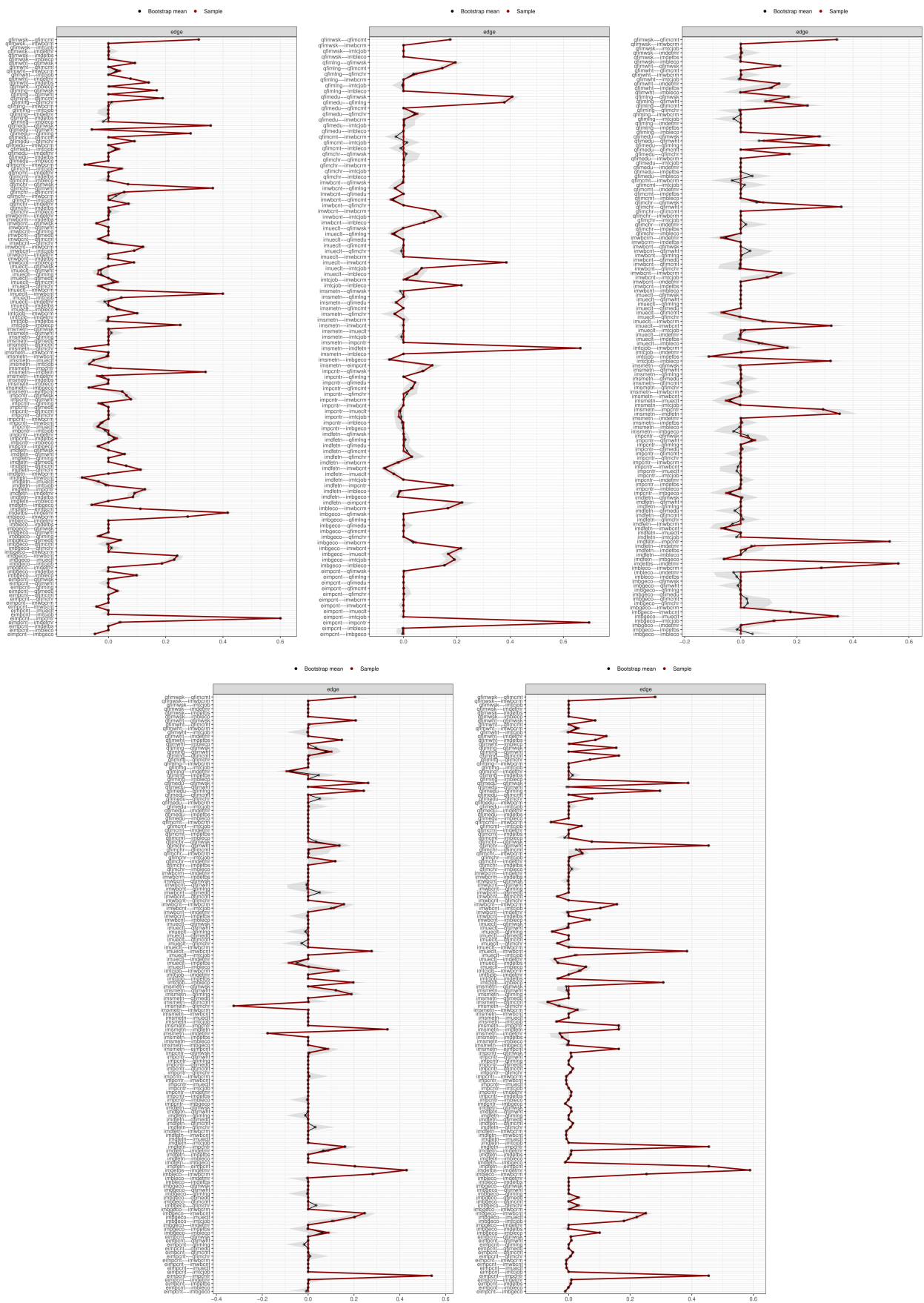
Figure 4.32: Stability analysis plots of the networks in Figure 4.31. From left to right and top to bottom: group 1, group 2, group 3, group 4, group 5.
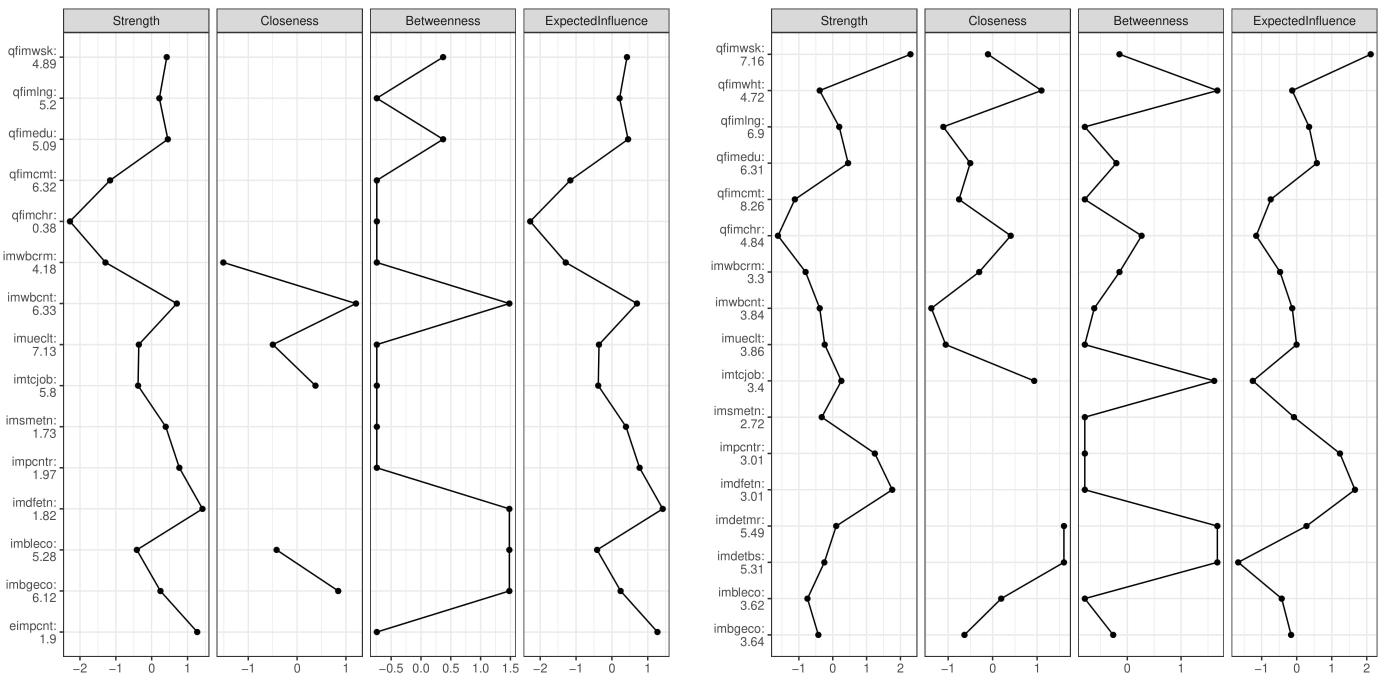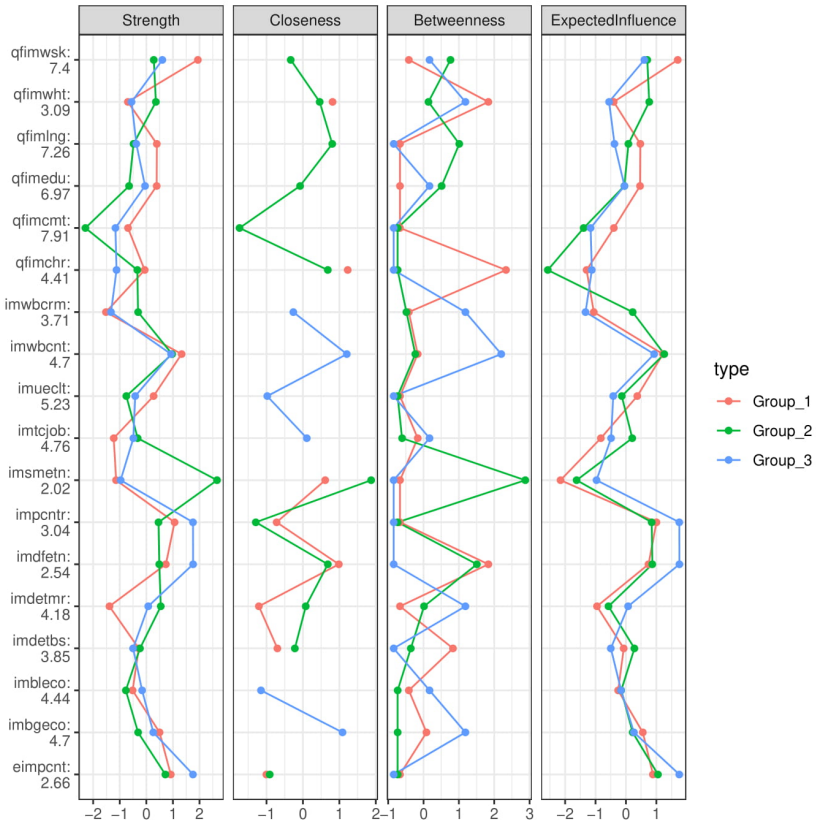
Figure 4.33: Standardised centrality plots for subgroups in the three-group model of the ESS data. Top panel: groups 1, 4 & 5. Bottom panel, left to right: group 2, group 3.

| Group | Strength | Expected Influence | Betweenness | Closeness |
|---|---|---|---|---|
| 1 | qfimwsk | qfimwsk | qfimchr | qfimchr |
| | imwbcnt | imwbcnt | qfimwht | imdfetn |
| | impcntr | impcntr | imdfetn | qfimwht |
| 2 | imdfetn | imdfetn | imbgeco | imwbcnt |
| | eimpcnt | eimpcnt | imbleco | imbgeco |
| | impcntr | impcntr | imdfetn | imtcjob |
| 3 | qfimwsk | qfimwsk | | |
| | imdfetn | imdfetn | NA | NA |
| | impcntr | impcntr | | |
| 4 | imsmetn | imwbcnt | | imsmetn |
| | imwbcnt | eimpcnt | NA | qfimchr |
| | eimpcnt | impcntr | | imdfetn |
| 5 | impcntr | eimpcnt | imwbcnt | imwbcnt |
| | imdfetn | imdfetn | imwbcrm | imbgeco |
| | eimpcnt | impcntr | imdetmr | imtcjob |

Table 4.11: Three most central nodes per group per centrality measure, for networks of subgroups in the five-group model.
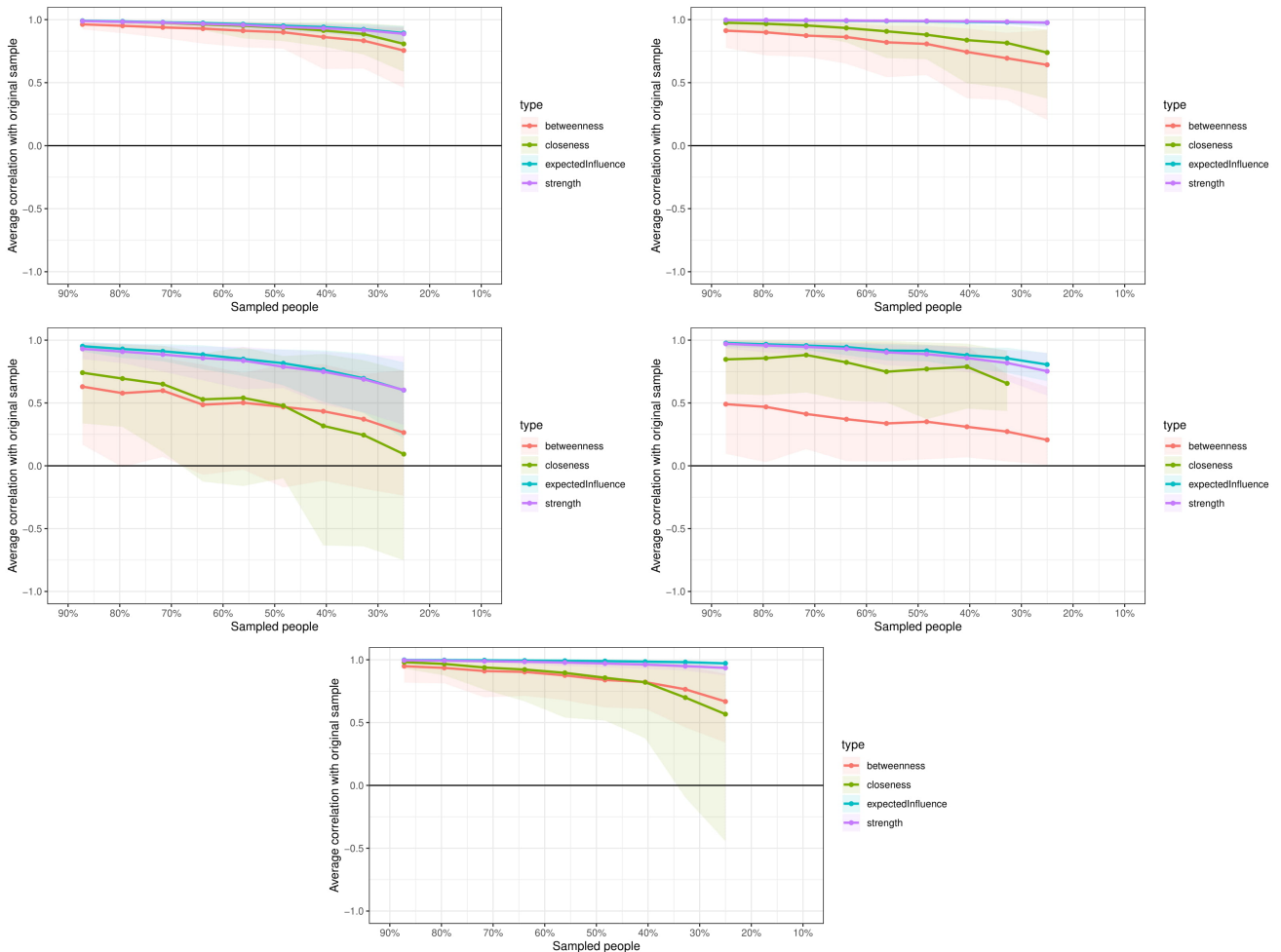


Figure 4.34: Centrality stability plots for subgroups in the three-group model of the SMD data. From left to right, from top to bottom: group 1, group 2, group 3, group 4, group 5.

**Summary.** Intuitively, the group 1 network is very similar in structure to group 1 in the three-group model (labeled 'indifferent), and to group 5 in this model. The main difference is the bridging role of `qfimchr` (qualification for immigration: be Christian). It is also the second largest group, after group 5. Another distinguishing feature is the relatively large discrepancy (compared to all other groups) between allowing immigrants from poor countries within Europe (`eimpcnt`) and outside Europe (`impcntr`). Immigrants from within Europe are slightly less welcomed by this subgroup. In addition, variables measuring importance of qualification criteria have slightly higher mean values than the comparison groups. This leads to an intuitive labelling of a *'euro-sceptic'* subgroup. This group is largely indifferent, but has a slightly higher bar for qualification and a slightly lower appreciation of European immigrants.

Intuitively, the group 2 network is very similar to the group 3 network in the three-group model (labelled 'favourable'). It also has removed the variable `qfimwht` (qualification for immigration: be white) from the model because all values were zero. It also has the lowest mean values for items tapping importance of qualification criteria and opposition towards immigrants, as well as the highest mean values for items measuring opinions of benefits of accepting immigrants. Not surprisingly, the structure of the network corresponds to these three categories. Therefore, this network can be labelled as *'favourable'*, corresponding to the group 3 network from the three-group model.

Intuitively, the network for group 3 structure suggests an *'anti-immigration'* subgroup. This group has the highest mean values on all variables measuring opposition towards immigrants, as well as all variables measuring importance of qualification criteria. In addition, it has the lowest mean values on all variables measuring opinions of beneficial effects of accepting immigrants. Especially interesting are the items tapping importance of race (`qfimwht`) and religion (`qfimchr`), which are highly correlated. Overall, this group is the most reluctant to accept immigrants of any kind.

Intuitively, the network structure for group 4 shows some resemblance with group 2 from the three-group model (labeled 'opposed'). From its properties, it suggests a *'religiously motivated'* subgroup. This is evidenced by the very central role and high mean value of the variable `qfimchr` (qualification for immigration: be Christian). This variable strongly influences whether immigrants from poor countries outside Europe would be accepted, as indicated by the negative partial correlation between those variables. In addition, the variable measuring opposition towards a close relative marrying an immigrant with a different ethnicity (`imdetmr`) has a much higher mean value than the variable measuring opposition towards having an immigrant with a different ethnicity as your boss (`imdetbs`). Interestingly, whether immigrants are white (`qfimwht`) has less impact (i.e. lower mean value and lower centrality). Furthermore, other variables about qualification for immigration have relatively low means compared to group 2 from the three-group model or group 3 from this model.

Finally, group 5 constitutes the largest group in this model, showing many similarities with group 1 from the three-group model (labelled 'indifferent'). As in the 'indifferent' group, this group shows no preference for immigrants from within or outside Europe, and does not have a central role for either race or religion. Furthermore, the variables measuring importance of qualification criteria are relatively high, but in a separate clique of the network, not impacted by other variables. The same holds for variables measuring opinions of benefits of immigrants. Therefore, this group is assumed to be the *'indifferent'* subgroup, corresponding to group 1 from the three-group model.

# Discussion

In this section, results and limitations of this thesis are discussed, after which suggestions for future research are made. Specifically, we discuss implications for the GGM/GMM method and evaluate its performance both on artificial and real world data.

## Conclusions

Results described in the previous section are discussed below. Specifically, the following paragraphs provide conclusions and inferences that can be drawn from the results. The section is divided according to the results. First, we discuss benchmarking results. Second, we discuss results of the Social Media Disorder data. Third, we discuss results of the Public Opinion of Immigration and Refugees data.

### Benchmarking

The benchmarking tests aimed to test the impact of a number of properties of the data on the performance of the method. Performance was evaluated as the ability to accurately retrieve the conditional independence structure of the groups in the data by comparing the estimated network structure to the true network structure. Specifically, three properties of the data were investigated: the degree of structural overlap between the groups, the sample size, and the presence of unbalanced group sizes. Furthermore, for all conditions, a full sample network was drawn. This network represents the data if no clustering is performed: the sample size is thus $N_{full} = \sum_k N_k$. The relation between the group structures and the full sample structure is also reported.

First, the degree of structural overlap did not limit the performance of the method. For all conditions, the true conditional independence structures were correctly estimated. This implies that the method is suitable for the detection of subgroups, given that these groups have heterogeneous conditional independence structures (CIS). Even if the structures of the groups have some variables in common (i.e. overlapping), this does not hinder estimation. In addition, with increasing structural overlap, the edge weight stability remained similar across conditions. This indicates a stable basis for interpretation, using centrality indices and stability measures. The full sample network is unable to represent all groups in a single graph, especially if there is structural overlap between the groups. It tends to show those edges that are shared by two or more groups, and may serve as an indicator of edges that are important for more than one group. Nonetheless, the full sample network cannot show all group structures, unless the the groups have no structural overlap.

Second, the sample size of the full sample is a strong limiting factor on the performance of the method. For all conditions with a group size lower than 500, we see some degree of error. For conditions with group sizes of 300 or less, these errors become severe, leading to empty networks being estimated for at least one of the groups. Furthermore, network stability decreased with decreasing sample size. This poses limitations on the use of centrality indices and stability measures in the interpretation of the resulting networks. Structures of small groups may still be estimated by lowering the hyperparameter $\gamma$, which controls the degree of regularisation. With a lower value of $\gamma$, more edges will be included in the graphs. However, this comes at the cost of precision and stability, further limiting interpretation of the structure. It should be noted that the stability of the *true* networks in conditions sample sizes below 300 was also low. Therefore, it seems that small group sizes impact the stability of the true data, which in turn clouds estimation. As a consequence, we suggest using this method when the expected group size is at minimum 400, and preferably 500 or larger. The full sample network becomes more sparse with decreasing sample size. For all conditions, the strongest edges in this network are those that are shared between groups. Therefore, the full sample network is indicative of edges that are shared between groups, even when sample size is low.

Third, the presence of unbalanced group sizes only impacts performance for the smallest group. Group sizes allocated to each group by the GGM/GMM method somewhat reflect the true group sizes. As an example, consider Figure 4.12, condition $N = \{1000, 1000, 500\}$, where the smallest true group (group 3, $N = 500$) is also the smallest estimated group (group 3, $N = 596$). Furthermore, it does not seem that the presence of one small group skews the estimation results of large groups. However, it does seem that

larger groups can 'impose' some of their properties on smaller groups (e.g. Figure 4.12, condition $N = \{1000, 1000, 500\}$, Group 3). This can be explained by the mixture model assigning (many) members of a large group to a small group, causing the graphical model to try and reconcile the two competing structures in one graph. In addition, if there is more than one small group, the method may not have enough data points to separate the two distributions. As a consequence, the method either estimates one empty network and one network that contains properties of both groups, or it estimates two networks that may share some properties of both groups. The networks then become confounded.

The network of the full sample is strongly biased towards the largest group. As an example, see figure 4.12, condition $N = \{1000, 1000, 500\}$, full sample network. Here, the network for the full sample shows clearly the structures of the two largest groups, but does not show any edges unique to the smallest group. Using the GGM/GMM method, however, the structure of the smallest group is retrieved with almost perfect accuracy.

In sum, a number of conclusions can be drawn regarding the performance of the GGM/GMM method. First, performance of the method (i.e. the ability to accurately estimate subgroup CISs) is highly dependent on group sizes. Specifically, for any group with a size smaller than 400, the results will be unstable and will deviate more from any true underlying structure.

Second, if a single network is drawn for all groups (i.e. the full sample network), it reflects the underlying subgroups in two ways. First, it is biased towards the largest subgroup(s): the emerging structure will look most like the structure of the largest subgroup(s) in the data. Structural elements of smaller groups are more likely to be 'outweighed' and excluded from the network. Second, it is biased towards shared edges, i.e. edges that are present in the networks of more than one subgroup. This property can help identify edges present in most or all of the underlying subgroups.

Third, centrality scores, edge weight stability and centrality stability that aid interpretation are less informative for graphs with low sample sizes, due to the fact that graphs with low sample sizes are unstable. Therefore, interpretation of networks estimated from a small sample size should be done with additional care.

# Social Media Disorder

Application of the outlined GMM/GGM method produces $k$ conditional independence structures indicating possible heterogeneous subgroups in the data. While the method does not test whether the subgroups produced by the methods are indeed empirically valid, it does allow building theory using this data-driven exploration. In this section, we discuss to what extent the results fits with existing social media disorder (SMD) theory.

## Three-group model

The networks resulting from the three-group model show marginally different structures. All groups have a strong, common structure that is supported by the full sample network. Specifically, the ADHD variables (attention problems, impulsivity and hyperactivity) and the quality of life measures (life satisfaction, self-esteem and physical self-esteem) form closely connected cliques in all groups. Interestingly, the cliques are connected through depressive symptoms. This observation is in line with known comorbidity between ADHD and depression (Yen, Ko, Yen, Wu, & Yang, 2007).

For the three-group model, we expected to see differences between groups based on the three levels of social media disorder: normal, risky, and problematic. However, differences observed between the three groups do not support this hypothesis. In fact, the role of social media disorder score seems very limited for all groups. Instead, observed differences seem to be a reflection of differences in quality of life. Group 2 (intuitively labeled 'insecure'), for example, shows lowest mean scores for quality of life measures, and highest mean scores for depressive symptoms and ADHD, while the opposite is true for group 3 (intuitively labeled 'happy & healthy'). Furthermore, group 1 (intuitively labeled 'average') shows no striking or unique structure that indicates how it differs from the other groups.

Although the results are not in line with the hypothesis about this model, the estimated subgroups and networks still show some interesting properties that may indicate potential avenues for further research of social media disorder. Specifically, the structure of the 'insecure' subgroup suggests a subgroup of adolescents that have a unique, negative correlation between restrictive parental rules and quality of communication with parents. van Den Eijnden et al. (2010) show that independently, quality of communication is a preventive factor of social media disorder, while restrictive parental rules are a risk factor of social media disorder.

In combination with the low quality of life measures of this group, one hypothesis may be that this group represents adolescents whose parents employ a authoritative parenting style, characterised by restrictive rules and low communication (Gray & Steinberg, 1999). This hypothesis would also explain the relatively high mean value of social media disorder in this group. In this way, the presented method provides data-driven insights that can lead to new hypotheses for further empirical investigation.

### Five-group model

The networks of the five-group model show more differences in structure and group sizes compared to the three-group model. However, all groups have relatively small group sizes, indicating instability. Therefore, results for this model should be interpreted with due care. Same as in the three-group model, ADHD symptoms (attention problems, impulsivity and hyperactivity) and quality of life measures (life satisfaction, self-esteem and physical self-esteem) form cliques in all groups.

The five-group model was expected to further break down the groups of the three-group model. Given the method's sensitivity to sample sizes, it is most likely that the largest group of the three-group model (i.e. the 'average' group) is broken up. Results suggest that, to some degree, this is indeed the case. Some informative features of the three-group model were carried over (such as the 'insecure' group), but due to the unbalanced nature of the data (further discussed under Limitations), and the relatively low sample sizes of all groups, the five-group model does not provide new, clearly interpretable subgroups compared to the three-group model.

# Public opinion of immigration and refugees

### Three-group model

At first glance, the three networks output by the three-group network seem to have quite different structures. One striking observation is the removal of three variables from the network of group 3, and another is the strongly connected structure of group 2 (see Figure 4.26).

In general, the high connectivity of the group 2 network and the varying mean values of different variables suggest an ambiguous structure. Previous work on anti-immigrant attitudes has revealed high prevalence of discrepant or contradicting opinions within anti-immigration attitudes (e.g. Katwala, Ballinger, & Rhodes, 2014; Pratto & Lemieux, 2001). An example would be acknowledgement of the economic value of immigrant labour, while also feeling that immigrants make the country a worse place to live (H. Dempster & Hargrave, 2017). Some unique structural properties of this group could be attributed to prejudicial attitudes, such as racial discrimination or euro-skepticism. However, this is impossible to determine from the conditional independence structure (and the mean values) alone.

Furthermore, the existence of a 'favourable' group has been reported before (for example, Ford and Lowles (2016) define a group of 'confident multiculturals' within the UK public). The final network structure of this group is not informative in itself. However, the exclusion of three variables that tap bias against immigrants is a significant structural difference.

Finally, group 1 (intuitively labelled 'indifferent') seems to represent a majority of Europeans. The majority 'ambiguous middle' group, reported by H. Dempster and Hargrave (2017), seems to be less ambiguous and more indifferent. The network for this group does not suggest any conflicting opinions, and does not suggest favour or opposition towards immigrants.

In sum, the network structures found in the three-group model do seem to represent the expected groups. However, ambiguity and discrepant opinions seem to be reserved mostly for the 'opposed' subgroup.

### Five-group model

The networks of the five-group model seem to further break down the subgroups found in the three-group model. Interestingly, the main themes of the groups in the three-group model are clearly carried over to the five-group model. First, the largest groups in both models (i.e. group 1 in the three-group model, and the group 5 in the five-group model) have nearly identical structures. Second, both models have clear pro-immigration and anti-immigration subgroups. The five-group model further breaks down public opinion and suggests smaller, more niche attitude profiles, such as a group that emphasises Christianity and a group that is Euro-sceptic. Such ideological distinctions are supported by previous work on attitudinal segmentation of public opinion of immigration in Europe (Green, 2007), and are also reflected in the European parliament (Lahav, 1997). The network structures for these groups may help identify values important to specific ideologies.

The results of the proposed method seem in line with existing literature on attitudinal segmentation. This work adds a way of exploring the structural relations between related topics for each of these categories, facilitating understanding of an ultimately complex attitude.

# Limitations

The conclusions presented in this thesis are a result of an ambitious project to combine unsupervised learning and structural modelling methods, applying the method to a variety of real and artificial test data. Due to the broad nature of the approach, the method can have many different parameter configurations, a few of which have been tested. However, this process was not exhaustive and was limited by a variety of process- and content-related factors. These limitations are discussed in this section.

## Method

While the method shows promise for the identification of subgroups based on conditional independence structure, it has a number of limitations. First, the current implementation of the method lacks an algorithmic way to select the value of $k$, the number of groups. Such methods do exist for Gaussian Mixture Modelling, for example greedy learning (Verbeek et al., 2003), or model selection criteria (Figueiredo & Jain, 2002). Such methods would increase the applicability of the method to datasets where there are no theoretical expectations for how many probable subgroups the data contains.

Second, due to time and space constraints, the benchmark tests do not include all possible parameter variations. For example, none of the internal parameters of the method (such as the EM parameters (log-likelihood threshold, number of iterations and minimum cluster size), or the model selection method (EBIC, BIC, cross-validation)) have been varied in the benchmarking tests, because these parameters have seen extensive testing in other works and common practice parameter values are available. In addition, not all possible properties of the data have been tested, such as combinations of the conditions in this work (e.g. sample size × structural overlap), or the type of data (i.e. ordinal vs. ratio data). Furthermore, some of the tests in this work could be expanded further (e.g. including negative weights in conditions with structural overlap).

Third, the conditional independence structures output by the model are a computational way to express structural heterogeneity of the data. Whether the resulting structures are theoretically feasible or useful in practice, needs to be evaluated by specialists in the field. In fact, the method can only provide a first step in finding theoretically sound subtypes. Therefore, the results should not be interpreted as definitive subtypes of the construct. Instead, they should be interpreted as indicators of the heterogeneity, which can be used to further investigate and discriminate potential subtypes.

## Social Media Disorder & Public opinion of immigration and refugees

Limitations for the results of the empirical cases lie in the data itself. The SMD data was stratified into three categories, based on the degree of SMD. However, the distribution of values for social media disorder (SMD) score is very skewed: the group with the highest level of SMD has a sample size of less than 50, making it too small to be properly detected by the GGM/GMM method. As a consequence, the majority of the participants can be considered 'healthy', which biases the results. Specifically, the groups identified in both models are most likely different subtypes of healthy participants. This conclusion is further supported by group 5 in the five-group model, which constitutes over one third of the sample, but has a SMD score of zero for the entire group. This limitation can be remedied by performing the same analyses on a large sample of data from the high SMD population. However, this data is not available at this time.

For the data on public opinion of immigration and refugees, data was retrieved from the European Social Survey. However, many potentially informative variables in this data were not used. For example, the effects of individual demographics (i.e. nationality, age, a.o.) are not taken into account explicitly. In addition, it is impossible to exclude any systematic errors or biases in the data, beyond what is reported by the authors of the ESS.

The main limitation for results in both domains is the fact that the number of subgroups to be identified by the method was set manually. In the SMD case, the three-group model was expected to capture the three categories of SMD. However, due to the highly unbalanced nature of these groups, this proved difficult. The five-group model served as a contrast to compare with the three-group model, but the choice for $k = 5$

is rather arbitrary. As a consequence, a meaningful interpretation of subgroups in the five-group model becomes difficult and arbitrary. This limitation can be remedied by providing a means of selecting $k$, as stated before. In the public opinion case, the three-group and five-group models serve as a nice comparison with the SMD data. However, it cannot be ignored that a model with a different number of subgroups may be more explanatory, and may be a better fit to model subtypes within this dataset.

# Future research

This work provides a novel application of data-driven, computational methods to topics that traditionally fall under the experimental approach of social sciences. Building on this work, there are a number of avenues that interested researchers may want to pursue. First is the development of a $k$-selection method. Expanding the current methodology to perform automatic model selection (and limiting the resulting computational cost) will improve the identification power of this method. For example, Verbeek et al. (2003) introduce a greedy learning algorithm for learning the optimal amount of mixtures for a GMM, and Figueiredo and Jain (2002) suggest an alternative algorithm that performs simultaneous selection of components and parameter fitting.

In the same vein, since the expectation-maximisation algorithm is known to be sensitive to initialisation values, the method could benefit from alternatives to the EM algorithm. For example, Pernkopf and Bouchaffra (2005) describe a novel algorithm combining genetic algorithms with the traditional EM algorithm in order to overcome the limitations of EM. Similarly, Benavent, Ruiz, and Sáez (2009) suggest an entropy-based EM algorithm and M.-S. Yang, Lai, and Lin (2012) develop a robust EM algorithm that overcomes limitations of traditional EM.

Second, the dual nature of the GMM/GGM method implies a complicated method that requires extensive testing. A start has been made in this thesis, but as discussed before, there are a number of more internal parameters that have not been extensively tested yet. These include internal parameters governing the EM algorithm (i.e. the log-likelihood difference threshold, minimum cluster size and number of iterations), as well as the hyperparameter $\gamma$ that determines model sparsity. In addition, more testing of data properties will further inform of the strengths and weaknesses of the GGM/GMM method. Of particular interest is the influence of the type of data (i.e. ordinal, normally distributed, or mixed type) on the performance of the model, as the type of data changes the properties of the covariance matrix on which the method depends. Testing this is especially important for social sciences, where differing measurement scales are commonplace.

Third, the method can be extended with additional functionality, such as feature selection. While this was not a topic of focus in the current work, the framework of the GMM/GGM model does allow for such functionality. For example, Zhou et al. (2009) apply regularisation penalties to means vectors within the GMM, thereby performing simultaneous model and feature selection. The addition of feature selection to the methodology is of particular interest in cases where the number of variables is large.

In sum, this thesis addressed the identification and interpretation of subgroups in data. To this end, clustering and structural estimation methods were combined in order to add explainability to the clustering results. The resulting GGM/GMM method distinguishes clusters in data by their conditional independence structure, which can be visualised in the form of a network. The method yields a network for each cluster, that can be analysed using network analytic tools. Clusters can then be compared to each other based on network structure and network analysis results. Comparing the networks can generate new insights and hypotheses about the studied construct.

The GGM/GMM method adds to existing works by using the conditional independence structure both as a feature in clustering and as a tool for interpretation of the results, addressing challenges in the interpretation of clustering results. By addressing these challenges, the method allows generation of new hypotheses about structurally different types of the construct. Such identification of subtypes is especially meaningful in domains that deal with heterogeneity in the data, such as social sciences.

# References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR* (4th ed., text rev. ed.). Washington, DC: Autor.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed. ed.). Washington, DC: Autor.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The'k'in k-fold cross validation. In *Esann.*

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 post-traumatic stress disorder symptoms and correlates in US military veterans. *Journal of anxiety disorders*, *45*, 49-59.

Baioletti, M., Busanello, G., & Vantaggi, B. (2009). Conditional independence structure and its closure: inferential rules and algorithms. *International Journal of Approximate Reasoning*, *50*(7), 1097–1114.

Bates, D., & Maechler, M. (2019). Matrix: Sparse and dense matrix classes and methods [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=Matrix` (R package version 1.2-18)

Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M., ... Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological medicine*, *46*(16), 3359–3369.

Bell, P., & King, S. (2007). Sparse gaussian graphical models for speech recognition.

Benavent, A. P., Ruiz, F. E., & Sáez, J. M. (2009). Learning gaussian mixture models with entropy-based criteria. *IEEE Transactions on Neural Networks*, *20*(11), 1756–1771.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Benjamins, J. S., Migliorati, F., Dekker, K., Wassing, R., Moens, S., Blanken, T. F., ... Van Someren, E. J. (2017). Insomnia heterogeneity: characteristics to consider for data-driven multivariate subtyping. *Sleep Medicine Reviews*, *36*, 71–81.

Bora, E., Aydın, A., Saraç, T., Kadak, M. T., & Köse, S. (2017). Heterogeneity of subclinical autistic traits among parents of children with autism spectrum disorder: Identifying the broader autism phenotype with a data-driven method. *Autism Research*, *10*(2), 321–326.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, 91–121.

Buhmann, J. (1995). Data clustering and learning. *The handbook of brain theory and neural networks*, 278–281.

Bukh, J., Miller, R. H., & Purcell, R. H. (1995). Genetic heterogeneity of hepatitis c virus: quasispecies and genotypes. In *Seminars in liver disease* (Vol. 15, pp. 41–63).

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759-771.

Chen, J., & Kalbfleisch, J. (1996). Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, *24*(2), 167–175.

Chu, J.-h., Weiss, S. T., Carey, V. J., & Raby, B. A. (2009). A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC systems biology*, *3*(1), 55.

Constantinopoulos, C., Titsias, M. K., & Likas, A. (2006). Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(6), 1013–1018.

Contreras, A., Nieto, I., Valiente, C., Espinosa, R., & Vazquez, C. (2019). The study of psychopathology from the network analysis perspective: a systematic review. *Psychotherapy and psychosomatics*, *88*(2), 71–83.

Corduneanu, A., & Bishop, C. M. (2001). Variational bayesian model selection for mixture distributions. In *Artificial intelligence and statistics* (Vol. 2001, pp. 27–34).

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13-29.

Costantini, G., & Perugini, M. (2016). The network of conscientiousness. *Journal of Research in Personality*, *65*, 68–88.

Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. a tutorial on recent developments in network psychometrics. *Personality and Individual Differences*, *136*, 68–78.

Cramer, A. O., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414–431.

Currie, S. (2016). Reflecting on Brexit: Migration myths and what comes next for EU migrants in the UK? *Journal of Social Welfare and Family Law*, *38*(3), 337-342.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, *123*(1), 2-22.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2017). Network analysis on attitudes: A tutorial. *Social psychological and personality science*, *8*(5), 528-537.

d'Aspremont, A., Banerjee, O., & El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, *30*(1), 56–66.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*(1), 157-175.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dempster, H., & Hargrave, K. (2017). *Understanding public attitudes towards refugees and migrants.* London: Overseas Development Institute & Chatham House.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, *49*(1), 71–75.

Dijkstra, E. W., et al. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, *90*(1), 196–212.

Doty, D. H., & Glick, W. H. (1994). Typologies as a unique form of theory building: Toward improved understanding and modeling. *Academy of management review*, *19*(2), 230–251.

Drton, M., & Perlman, M. D. (2004). Model selection for gaussian concentration graphs. *Biometrika*, *91*(3), 591–602. Retrieved from `http://www.jstor.org/stable/20441125`

Epskamp, S. (2016a). Brief report on estimating regularized gaussian networks from continuous and ordinal data. *arXiv preprint arXiv:1606.05771*.

Epskamp, S. (2016b). Regularized gaussian psychological networks: Brief report on the performance of extended bic model selection.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195-212.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617-634.

Epskamp, S., Maris, G. K., Waldorp, L. J., & Borsboom, D. (2016). Network psychometrics. *arXiv preprint arXiv:1609.02818*.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927.

ESS round 7. (2014). European social survey round 7 data (2014), data file. doi: 10.21338/NSD-ESS7-2014

Farnoush, R., & Zar, P. B. (2008). Image segmentation using gaussian mixture model.

Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(3), 381–396.

Firoozbakht, F., Rezaeian, I., Porter, L., & Rueda, L. (2014). Breast cancer subtype identification using machine learning techniques. In *2014 ieee 4th international conference on computational advances in bio and medical sciences (iccabs)* (pp. 1–2).

Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, *108*(3), 479–485.

Ford, R., & Lowles, N. (2016). *Fear & hope 2016: Race, Faith and Belonging in Todayś England.* London: HOPE not hate.

Forrest, L. N., Jones, P. J., Ortiz, S. N., & Smith, A. R. (2018). Core psychopathology in anorexia nervosa and bulimia nervosa: A network analysis. *International Journal of Eating Disorders*, *51*(7), 668–679.

Foygel, R., & Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems* (pp. 604–612).

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314-320.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432-441.

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129-1164.

Giudici, P., & Spelta, A. (2016). Graphical network models for international financial flows. *Journal of Business & Economic Statistics*, *34*(1), 128–138.

Gray, M. R., & Steinberg, L. (1999). Unpacking authoritative parenting: Reassessing a multidimensional construct. *Journal of Marriage and the Family*, 574–587.

Green, E. G. (2007). Guarding the gates of europe: A typological analysis of immigration attitudes across 21 countries. *International Journal of Psychology*, *42*(6), 365–379.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 2125–2126).

Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *JSTOR: Applied Statistics*, *28*(1), 100–108.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1–12.

Hevey, D. (2018). Network analysis: a brief overview and tutorial. *Health Psychology and Behavioral Medicine*, *6*(1), 301–328.

Hill, S. M., & Mukherjee, S. (2013). Network-based clustering with mixtures of l1-penalized gaussian graphical models: an empirical investigation. *arXiv preprint arXiv:1301.2194*.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Isacson, A., Meyer, M., & Morales, G. (2014). Mexicoś other border: Security, migration, and the humanitarian crisis at the line with Central America. *Washington Office on Latin America*, 2-44.

Ising, E. (1925, February). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, *31*(1), 253-258. doi: 10.1007/BF02980577

James, L. F., Priebe, C. E., & Marchette, D. J. (2001). Consistent estimation of mixture complexity. *Annals of Statistics*, 1281–1296.

Jones, P. J., Mair, P., & McNally, R. J. (2018). Visualizing psychological networks: A tutorial in r. *Frontiers in Psychology*, *9*, 1742.

Kandel, D. B., & Davies, M. (1982). Epidemiology of depressive mood in adolescents: An empirical study. *Archives of general psychiatry*, *39*(10), 1205–1212.

Katwala, S., Ballinger, S., & Rhodes, M. (2014). How to talk about immigration. *London: British Future*.

Kearns, D. B., & Losick, R. (2005). Cell population heterogeneity during growth of bacillus subtilis. *Genes & development*, *19*(24), 3083–3094.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 49–66.

Kim, S. C., & Kang, T. J. (2007). Texture classification and segmentation using wavelet packet frame and gaussian mixture model. *Pattern Recognition*, *40*(4), 1207–1221.

Knief, U., & Forstmeier, W. (2018). Violating the normality assumption may be the lesser of two evils. *bioRxiv*, 498931.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality indices. In *Network analysis* (pp. 16–61). Springer.

Krämer, N., Schäfer, J., & Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, *10*(1), 1-24.

Lahav, G. (1997). Ideological and party constraints on immigration attitudes in europe. *JCMS: Journal of Common Market Studies*, *35*(3), 377–406.

Lauritzen, S. L. (1996). *Graphical models.* Oxford, UK: Clarendon Press.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., ... others (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218.

Lee, J. W., Moen, E. L., Punshon, T., Hoen, A. G., Stewart, D., Li, H., ... Gui, J. (2019). An integrated gaussian graphical model to evaluate the impact of exposures on metabolic networks. *Computers in biology and medicine*, *114*, 103417.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 1350–1360.

Lim, T. C. (2018, April). *Clustering: k-means, k-means ++ and gganimate.* Retrieved 2018-04-19, from `https://theanlim.rbind.io/post/clustering-k-means-k-means-and-gganimate/`

Lombardo, M. V., Lai, M.-C., Auyeung, B., Holt, R. J., Allison, C., Smith, P., ... others (2016). Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Reports*, *6*(1), 1–15.

Ma, S., Gong, Q., & Bohnert, H. J. (2007). An arabidopsis gene network based on the graphical gaussian model. *Genome research*, *17*(11), 1614–1625.

Martin, P. L. (2017). Election of Donald Trump and migration. *Migration Letters*, *14*(1), 161-171.

McClellan, J., & King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, *141*(2), 210–217.

Meinshausen, N., & Bühlmann, P. (2006). Highdimensional graphs and variable selection with the lasso. *The annals of statistics*, *34*(3), 1436-1462.

Meredith, J. (1993). Theory building through conceptual methods. *International Journal of Operations & Production Management*.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology*, *2*, 80–203.

Mueller, S., & Rungie, C. (2009). Is there more information in best-worst choice data?: Using the attitude heterogeneity structure to identify consumer segments. *International Journal of Wine Business Research*, *21*(1), 24–40.

Müller, A. C. (2018, March). *Clustering and mixture models.* Retrieved 2018-03-26, from `https://amueller.github.io/COMS4995-s18/slides/aml-16-032118-clustering-and-mixture-models/#1`

Nandi, A., Beard, J. R., & Galea, S. (2009). Epidemiologic heterogeneity of common mood and anxiety disorders over the lifecourse in the general population: a systematic review. *BMC psychiatry*, *9*(1), 31.

Nigg, J. T., Blaskey, L. G., Huang-Pollock, C. L., & Rappley, M. D. (2002). Neuropsychological executive functions and dsm-iv adhd subtypes. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(1), 59–66.

Nigg, J. T., Willcutt, E. G., Doyle, A. E., & Sonuga-Barke, E. J. (2005). Causal heterogeneity in attention-deficit/hyperactivity disorder: do we need neuropsychologically impaired subtypes? *Biological psychiatry*, *57*(11), 1224–1230.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*(3), 337–347.

Oquendo, M., Baca-Garcia, E., Artes-Rodriguez, A., Perez-Cruz, F., Galfalvy, H., Blasco-Fontecilla, H., ... Duan, N. (2012). Machine learning and data mining: strategies for hypothesis generation. *Molecular psychiatry*, *17*(10), 956–959.

Parnini, S. N., Othman, M. R., & Ghazali, A. S. (2013). The Rohingya refugee crisis and Bangladesh-Myanmar relations. *Asian and Pacific Migration Journal*, *22*(1), 133-146.

Pernkopf, F., & Bouchaffra, D. (2005). Genetic-based em algorithm for learning gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1344–1348.

Poddar, S. (2016). European migrant crisis: Financial burden or economic opportunity? *Social Impact Research Experience (SIRE)*, *43*.

Pratto, F., & Lemieux, A. F. (2001). The psychological ambiguity of immigration and its implications for promoting immigration policy. *Journal of Social Issues*, *57*(3), 413–430.

Przybylski, A. K., Murayama, K., DeHaan, C. R., & Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*, *29*(4), 1841–1848.

Purpose. (2017). *Germany and France segmentation research study, in collaboration with Ipsos MORI and IFOP.* (Research directed by Stephen Hawkins and Tim Dixon)

Ramalho, T. (2014, April). *Quick introduction to gaussian mixture models with python.* Retrieved 2014-04-03, from `https://tmramalho.github.io/blog/2014/04/03/quick-introduction-to-gaussian-mixture-models-with-python/`

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*.

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, *125*(6), 747.

Roeder, K., & Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, *92*(439), 894–902.

Rosenberg, M., Schooler, C., & Schoenbach, C. (1989). Self-esteem and adolescent problems: Modeling reciprocal effects. *American sociological review*, 1004–1018.

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

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, *4*(1).

Scholte, E., & Van der Ploeg, J. (1999). De ontwikkeling en toetsing van de adhd-vragenlijst. *Kind en adolescent*, *20*(2), 50–60.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, *88*(422), 486–494.

Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, *49*(267), 467–479.

Smith, K. E., Crosby, R. D., Wonderlich, S. A., Forbush, K. T., Mason, T. B., & Moessner, M. (2018). Network analysis: An innovative framework for understanding eating disorder psychopathology. *International Journal of Eating Disorders*, *51*(3), 214–222.

Steele, R. J., & Raftery, A. E. (2010). Performance of bayesian model selection criteria for gaussian mixture models. *Frontiers of statistical decision making and bayesian analysis*, *2*, 113–130.

Studeny, M. (2006). *Probabilistic conditional independence structures*. Springer Science & Business Media.

Thomaes, S., Stegge, H., Bushman, B. J., Olthof, T., & Denissen, J. (2008). Development and validation of the childhood narcissism scale. *Journal of personality assessment*, *90*(4), 382–391.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Toh, H., & Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics*, *18*(2), 287–297.

Treffers, D., Goedhart, A. W., Van den Bergh, B., Veerman, J., Ackaert, L., & De Rycke, L. (2002). Competentiebelevingsschaal voor adolescenten (cbsa): handleiding.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific reports*, *4*(5918), 1-10.

Van Den Eijnden, R., Koning, I., Doornwaard, S., Van Gurp, F., & Ter Bogt, T. (2018). The impact of heavy and disordered use of games and social media on adolescents' psychological, social, and school functioning. *Journal of behavioral addictions*, *7*(3), 697–706.

Van Den Eijnden, R. J., Lemmens, J. S., & Valkenburg, P. M. (2016). The social media disorder scale. *Computers in Human Behavior*, *61*, 478–487.

van Den Eijnden, R. J., Spijkerman, R., Vermulst, A. A., van Rooij, T. J., & Engels, R. C. (2010). Compulsive internet use among adolescents: Bidirectional parent–child relationships. *Journal of abnormal child psychology*, *38*(1), 77–89.

Verbeek, J. J., Vlassis, N., & Kröse, B. (2003). Efficient greedy learning of gaussian mixture models. *Neural computation*, *15*(2), 469–485.

Wardenaar, K. J., & de Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Medicine*, *11*(1), 201.

Wermuth, N., & Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society: Series B (methodological)*, *52*(1), 21–50.

Wille, A., & Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical applications in genetics and molecular biology*, *5*(1).

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., . . . others (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, *5*(11), R92.

Williams, D. R., & Rast, P. (2018). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, 1-15.

Wong, F., Carter, C. K., & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, *90*(4), 809–830.

Woo, M.-J., & Sriram, T. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, *101*(476), 1475–1486.

Yang, M.-H., & Ahuja, N. (1998). Gaussian mixture model for human skin color and its applications in image and video databases. In *Storage and retrieval for image and video databases vii* (Vol. 3656, pp. 458–466).

Yang, M.-S., Lai, C.-Y., & Lin, C.-Y. (2012). A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, *45*(11), 3950–3961.

Yen, J.-Y., Ko, C.-H., Yen, C.-F., Wu, H.-Y., & Yang, M.-J. (2007). The comorbid psychiatric symptoms of internet addiction: attention deficit and hyperactivity disorder (adhd), depression, social phobia, and hostility. *Journal of adolescent health*, *41*(1), 93–98.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, *94*(1), 19–35.

Zenetti, G., & Klapper, D. (2016). Advertising effects under consumer heterogeneity–the moderating role of brand experience, advertising recall and attitude. *Journal of Retailing*, *92*(3), 352–372.

Zhao, H., & Duan, Z.-H. (2019). Cancer genetic network inference using gaussian graphical models. *Bioinformatics and biology insights*, *13*, 1177932219839402.

Zhou, H., Pan, W., & Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic journal of statistics*, *3*, 1473.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th international conference on pattern recognition, 2004. icpr 2004.* (Vol. 2, pp. 28–31).

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.