Introducing SNAC: Sparse Network and Component model for integration of multi-source data

Pia Tio
Tilburg University
University of Amsterdam

Lourens J Waldorp
University of Amsterdam

Katrijn Van Deun
Tilburg University

Abstract

Gaussian graphical models (GGMs) are a popular method for analysing complex data by modelling the unique relationships between variables. Recently, a shift in interest has taken place from investigating relationships within a discipline (e.g. genetics) to estimating relationships between variables from various disciplines (e.g. how gene expression relates to cognitive performance). It is thus not surprising that there is an increasing need for analysing large, so-called *multi-source* datasets, each containing detailed information from many data sources on the same individuals. GGMs are a straightforward statistical candidate for estimating *unique cross-source relationships* from such network-oriented data. However, the multi-source nature of the data poses two challenges: First, different sources may inherently differ from one another, biasing the estimated relations. Second, GGMs are not cut out for separating cross-source relationships from all other, source-specific relationships. In this paper we propose adding a simultaneous-component-model as a pre-pocessing step to the GGM, the combination of which is suitable for estimating cross-source relationships from multi-source data. Compared to the graphical lasso (a commonly used GGM technique), this Sparse Network And Component (SNAC) model more accurately estimates the unique cross-source relationships from multi-source data. This holds in particular when the data contains more variables than observations ($p > n$). Neither differences in sparseness of the underlying component structure of the data nor in the relative dominance of the cross-source compared to source-specific relationships strongly affect the relationship estimates. Sparse Network And Component analysis, a hybrid component-graphical model, is a promising tool for modelling unique relationships between different data sources, thus providing insight in how various disciplines are connected to one another.

Background

Gaussian graphical models (GGMs) are a popular method for analysing complex data by modelling the unique relationships between variables (Koller & Friedman, 2009). Both the estimates and their visualisation as a network provide valuable insights in the underlying structure of the data. This holds true even for those datasets where detailed information is gathered on the same individuals, often resulting in datasets with more variables than observations ($p > n$; Schäfer et al., 2005; Krämer et al., 2009). Within the broad field of bioinformatics GGMs have been applied to various subdisciplines, including transcriptomics (Ingkasuwan et al., 2012; López-Kleine et al., 2013), genomics (Ma et al., 2007; Chu et al., 2011), ecotoxicogenomics (Villeneuve et al., 2007), and metabolomics (Krumsiek et al., 2011).

However, understanding the pathways from genotype or physiology to phenotype or behaviour requires more than estimating GGMs of each individual discipline; one also needs to know how these various isolated fields are connected to one another via *cross-source relationships* between variables from different data sources. Take the fields of genetics and human cognition (Davies et al., 2011). In their paper Johnson et al. (2015) identify gene co-expression networks associated with both healthy cognitive abilities, and cognitive and neurodevelopmental disorders. These findings can be further extended by adding information from additional sources, such as (functional) brain data, in order to further our understanding of human functioning, opening new treatment venues and increasing prediction accuracy for whom is at risk for developing said pathologies.

It is thus not surprising that emerging fields such as systems biology and network science emphasise the need for collecting and analysing large, so-called *multi-source* datasets, each containing detailed information from many data sources on the same individuals (Silverman & Loscalzo, 2012; Bartel et al., 2013). Luckily, with increasingly more sophisticated instruments and growing interdisciplinary cooperation, multi-source datasets become more common. Take for example the Avon Longitudinal Study of Parents and Children (Golding, 1990), which includes questionnaire data, fMRI scans, genetic information (SNPs), and physiological measurements available for the same set of individuals. However, availability of multi-source data alone is not enough to answer questions about cross-source relationships; appropriate statistical tools are needed too.

GGMs are a straightforward statistical candidate for analysing such network-oriented data. However, the multi-source nature of our data poses a challenge. In non-multi-source data, unique linear relationships can usually be estimated straightforwardly using partial correlations. However, if the data come from multiple sources, it is possible that groups of variables have different characteristics. For example, one group of variables may contain more noise than another group because of different measuring techniques or lower granularity; or variables within a group may be highly correlated to one another (e.g., positive correlations amongst cognitive variables) compared to other

The authors wish to thank Jeroen Vermunt and Denny Borsboom for their valuable contributions to earlier versions of the manuscript.

groups of variables (e.g., genetic information). Disregarding any of the inherent differences between sources and relationship-strength can lead to biased correlation estimates. Furthermore, it could also be the case that the cross-source relationships that we are interested in are weaker than the source-specific relationships. This problem is exacerbated by the fact that only few variables are relevant for the common mechanism. GGMs, in particular those adapted to deal with high-dimensional data ($p > n$), are well suited for estimating the strongest relationships amongst variables and thus less cut out for separating weaker cross-source relationships from all other, stronger source-specific relationships. Given that we already know that we want to estimate cross-source relationships of data from multiple sources, it makes sense to use statistical analyses that can disentangle relevant cross-source information from irrelevant source-specific information while incorporating the multi-source data structure.

In this paper we propose determining the set of variables from different sources that are most likely to be connected, the resulting network focusing on this subset of cross-source variables will be more accurate than without such a variable selection. We use a variant of sparse simultaneous component analysis to assess the subset of cross-source variables that are connected. In Section 2 we will describe GGM and the multi-source data structure in more detail, followed by introducing sparse simultaneous component analysis as a pre-processing step that enables GGMs to estimate cross-source relationships from multi-source data. Section 3 reports a simulation study investigating whether this Sparse Network And Component (SNAC) model outperforms regular GGMs. Lastly, in Section 4 we discuss current restrictions and possible improvements of SNAC.

Methods

First we present the graphical model as it appears in the literature, this is for data from a single source; second, we introduce the assumed data generating model for multi-source data and show how to isolate the cross-source relations; and third, we show how to apply the graphical model (for single source data) to these cross-source relations in order to obtain the desired network.

Graphical models

Graphical models (GMs) estimate conditional dependency relationships between variables using probability theory. Conditional dependence indicates that none of the remaining variables can explain away the relation between the two variables. As such conditional dependence relations can be said to be *unique* to the pair of variables. A GM can be visualised as a graph whose nodes and edges represent variables and conditional dependency relationships respectively. Many types of graphical models have been formulated; here we focus on Gaussian Graphical models (GGMs), which model undirected unique relationships in a multivariate Gaussian setting (Koller & Friedman, 2009).

Let $\mu = \mathbf{0}$ be a p-dimensional zero mean vector and $\mathbf{\Sigma}$ be a $p \times p$ positive definite covariance matrix. For a p-dimensional vector $\mathbf{x}$, the multivariate Gaussian density is

defined as

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\mathbf{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right], \quad (1)$$

with $|\mathbf{\Sigma}|$ the determinant of $\mathbf{\Sigma}$. Note that equation (1) is often expressed as $f(\mathbf{x}|\mu, \mathbf{\Sigma}) \sim \mathcal{N}_p(0, \mathbf{\Sigma})$. Under the assumption of normality a zero off-diagonal element in the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ is equivalent to the two corresponding variables being conditionally independent given all remaining variables, and a non-zero entry means conditional dependence (Lauritzen, 1996; Koller & Friedman, 2009).

Gaussian graphical models thus provide a mathematical and visual representation of unique dependency relationships between variables that is straightforward to interpret. These dependency relations can be estimated using maximum likelihood. However, this requires more observations than variables ($n > p$) otherwise it is not possible to obtain a positive definite covariance matrix (Bilodeau & Brenner, 1999). In multi-source data, it is very likely that there will be fewer observations than variables ($n < p$). A popular solution to deal with this situation is applying sparse modelling through an ℓ_1 penalty. This Least Absolute Shrinkage and Selection Operator (lasso) results in shrinkage of the parameters to zero with small values set exactly to zero (Tibshirani, 1996). The underlying assumption is that only a small number of all possible parameters is non-zero, i.e., there is a sparse solution including only the *relevant* relations and variables. One of the algorithms used to obtain such a lasso estimate of $\mathbf{\Sigma}^{-1}$ is the graphical lasso (Friedman et al., 2008). Let $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}^T$ be the empirical covariance matrix; the graphical lasso optimises the function

$$\log|\mathbf{\Sigma}^{-1}| + \mathrm{tr}\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} + \lambda\sum_{k\neq j}^{p}|(\mathbf{\Sigma}^{-1})_{jk}| \quad (2)$$

where $|(\mathbf{\Sigma}^{-1})_{jk}|$ is the absolute value of the jkth $(j, k = 1, \ldots, p)$ entry of $\mathbf{\Sigma}^{-1}$ and $\lambda \geq 0$ is a tuning parameter for the lasso penalty. Note that we do not use the diagonal elements in the penalty as this implies shrinking the variance (Bühlmann & Van De Geer, 2011).

Multisource data

Of interest to this paper are multi-source data and, in particular, the cross-source relationships existing between the variables of different sources. We assume that there are several sources of structural variation that give rise to the interconnections within and possibly between different sources. Together, these sources form a low-rank representation of the correlation matrix as used in factor and principal component models where the correlations can be reproduced on the basis of the loadings of the variables on the factors (or, latent variables). Let $\mathbf{\Lambda}$ be the matrix of loadings of the p variables on the R components; then, the factor analytic model assumes the following data generating mechanism for the observed data $\mathbf{x}$:

$$f(\mathbf{x}|\mu, \sigma) \sim \mathcal{N}(0, \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathrm{diag}([\sigma_1^2, \ldots, \sigma_j^2, \ldots, \sigma_p^2]), \quad (3)$$

with σ_j^2 the residual variance of the jth variable and $\mathbf{\Lambda}\mathbf{\Lambda}^T$ the covariance (correlation) matrix as reproduced by the factor analytic model. Here, to account for the multi-source structure $p = \sum_k p_k$ for $k = 1, \ldots, K$ sources with p_k the number of variables in source k; and also, $\mathbf{x} = [\mathbf{x}_1^T \ldots \mathbf{x}_K^T]^T$ so $\mathbf{\Lambda} = [\mathbf{\Lambda}_1^T \ldots \mathbf{\Lambda}_K^T]^T$. Under this model the covariance matrix $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \text{diag}([\sigma_1^2, \ldots, \sigma_j^2, \ldots, \sigma_p^2]$ and, making use of the Woodbury formula, its inverse is

$$\mathbf{\Sigma}^{-1} = \mathbf{D} - \mathbf{D}\mathbf{\Lambda}\left(\mathbf{I} - \mathbf{\Lambda}^T\mathbf{D}\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^T\mathbf{D} \tag{4}$$

with $\mathbf{D} = \text{diag}([\sigma_1^{-2}, \ldots, \sigma_j^{-2}, \ldots, \sigma_p^{-2}]$ and $\mathbf{I}$ the $R \times R$ identity matrix.

What we aim for is to determine the connection strength of the bridge nodes (variables) that connect between different sources. Application of the GGM to the observed data will not reach this aim. The reason for this is twofold. First, connection strengths between variables of different sources are low because of the different nature of the data between different sources. Under sparseness restrictions (see equation (2)), such weak correlations will be set equal to zero. Second, several sources of structural variation may underlie a variable whereby it may be involved both in cross-source and within-source relations. Let $\mathbf{\Lambda}_C$ represent the factors associated to the cross-source structural variation and $\mathbf{\Lambda}_S$ the source-specific variation, then $\mathbf{\Lambda} = [\mathbf{\Lambda}_C|\mathbf{\Lambda}_S]$ and

$$\begin{aligned} \mathbf{\Lambda}\mathbf{\Lambda}^T &= [\mathbf{\Lambda}_C|\mathbf{\Lambda}_S][\,\mathbf{\Lambda}_C|\mathbf{\Lambda}_S]^T \\ &= \mathbf{\Lambda}_C\mathbf{\Lambda}_C^T + \mathbf{\Lambda}_S\mathbf{\Lambda}_S^T, \end{aligned} \tag{5}$$

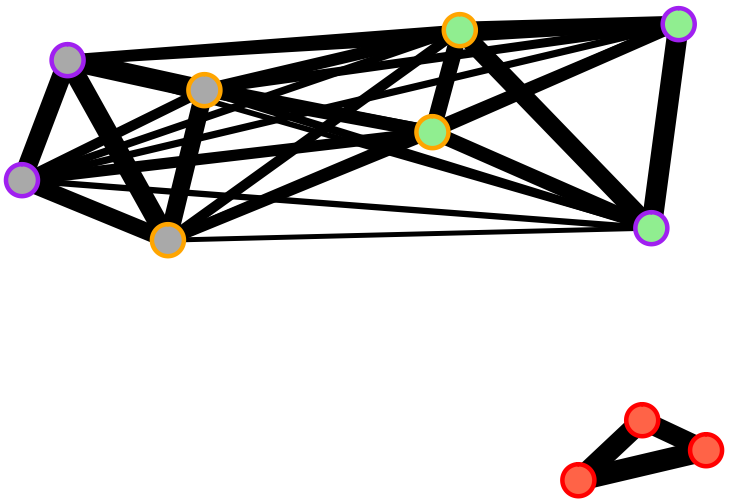
showing that the covariances consist of both cross-source and source-specific contributions.

To illustrate this, think of a hypothetical dataset containing data from three sources: four cognitive, four genetic, and three cardiovascular variables (Figure 1a). Here we not only need to a) differentiate relevant (purple outer circle) from irrelevant (red outer circle) variables and b) identify which of the relevant variables form cross-source relationships (orange outer circle), but additionally c) separate source-specific (S) from shared (or common; C) variation (see Figure 1b). In the next paragraph we show how the cross-source relations can be singled out.

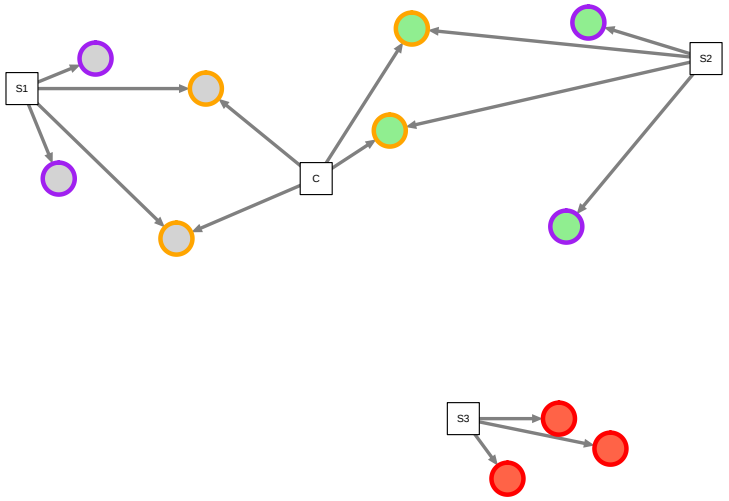
Isolating the cross-source relations

To single out the cross-source relations, we need to disentangle the common sources of variation shared between the different data blocks (with each block containing the data or variables of one source) from the sources of variation that are specific for a single or a few data blocks only. To do this, we perform a so called sparse DIStinctive and COmmon Simultaneus Component Analysis decomposition of the data (sparse DISCO SCA; see (Gu & Van Deun, 2016, 2017)), a method that was developed for the integrated analysis of multi-source data with the specific aim of separating block-specific sources of variation from common sources of variation. Sparse DISCO SCA models common and specific components by using specific constraints on the loadings of each of the components in $\mathbf{\Lambda}$. Next, we explain how this is done.

First, note that the loading of a variable on the component reflects the correlation of that variable with the component and this property is used to define common and



(a)



(b)

Figure 1. **A** Visualisation of hypothetical relationships between 4 cognitive (green), 4 genetic (purple), and 3 cardiovascular (orange) variables. Thickness of the edge is proportional to the strength of the relationship. **B** Relationships can be due to source specific (S) or common (C) sources of variation When estimating cross-source relationships, one needs to differentiate between irrelevant variables (red outer circle), relevant variables (purple outer circle), and those variables that are involved in cross-source relationships (orange outer circle).

specific components: A **common component** is associated to all data blocks and thus *each of the data blocks should have some variables with non-zero loadings* on this component; on the other hand, a **specific component** has no association at all with one (some) of the blocks and, consequently, for that (these) blocks *all variables of this block (these blocks)* should have a zero loading on this component. We give an illustration of such a loading matrix in Table 1 for the case of two data blocks (or, sources) having one common component (c), one component specific for the first source (s_1), and one specific for the second source (s_2): $\mathbf{\Lambda} = [\lambda_c \,|\lambda_{s1}\,|\lambda_{s2}]$. Note that sparse components are assumed, meaning that only a few relevant variables make up the component and thus also for the common component zero loadings do show up as well as in the non-zero part of the specific components.

Table 1: Sparse 3-component structure with one common component (non-zero loadings of both groups 1 and 2) and two source-specific components (non-zero loadings only from one group of variables).

	Common	Source 1-specific	Source 2-specific
Source 1	0	0	0
	$x_{2,1}$	$x_{2,2}$	0
	$x_{3,1}$	$x_{3,2}$	0
	0	0	0
	$x_{5,1}$	$x_{5,2}$	0
	0	0	0
	0	$x_{7,2}$	0
	0	0	0
	$x_{9,1}$	0	0
	$x_{10,1}$	$x_{10,2}$	0
	0	0	0
	$x_{12,1}$	$x_{12,2}$	0
Source 2	$x_{13,1}$	0	0
	$x_{14,1}$	0	$x_{14,3}$
	0	0	0
	$x_{16,1}$	0	0
	0	0	$x_{17,3}$
	$x_{18,1}$	0	$x_{18,3}$
	$x_{19,1}$	0	$x_{19,3}$
	0	0	$x_{20,3}$
	0	0	0
	$x_{22,1}$	0	0
	$x_{23,1}$	0	0
	$x_{24,1}$	0	$x_{24,3}$

To obtain such constrained structures, sparse DISCO SCA makes use of a com-

bination of penalties and/or hard constraints. Sparseness of the common component and of the non-zero part of the specific components is imposed by a lasso penalty on the loadings. The blocks of zero loadings of the specific components can be obtained in two ways: either by using a constrained approach or by using a group lasso penalty (Van Deun et al., 2011). The former approach requires prior knowledge of the structure of the components. In absence of such prior knowledge, when the number of blocks and components is small, -which is often the case in empirical applications - an exhaustive strategy can be used that compares all possible combinations of common and specific components. We refer to (Schouteden et al., 2014; Gu & Van Deun, 2017) for further discussion of this model selection issue.

Once the data have been modelled with common and specific components through a sparse DISCO SCA analysis, the cross-source relationships can be obtained from the common components as follows: Let $\hat{\mathbf{\Lambda}}_C$ denote the estimated loadings associated to the common components, then $\hat{\mathbf{\Sigma}}_C = \hat{\mathbf{\Lambda}}_C\hat{\mathbf{\Lambda}}_C^T$ reflects the cross-source correlation. For computational efficiency, only the variables containing at least one non-zero loading on a common component should be included.

SNAC: Sparse Network And Component model

When estimating unique cross-source relationships from multi-source data, a statistical procedure is needed that a) combines data from different sources where each source may have different characteristics, b) selects variables to determine which variables are involved in cross-source relationships, c) estimates these unique cross-source relationships, and d) presents results that can be interpreted in a meaningful, substantive way. As such graphical models, which covers the latter two criteria, are expected to be complemented by sparse DISCO SCA, which covers the first two criteria.

We therefore introduce Sparse Network And Component model (SNAC; Figure 2e), a two-step component-graphical model for estimating cross-source relationships from multi-source, multivariate Gaussian distributed data. First, sparse DISCO SCA is used to reveal the underlying common and source-specific sources of variation; as discussed in the previous section this allows us to single out the common source source of variation by calculating $\hat{\mathbf{\Lambda}}_C\hat{\mathbf{\Lambda}}_C^T$ which is the matrix containing the non-unique cross-source relations (see Figure 2d) while our interest is in the unique cross-source relations. A straightforward way to obtain these may seem to calculate the inverse of $\hat{\mathbf{\Lambda}}_C\hat{\mathbf{\Lambda}}_C^T$. Yet, $\hat{\mathbf{\Lambda}}_C$ is a low rank matrix of rank R_C, this is the number of common components. Furthermore, generalized inverses such as the Moore-Penrose inverse and the regularized estimation of inverse covariance matrices (shrinkage estimator; Schäfer et al., 2005) - although able to deal with the singularity incurred by the low rank - also do not give the desired result. This is because the pre-processing of the data in the sparse DISCO SCA removes the information in the data on the residual variances σ_j^2 ; see the data generating model for multisource data given by expression (3). Sparse DISCO SCA models the covariance matrix as $\mathbf{\Lambda}\mathbf{\Lambda}^T$ with inverse $(\mathbf{\Lambda}\mathbf{\Lambda}^T)^{-1}$ while the inverse of the population covariance matrix is given by expression (4). Considering the decomposition into common and

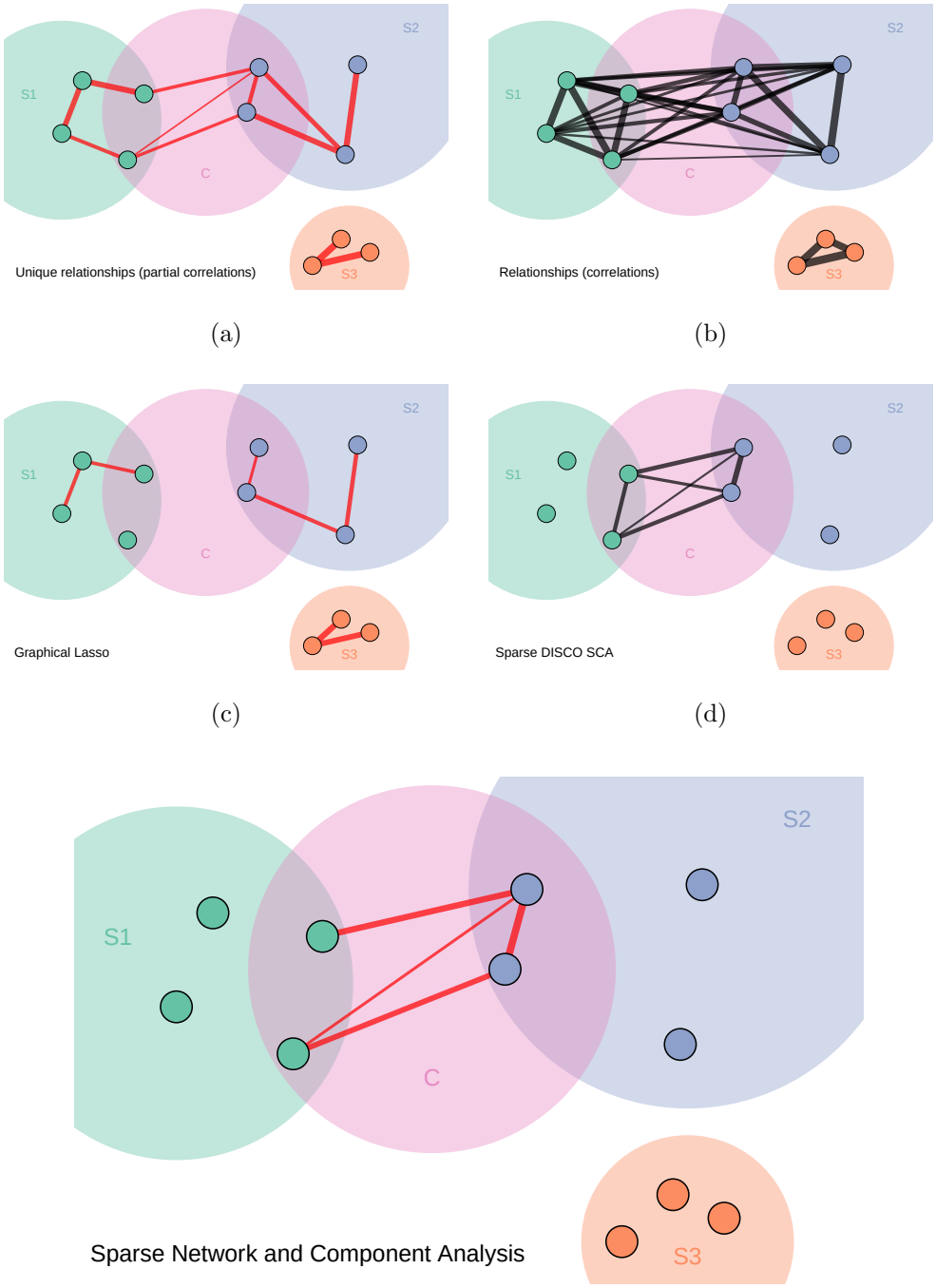


Figure 2. Different models capture different parts of the original data structure. **A** visualises the unique hypothetical relationships between 4 cognitive (green), 4 genetic (purple), and 3 cardiovascular (orange) variables. Variables involved in cross-source relationships are in the pink circle. **B** The corresponding correlation structure. **C** Graphical lasso recovers a sparse solution of the unique relationships structure containing the strongest relationships. This solution does not contain the unique cross-source relationships. **D** Sparse DISCO SCA disentangles source-specific (S) from common (C) sources of variation, retaining only the cross-source relationships. However, these relationships are not unique. **E** Sparse Network and Component Analysis combines the strength of both graphical lasso and sparse DISCO SCA and can therefor estimate the unique cross-source relationships. Thickness of the edge is proportional to the strength of the relationship.

source-specific components $\mathbf{\Lambda} = [\mathbf{\Lambda}_C|\mathbf{\Lambda}_S]$, we find for the inverse

$$(\mathbf{\Lambda}\mathbf{\Lambda}^T)^{-1} = \mathbf{B}^+ - \mathbf{B}^+\mathbf{\Lambda}_C(\mathbf{I} - \mathbf{\Lambda}_C^T(\mathbf{\Lambda}_S\mathbf{\Lambda}_S^T)^{-1}\mathbf{\Lambda}_C)^{-1}\mathbf{\Lambda}_C^T\mathbf{B}^+ \tag{6}$$

where $\mathbf{B} = \mathbf{\Lambda}_S\mathbf{\Lambda}_S^T$ and $^+$ denotes the Moore-Penrose inverse. We can see that the source-specific components scale the common component. This will lead to some confounding of the two different components.

Let us now inspect the generalized inverses of this covariance matrix a bit closer. First, to study the Moore-Penrose inverse, we consider the eigenvalue decomposition: $\mathbf{\Lambda}\mathbf{\Lambda}^T = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$ with $\mathbf{V}$ containing the R_C eigenvectors and $\mathbf{S}^2$ a diagonal matrix containing the eigenvalues. The Moore-Penrose inverse of $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is then equal to $\mathbf{V}\mathbf{S}^{-2}\mathbf{V}^T$ while for the population covariance matrix, replacing $\mathbf{\Lambda}$ by $\mathbf{VS}$ in expression (4), is given by

$$\mathbf{\Sigma}^{-1} = \mathbf{D} - \mathbf{DVS}\left(\mathbf{I} - \mathbf{SV}^T\mathbf{DVS}\right)^{-1}\mathbf{SV}^T\mathbf{D}. \tag{7}$$

This implies that the Moore-Penrose inverse is not suitable to estimate the direct cross-source relations. The regularized estimation of inverse covariance matrices as presented by (Schäfer et al., 2005) is not suitable either as it estimates the variances on the basis of the given covariance matrix and, these are biased downwards as a result of the sparse DISCO SCA step. The graphical lasso, on the other hand, inflates the given variances by fixing the variances as follows in the first step of the iterative estimation procedure used to estimate the off-diagonal elements of the inverse covariance matrix:

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \rho\mathbf{I}, \tag{8}$$

with $\rho \geq 0$, see page 3 in Friedman et al. (2008). Note that this expression is very similar to the expression of the population covariance matrix (3). The inverse of this covariance matrix closely resembles the inverse population covariance matrix:

$$\hat{\mathbf{\Sigma}}^{-1} = (\rho\mathbf{I}) - (\rho\mathbf{I})\mathbf{VS}\left(\mathbf{I} - \mathbf{SV}^T(\rho\mathbf{I})\mathbf{VS}\right)^{-1}\mathbf{SV}^T(\rho\mathbf{I}), \tag{9}$$

showing that the GGM will correctly estimate the direct relations when $\sigma_j^2 = \rho$ for all j. Hence, the way to go in constructing networks for cross-source relations is to combine a sparse DISCO SCA analysis with a Gaussian Graphical model. A summary of the steps needed to model the direct relations of the cross-source relations, is shown as pseudo code in Algorithm 1: SNAC, a Sparse Network And Component model.

Algorithm 1 SNAC: Sparse Network And Component model

1. Apply sparse DISCO SCA to obtain the loadings of the common component $\hat{\mathbf{\Lambda}}_C$
2. Calculate the cross-source relations $\hat{\mathbf{\Sigma}_C} = \hat{\mathbf{\Lambda}}_C\hat{\mathbf{\Lambda}}_C^T$
3. Estimate the direct cross-source relations by applying the graphical lasso to $\hat{\mathbf{\Sigma}}_C$

Simulation study

In this section, we describe the simulation study designed to compare the performance of graphical lasso, a commonly used GGM technique, and SNAC in estimating the unique cross-source relationships amongst many variables from two multivariate normal sources. Additionally we demonstrate that a different covariance-inversion technique such as Moore-Penrose or shrinkage estimator inaccurately estimates unique (cross-source) relationships, and thus are inappropriate statistical analyses for such research questions.

Design

Three factors were systematically manipulated in a factorial design:

- *Sparsity* of the component structure. Either 20 or 50 percent of all non-zero component loadings were set to 0.
- *Common component importance ratio.* It is not unlikely that, in comparison to the common source of variation, the source-specific sources of variation dominate the relationships between variables. To investigate whether cross-source relationships from a relative weaker common source of variation can be detected amongst stronger source-specific relationships, we manipulate the common component such that it is either equally strong (S1) or weaker (S2) compared to the source-specific components.
- *n/p ratio.* The ratio of number of individuals (n) to number of variables (p): 2/3 ($n = 200$, $p = 300$; $n < p$), 1 ($n = 300$, $p = 300$; $n = p$), and 3/2 ($n = 300$, $p = 200$; $n > p$).

Data generation

For this simulation study we generated multivariate normal data with variance-covariance matrix $\mathbf{\Sigma}$, which reflects a sparse three-component, two-source structure (see Table 1 and equation 5). On two of the three components, only the variables from one of two groups load; these two components reflect source-specific sources of variation (*source-specific or S-components*). The third component contains loadings for variables of both groups and reflects multi-source or common sources of variation (*common or C-component*).

The sparse three-component structure is generated as follows. First, a singular value decomposition is applied to a randomly generated, standard-normal distributed data matrix $\mathbf{L} = \mathbf{UWV}^T$ with dimensions $n \times p$, where the first half of the variables were set to belong to the first data source and the rest to the second source. From the resulting decomposition we derive orthogonal component loadings $\mathbf{\Lambda}$ by setting them equal to the three right singular vectors associated with the three largest singular values. Component loadings in $\mathbf{\Lambda}$ are set to zero such that its first component reflects a common source of variation and the other two components reflect source-specific sources of variation. Lastly, sparseness was introduced on the remaining non-zero loadings.

A 3×3 diagonal matrix served as singular values matrix $\mathbf{S}$, and indicates how relative important each component is. When all components are equally important, $\mathbf{S}$ is an identity matrix. To create a data structure that is dominated by source-specific sources of variation, their singular values will be set to 2 (resulting in the diagonal 1 2 2). In empirical data it is unlikely that two groups of variables have source-specific correlation structures of equal strength. To incorporate this complexity in the data, the scaling vector g was manipulated to be $g_1 = 0.8$ for source 1 and $g_2 = 0.3$ for source 2:

$$\lambda_{kr}^{true} = \sqrt{\frac{g_k}{\sum_{r=1}^{R} \lambda_{kr}^2 s_{rr}^2}} \lambda_{kr}, \tag{10}$$

where s_{rr} is the singular value of the r^{th} component and $k = 1, 2$. Using $\mathbf{\Lambda}^{true}$ and $\mathbf{S}$, the true variance-covariance matrix $\mathbf{\Sigma}^{true}$ was calculated,

$$\mathbf{\Sigma}^{true} = \mathbf{\Lambda}^{true}\mathbf{S}^2(\mathbf{\Lambda}^{true})^T, \tag{11}$$

where the diagonal of $\mathbf{\Sigma}^{true}$ was manually set to 1 (see also equation 3), resulting in lower correlations. The true off-diagonal values in the covariance matrix representing the source-specific sources can be calculated in similar fashion

$$\mathbf{\Sigma}_S^{true} = \mathbf{V}_S^{true}\mathbf{S}_S^2(\mathbf{V}_S^{true})^T; \tag{12}$$

noting that the diagonal values in $\mathbf{\Sigma}_S^{true}$ are not the variances of the variables. The true covariance matrix of the common source of variability between variables (the common component) is the difference between equations (11) and (12)

$$\mathbf{\Sigma}_C^{true} = \mathbf{\Sigma}^{true} - \mathbf{\Sigma}_S^{true}. \tag{13}$$

Given our interest in unique relationships amongst variables, we are particularly interested in the inverse covariance matrices. Both $\mathbf{\Sigma}^{true}$ and $\mathbf{\Sigma}_C^{true}$ are invertible, leading to the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ and the desired inverse covariance matrix of the common source of variability $[\mathbf{\Sigma}_C^{true}]^{-1}$.

Finally, data was generated from a multi-variate normal distribution with $\mu = \mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}^{true}$.

Analyses

The recovery performance of graphical lasso, SNAC and the combination of sparse DISCO SCA with either Moore-Penrose or shrinkage estimator was assessed on four covariance matrices: Covariance matrix $\mathbf{\Sigma}$, inverse covariance matrix $\mathbf{\Sigma}^{-1}$, covariance matrix of the information of the common component $\mathbf{\Sigma}_C$, and inverse covariance matrix of the information of the common component $\mathbf{\Sigma}_C^{-1}$. Especially the performance on estimating the last is of importance, since this covariance matrix models the unique cross-source relationships.

Two fit indices were used: (i) the percentage of correctly estimated zeros and non-zeros (Selection status recovery; SSR)

$$SSR = \frac{\text{number of correct 0 and non-0 edges}}{p(p-1)/2}, \tag{14}$$

where p is the number of variables, and (ii) Tucker's coefficient (TC; Tucker, 1951)

$$TC = \frac{\mathbf{U}^T\mathbf{Z}}{\sqrt{(\mathbf{U}^T\mathbf{U})(\mathbf{Z}^T\mathbf{Z})}}, \tag{15}$$

where $\mathbf{U}$ is the upper triangle of the population (inverse) covariance matrix, and $\mathbf{Z}$ is the upper triangle of the estimated (inverse) covariance matrix. Values between .85 and .95 indicate "a fair similarity" and values above .95 indicate that the two covariance matrices can be considered equal (Abdi, 2007).

With the exception of the Moore-Penrose inverse, all statistical methods used in this study require input for a lasso tuning parameter. Additionally, sparse DISCO SCA requires the number of common and distinctive components, and their sparseness. To avoid confounding influence of potentially mis-specifying the lasso-tuning parameter, we set it such that the estimated (inverse) covariance matrix recovers the true amount of zeros as much as possible. The range of selected tuning parameters for graphical lasso, SNAC and shrinkage estimator can be found in Appendix 1. As for sparse DISCO SCA, given that the component structure was assumed known tuning parameters for the lasso are not needed.

Graphical lasso was performed using the function *glasso* (*R*-package *glasso*, (Friedman et al., 2014)). Sparse DISCO SCA was performed using adjusted code from *R*-package *RegularizedSCA* (Gu & Van Deun, 2017). Moore-Penrose inverse was performed using the function *ginv* (*R*-package MASS). Shrinkage estimation was performed using the function *pcor.shrink* (*R*-package *corpcor*, Schäfer et al. (2005)).

Results

Graphical lasso

The graphical lasso has been developed to estimate sparse (inverse) covariance matrices from datasets. It is therefore not surprising that it estimates the covariance matrix $\mathbf{\Sigma}$ very well (SSR .81 - .93; TC .88 - .97, see Table 2) even when $p > n$ (SSR .76 - .83; TC .93 - .97). The Inverse covariance matrix $\mathbf{\Sigma^{-1}}$ is estimated less accurately (SSR .57 - .84; TC .86 - 98) especially when $p > n$ (SSR .47 - .70; TC .18 - .28).

Compared to the full covariance matrix, the graphical lasso estimates the covariance matrix of the *common variation* $\mathbf{\Sigma_C}$ less well, though there is not much difference between $p > n$ (SSR .64 - .80; TC .46 - .70) and other n/p-ratios (SSR .62 - .85; TC .42 - .64). The inverse covariance matrix of the common source of variance $\mathbf{\Sigma_C^{-1}}$ is also estimated less accurately compared to the full inverse covariance matrix (SSR .64 - .91; TC .59 - 64), with similar poor results for $p > n$ (SSR .50 - .65; TC .17 - .28). This is

Table 2: Performance of graphical glasso on estimating (unique) (cross-source) relationships. A priori indicates that information on which variables are part of cross-source relationships is known. In bold the results for the common (C) inverse covariance matrix. SSR = Selection status recovery; TC = Tucker's congruence; Sparse = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio.

		Selection status recovery				Tucker's congruence			
		Sparse 20		Sparse 50		Sparse 20		Sparse 50	
n/p		S1	S2	S1	S2	S1	S2	S1	S2
2/3	Covariance	0.76	0.83	0.83	0.82	0.94	0.97	0.93	0.94
	Covariance C	0.72	0.64	0.80	0.79	0.70	0.51	0.53	0.46
	Inverse	0.70	0.70	0.47	0.47	0.18	0.18	0.28	0.28
	Inverse C	**0.50**	**0.50**	**0.65**	**0.63**	**0.17**	**0.17**	**0.28**	**0.26**
	Inverse C a priori	**0.54**	**0.54**	**0.24**	**0.29**	**0.21**	**0.21**	**0.42**	**0.41**
1	Covariance	0.92	0.91	0.93	0.93	0.95	0.96	0.88	0.89
	Covariance C	0.85	0.84	0.88	0.88	0.61	0.44	0.50	0.42
	Inverse	0.78	0.77	0.84	0.84	0.86	0.87	0.88	0.88
	Inverse C	**0.80**	**0.79**	**0.91**	**0.90**	**0.59**	**0.59**	**0.62**	**0.62**
	Inverse C a priori	**0.60**	**0.55**	**0.59**	**0.56**	**0.71**	**0.56**	**0.82**	**0.81**
3/2	Covariance	0.81	0.83	0.83	0.81	0.96	0.97	0.93	0.93
	Covariance C	0.69	0.62	0.81	0.80	0.64	0.50	0.61	0.53
	Inverse	0.61	0.57	0.76	0.74	0.96	0.97	0.98	0.98
	Inverse C	**0.69**	**0.64**	**0.83**	**0.81**	**0.64**	**0.64**	**0.63**	**0.63**
	Inverse C a priori	**0.67**	**0.60**	**0.65**	**0.62**	**0.76**	**0.74**	**0.84**	**0.84**

expected given that it has no way of identifying which information is part of the cross-source relationships (captured in the common component) and which is not. However, what if based on theory or earlier findings one has information on which variables are part of cross-source relationships. Would taking this *a priori* information into account improve the graphical lasso estimation of the (inverse) covariance matrix of the common variation? Adding such information to the graphical lasso results in a decrease of the percentage correctly estimated zeros (SSR 0.55 - 0.62) but the accuracy of the estimates increases (TC 0.56 - 0.84). Again, for $p > n$ estimates of the inverse covariance matrix are less accurate (SSR 0.21 - 0.81; TC .21 - .42).

Table 3: Performance of Sparse Network and Component Analysis on estimating (unique) (cross-source) relationships. In bold the results for the common (C) inverse covariance matrix. SSR = Selection status recovery; TC = Tucker's congruence; Sparse = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio.

		SSR				TC			
		Sparse 20		Sparse 50		Sparse 20		Sparse 50	
n/p-ratio		S1	S2	S1	S2	S1	S2	S1	S2
2/3	Covariance	0.83	0.83	0.81	0.80	0.93	0.92	0.89	0.90
	Covariance C	0.98	0.96	1.00	0.99	1.00	1.00	1.00	1.00
	Inverse	0.31	0.31	0.62	0.61	0.63	0.63	0.66	0.66
	Inverse C	**0.75**	**0.69**	**0.90**	**0.89**	**0.86**	**0.89**	**0.87**	**0.83**
1	Covariance	0.83	0.81	0.81	0.80	0.93	0.92	0.89	0.90
	Covariance C	0.95	0.91	0.98	0.96	1.00	1.00	1.00	1.00
	Inverse	0.29	0.29	0.60	0.60	0.60	0.60	0.64	0.64
	Inverse C	**0.68**	**0.62**	**0.86**	**0.84**	**0.73**	**0.75**	**0.75**	**0.76**
3/2	Covariance	0.83	0.82	0.81	0.80	0.93	0.92	0.89	0.90
	Covariance C	0.98	0.95	1.00	0.99	1.00	1.00	1.00	1.00
	Inverse	0.31	0.31	0.61	0.61	0.63	0.63	0.66	0.67
	Inverse C	**0.75**	**0.69**	**0.90**	**0.89**	**0.86**	**0.89**	**0.87**	**0.83**

SNAC

The Sparse Network And Component analysis performs very well when estimating covariance matrices $\mathbf{\Sigma}$ (SSR .80 - .83; TC .89 - .93, see Table 3) and $\mathbf{\Sigma_C}$ (SSR .95 - 1.00; TC 1.00) regardless of the n/p-ratio. The under performance on estimating inverse covariance matrix $\mathbf{\Sigma^{-1}}$ (SSR 0.29 - 0.62; TC 0.60 - 0.67) might seem counterintuitive

unless one realises that SNAC applies a lasso twice, once during the sparse DISCO SCA, and once during the graphical lasso procedure. This together with the already sparse component structure we are trying to recover is likely the cause of the too sparse solution with less accurate estimates. Most importantly however, SNAC adequately estimates inverse covariance matrix of the *common source of variation* $\mathbf{\Sigma_C^{-1}}$ (SSR 0.62 - 0.90; TC 0.73 - 0.89). In the case of more variables than observations, the most likely scenario when dealing with multi-source data, it outperforms the graphical lasso (see Figure 3).

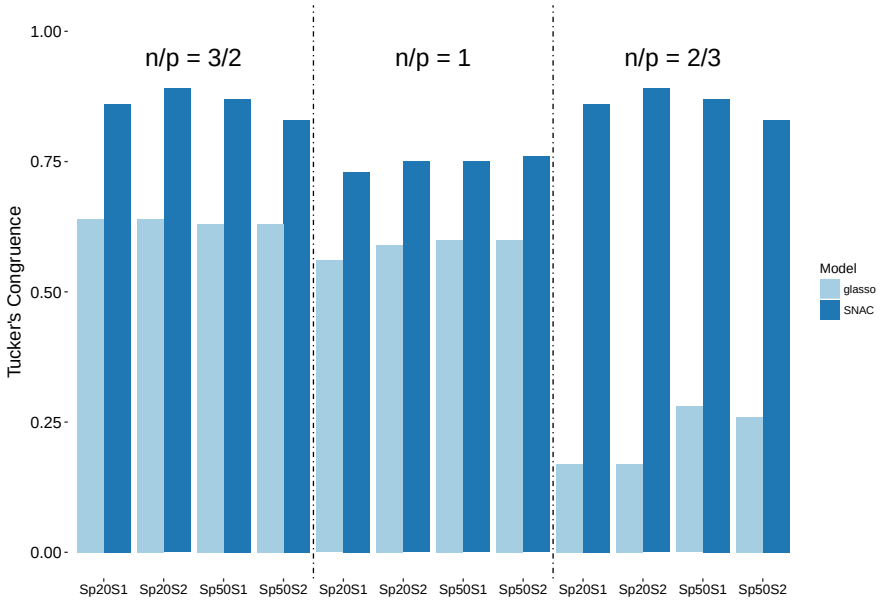
Sparse DISCO SCA

Sparse DISCO Simultaneous Component Analysis has been developed to find source-specific and common components that together maximise the amount of variance explained. As such it comes to no surprise that this statistical method accurately estimates the (off-diagonal part of the) covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_C$ (SSR 0.78 - 1; TC 0.97 - 1; see Table 4), regardless of component sparseness, component importance ratio, and observation-variable ratio. Due to the fact that these covariance matrices are based on either three ($\mathbf{\Sigma}$) or one component ($\mathbf{\Sigma}_C$), these covariance matrices are singular and thus cannot be inverted. In such cases, however, it is possible to calculate a pseudoinverse such as the Moore-Penrose pseudoinverse (MP) and shrunken partial correlations (Shrink). While both methods almost perfectly identify the zeroes (SSR 0.90 - 1), the actual estimates are misspecified (TC-0.51 - -0.86). One may think that an explanation for these negative Tucker congruences might be a change in signs; however, this seems unlikely given that the component model is sign invariant (the change of sign in loadings is compensated by a change of sign in the component scores).

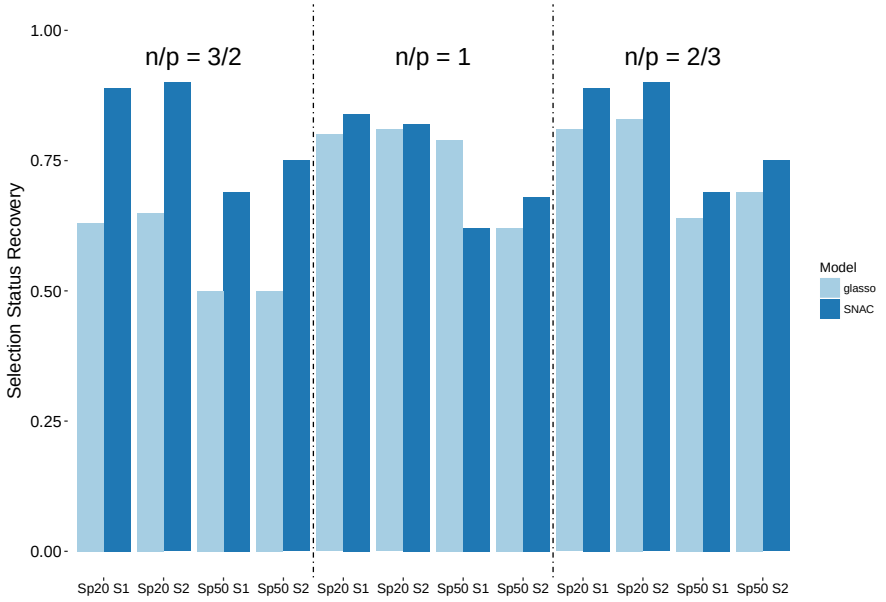
Conclusion and Discussion

In this paper, we propose a component-network hybrid method that is suited for estimating the unique cross-source relationships amongst multivariate normally distributed variables found in multi-source datasets. This Sparse Network And Component (SNAC) model combines the strengths of two existing methods: the variable selecting and information dis-entangling properties of sparse DISCO Simultaneous Component Analysis (SCA) with the conditional dependence focused theoretical framework of Gaussian graphical models (GGMs). We have shown that SNAC outperforms a regular graphical model technique in accurately estimating unique cross-source relationships, especially when the data consists of more variables than observations. These results hold even when a priori knowledge about which variables are part of cross-source relations is available. Neither variations in sparseness of the estimated structure nor in the relative dominance of source-specific sources of variation influence SNACs performance.

Using the GGM framework, we can model unique relationships by estimating the inverse covariance matrix. However, adding sparse DISCO SCA to GGM has two consequences. First, it is not possible to use GGM to estimate an inverse covariance matrix based on a limited number of SCA-estimated components. This means that an pseudoinverse is required to gain insight in the relationships between variables. Second,



(a)



(b)

Figure 3. Tucker's congruence and Selection status recovery by graphical lasso (glasso) and sparse network and component analysis (SNAC) for estimating the inverse covariance matrix of the common source of variation. Sp = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio.

Table 4: Performance of Sparse DISCO Simultaneous Component Analysis on estimating (unique) (cross-source) relationships. In bold the results for the common (C) inverse covariance matrix. For the inverse, Moore-Penrose (MR) inverse and Shrinkage estimator (Shrink) are reported. SSR = Selection status recovery; TC = Tucker's congruence; Sparse = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio.

		SSR				TC			
		Sparse 20		Sparse 50		Sparse 20		Sparse 50	
n/p-ratio		S1	S2	S1	S2	S1	S2	S1	S2
2/3	Covariance	0.83	0.82	0.78	0.78	1.00	0.98	1.00	1.00
	Covariance C	0.99	0.89	1.00	1.00	1.00	1.00	1.00	1.00
	Inverse MP	1.00	0.99	1.00	1.00	-0.57	-0.63	-0.58	-0.60
	Inverse Shrink	1.00	0.99	1.00	0.99	-0.79	-0.78	-0.78	-0.78
	Inverse C MP	**0.99**	**0.91**	**1.00**	**1.00**	**-0.86**	**-0.85**	**-0.86**	**-0.86**
	Inverse C Shrink	**0.99**	**0.91**	**1.00**	**0.99**	**-0.57**	**-0.52**	**-0.60**	**-0.56**
1	Covariance	0.83	0.82	0.78	0.78	1.00	0.99	1.00	1.00
	Covariance C	0.99	0.90	1.00	1.00	1.00	0.97	1.00	1.00
	Inverse MP	1.00	0.99	1.00	1.00	-0.57	-0.63	-0.58	-0.60
	Inverse Shrink	1.00	0.99	1.00	0.99	-0.79	-0.78	-0.79	-0.79
	Inverse C MP	**0.99**	**0.91**	**1.00**	**1.00**	**-0.86**	**-0.85**	**-0.86**	**-0.86**
	Inverse C Shrink	**0.99**	**0.91**	**1.00**	**0.99**	**-0.56**	**-0.52**	**-0.60**	**-0.56**
3/2	Covariance	0.83	0.83	0.80	0.80	1.00	0.99	1.00	1.00
	Covariance C	0.99	0.91	1.00	1.00	1.00	0.98	1.00	1.00
	Inverse MP	1.00	0.98	1.00	1.00	-0.58	-0.63	-0.57	-0.60
	Inverse Shrink	1.00	0.98	0.99	1.00	-0.79	-0.78	-0.78	-0.78
	Inverse C MP	**0.99**	**0.90**	**1.00**	**0.99**	**-0.86**	**-0.84**	**-0.86**	**-0.86**
	Inverse C Shrink	**0.99**	**0.90**	**0.99**	**0.99**	**-0.57**	**-0.51**	**-0.60**	**-0.56**

adding the component pre-processing step changes the assumed underlying data generating structure of our model. As demonstrated, inverse techniques that assume correctly estimated variances, such as Moore-Penrose and shrinkage estimator, are unsuitable as they do not take the assumed underlying factor analytic structure, this is including residual variances in the data generating model, into account.

Applying the SNAC model to data requires input for several hyper-parameters: the number of common and source-specific components, and tuning parameters for several lassos. Because the purpose of this paper is to provide a proof of concept, we set these parameters as close to their optimal value as possible. Selecting the proper parameters in a non-simulation study, especially when analysing high-dimensional data, is challenging. As demonstrated by (Gu & Van Deun, 2016), the tuning parameter of *l1*-Lasso can successfully be selecting using (Meinshausen & Bühlmann, 2010)'s resample-based stability selection method, although the multi-component structure of both (Gu & Van Deun, 2016)'s and our work complicates matters further.

In its current form, SNAC is a two-step procedure: first we apply sparse DISCO SCA to the data, whose results we then use as input for the graphical lasso. This step-by-step operation might unintentionally introduce bias, one form of which is the double shrinkage to zero by applying twice a penalized approach; whether we can optimally estimate the unique cross-source relationships depends fully on how accurate the information dis-entanglement during the first step has been. One way to decrease this bias is by performing the information entanglement and unique relationship estimation in an iterative fashion. Such simultaneous SNAC, however, requires a mathematical equivalence between component and graphical model. One possibility is interpreting the common component as identified by SCA as a latent variable which again is equivalent, under certain assumptions, to a network clique (a subset of an undirected graph that is complete; every pair of nodes is connected by a unique edge).

Finally, in this paper we have only considered Gaussian distributed data, which while common are not the only data-type in multi-source data. Both sparse DISCO Simultaneous Component Analysis and graphical lasso are able to handle non-Gaussian variables, and as such we expect that this characteristic will be transferred to Sparse Network And Component analysis. Further research will have to investigate the translation from non-Gaussian-based component scores to non-Gaussian-based partial correlations.

P. Tio (Corresponding author)
Department of Psychological Methods, University of Amsterdam, the Netherlands
Department of Methodology and Statistics, Tilburg University, the Netherlands
E-mail address: piatio@gmail.com

L.J. Waldorp
Department of Psychological Methods, University of Amsterdam, the Netherlands

K. Van Deun
Department of Methodology and Statistics, Tilburg University, the Netherlands

References

Abdi, H. (2007). Rv coefficient and congruence coefficient. *Encyclopedia of measurement and statistics*, 849–853.

Bartel, J., Krumsiek, J., & Theis, F. J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. *Computational and structural biotechnology journal*, *4*(5), 1–9.

Bilodeau, M., & Brenner, D. (1999). *Theory of multivariate statistics.* Springer.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Chu, J.-h., Lazarus, R., Carey, V. J., & Raby, B. A. (2011). Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. *BMC systems biology*, *5*(1), 89.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., ... others (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular psychiatry*, *16*(10), 996–1005.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2014). glasso: Graphical lasso- estimation of gaussian graphical models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=glasso` (R package version 1.8)

Golding, J. (1990). Children of the nineties. a longitudinal study of pregnancy and childhood based on the population of avon (alspac). *West of England medical journal*, *105*(3), 80–82.

Gu, Z., & Van Deun, K. (2016). A variable selection method for simultaneous component based data integration. *Chemometrics and Intelligent Laboratory Systems*, *158*, 187–199.

Gu, Z., & Van Deun, K. (2017). Rsca: Regularized simultaneous component analysis for data integration in r. *submitted to Journal of Statistical Software*.

Ingkasuwan, P., Netrphan, S., Prasitwattanaseree, S., Tanticharoen, M., Bhumiratana, S., Meechai, A., ... Cheevadhanarak, S. (2012). Inferring transcriptional gene regulation network of starch metabolism in arabidopsis thaliana leaves using graphical gaussian model. *BMC systems biology*, *6*(1), 100.

Johnson, M. R., Shkura, K., Langley, S. R., Delahaye-Duriez, A., Srivastava, P., Hill, W. D., ... others (2015). Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease. *Nature neuroscience*.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques.* MIT press.

Krämer, N., Schäfer, J., & Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC bioinformatics*, *10*(1), 384.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, *5*(1), 21.

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

López-Kleine, L., Leal, L., & López, C. (2013). Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. *Briefings in functional genomics*, *12*(5), 457–467.

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

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Schäfer, J., Strimmer, K., et al. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, *4*(1), 32.

Schouteden, M., Van Deun, K., Wilderjans, T. F., & Van Mechelen, I. (2014). Performing disco-sca to search for distinctive and common information in linked data. *Behavior research methods*, *46*(2), 576–587.

Silverman, E. K., & Loscalzo, J. (2012). Network medicine approaches to the genetics of complex diseases. *Discovery medicine*, *14*(75), 143.

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

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Tech. Rep.). DTIC Document.

Van Deun, K., Wilderjans, T. F., Van den Berg, R. A., Antoniadis, A., & Van Mechelen, I. (2011). A flexible framework for sparse simultaneous component based data integration. *BMC bioinformatics*, *12*(1), 448.

Villeneuve, D. L., Larkin, P., Knoebl, I., Miracle, A. L., Kahl, M. D., Jensen, K. M., ... others (2007). A graphical systems model to facilitate hypothesis-driven ecotoxicogenomics research on the teleost brain- pituitary- gonadal axis. *Environmental science & technology*, *41*(1), 321–330.

Appendix

Table 5: Range of selected tuning parameter lambda for graphical lasso. Sparse = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio; C = common (inverse) covariance matrix.

		Sparse 20		Sparse 50	
n/p-ratio		S1	S2	S1	S2
2/3	Covariance	0.2 - 0.3	0.2 - 0.3	0.3 - 0.4	0.3 - 0.4
	Covariance C	0.3	0.3	0.5	0.4 - 0.5
	Inverse	0.005	0.005	0.01 - 0.015	0.01 - 0.15
	Inverse C	0.01	0.01	0.03	0.025
1	Covariance	0.3	0.2 - 0.3	0.4	0.4
	Covariance C	0.4	0.3 - 0.4	0.5	0.5
	Inverse	0.005	0.005	0.005	0.005
	Inverse C	0.005	0.005	0.01 - 0.025	0.01 - 0.025
3/2	Covariance	0.2 - 0.3	0.2	0.3 - 0.4	0.3
	Covariance C	0.3	0.2 - 0.3	0.3 - 0.7	0.3 - 0.4
	Inverse	0.005	0.005	0.005	0.005
	Inverse C	0.005	0.005	0.01 - 0.025	0.005 - 0.025

Table 6: Range of selected tuning parameter lambda for Sparse Network and Component Analysis (SNAC). Sparse = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio; C = common inverse covariance matrix.

		Sparse 20		Sparse 50	
n/p-ratio		S1	S2	S1	S2
2/3	Covariance	0.3	0.1 - 0.3	0.3 - 0.4	0.1 - 0.4
	Covariance C	0.02 - 0.05	0.01 - 0.04	0.02 - 0.1	0.005 - 0.05
	Inverse	0.005 - 0.05	0.005 - 0.05	0.005 - 0.05	0.02 - 0.06
	Inverse C	0.005	0.005	0.005 - 0.01	0.005 - 0.01
1	Covariance	0.3	0.3 - 0.4	0.3 - 0.4	0.1 - 0.4
	Covariance C	0.02 - 0.06	0.03 - 0.10	0.03 - 0.1	0.03 - 0.06
	Inverse	0.005 - 0.05	0.005 - 0.025	0.005 - 0.025	0.005 - 0.025
	Inverse C	0.005	0.005 - 0.10	0.005 - 0.01	0.005
3/2	Covariance	0.4	0.3 - 0.4	0.3 - 0.4	0.3 - 0.4
	Covariance C	0.03 - 0.10	0.03 - 0.10	0.02 - 0.10	0.02 - 0.06
	Inverse	0.005	0.005 - 0.025	0.005 - 0.025	0.005 - 0.025
	Inverse C	0.01 - 0.015	0.005 - 0.010	0.005 - 0.010	0.005 - 0.01

Table 7: Range of selected tuning parameter Lambda for Shrinkage estimator (Schfer *et al.*, 2005) executed on the results of Sparse DISCO Simultaneous Component Analysis (SCA). Sparse = component sparseness; S1 = common and source-specific sources of variation are equally important; S2 = source-specific sources of variation dominate data structure; n/p-ratio = observation-variables ratio; C = common (inverse) covariance matrix.

		Sparse 20		Sparse 50	
n/p-ratio		S1	S2	S1	S2
2/3	Inverse	0.016 - 0.022	0.017 - 0.024	0.015 - 0.026	0.016 - 0.025
	Inverse C	0.006 - 0.015	0.006 - 0.017	0.006 - 0.015	0.007 - 0.015
1	Inverse	0.11 - 0.015	0.011 - 0.015	0.01 - 0.015	0.011 - 0.15
	Inverse C	0.005 - 0.020	0.005 - 0.011	0.005 - 0.011	0.005 - 0.010
3/2	Inverse	0.010 - 0.014	0.011 - 0.016	0.010 - 0.014	0.010 - 0.015
	Inverse C	0.004 - 0.010	0.004 - 0.009	0.005 - 0.009	0.005 - 0.010