OPEN ACCESS Freely available online

PLOS ONE

Iterative Reconstruction of High-Dimensional Gaussian Graphical Models Based on a New Method to Estimate Partial Correlations under Constraints

Vincent Guillemot, Andreas Bender, Anne-Laure Boulesteix*

Department of Medical Informatics, Biometry and Epidemiology of the Faculty of Medicine, University of Munich, Munich, Germany

Abstract

In the context of Gaussian Graphical Models (GGMs) with high-dimensional small sample data, we present a simple procedure, called PACOSE – standing for PArtial COrrelation SElection – to estimate partial correlations under the constraint that some of them are strictly zero. This method can also be extended to covariance selection. If the goal is to estimate a GGM, our new procedure can be applied to re-estimate the partial correlations after a first graph has been estimated in the hope to improve the estimation of non-zero coefficients. This iterated version of PACOSE is called iPACOSE. In a simulation study, we compare PACOSE to existing methods and show that the re-estimated partial correlation coefficients may be closer to the real values in important cases. Plus, we show on simulated and real data that iPACOSE shows very interesting properties with regards to sensitivity, positive predictive value and stability.

Citation: Guillemot V, Bender A, Boulesteix A-L (2013) Iterative Reconstruction of High-Dimensional Gaussian Graphical Models Based on a New Method to Estimate Partial Correlations under Constraints. PLoS ONE 8(4): e60536. doi:10.1371/journal.pone.0060536

Editor: Francesco Pappalardo, University of Catania, Italy

Received December 6, 2012; **Accepted** February 27, 2013; **Published** April 11, 2013

Funding: VG was supported by grant BO3139/2-1 from the German Science Foundation (DFG: http://www.dfg.de/en) to ALB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: boulesteix@ibe.med.uni-muenchen.de

Introduction

The robust estimation of the inverse covariance matrix is crucial in many multivariate statistical methods such as discriminant analysis or linear regression [1]. Many variants of these multivariate methods aim at somehow "regularizing" the estimation of the covariance matrix to make it invertible or better conditioned, *e.g.* ridge regression (RR), diagonal discriminant analysis or regularized discriminant analysis [2]. A large body of literature is devoted to the estimation of the inverse covariance matrix in high-dimensional small sample settings, *i.e.* when the number of observations n is much smaller than the number of variables p. A well-known example is the shrinkage estimator by Shäfer & Strimmer [3] which is defined as a weighted sum of the sample covariance matrix and a fixed (invertible) target matrix. This method can be considered as "agnostic" in the sense that it estimates the covariance matrix in a completely data-driven way, *i.e.* without prior knowledge.

In this article, we first propose a method that directly estimates the partial correlation matrix while taking into account prior information on the dependencies between variables materialized by a given undirected graph. In a nutshell, our new method takes such a graph – called "independence graph" – as input and estimates the non-zero coefficients of the partial correlation matrix by regularized linear regression using the regression-based definition of partial correlation. The inverse covariance matrix can then be simply obtained from the partial correlation matrix by incorporating estimates of the variances. In this sense, our method can be seen as a *covariance selection* algorithm [4]. Although many covariance selection methods have been proposed in the literature (see below for details), none of these methods is designed to estimate the partial correlation matrix in high-dimensional settings while incorporating a *non-decomposable* independence graph. In reference to covariance selection, we called this first method "PACOSE', standing for **PA**rtial **CO**rrelation **SE**lection.

Furthermore, we suggest a new iterative algorithm called "iPACOSE" – standing for iterative PACOSE – that estimates an independence graph from a dataset using our new partial correlation estimate in a recursive way. Briefly, iPACOSE takes as inputs a dataset and a significance level for the partial correlation and gives as an output an estimated independence graph. We show on simulated datasets that recursive reestimation of the partial correlation coefficients yields graphs closer to the true graph than a simple thresholding of an estimated partial correlation matrix.

The rest of the paper is structured as follows. We first present our iterative method and the associated covariance selection and also briefly reviews existing covariance selection methods. Then, we compare our new method to existing estimation algorithms for Gaussian Graphical Models (GGM) on simulated data. Finally, we apply our method to real datasets.

For the sake of reproducibility, we made our code available in the form of:

- An R package called pacose, available on the CRAN http://cran.r-project.org/web/packages/pacose/index.html (Accessed 2013 March 13),
- A set of R programs for the reproduction of our results, available online at http://www.ibe.med.uni-muenchen.de/

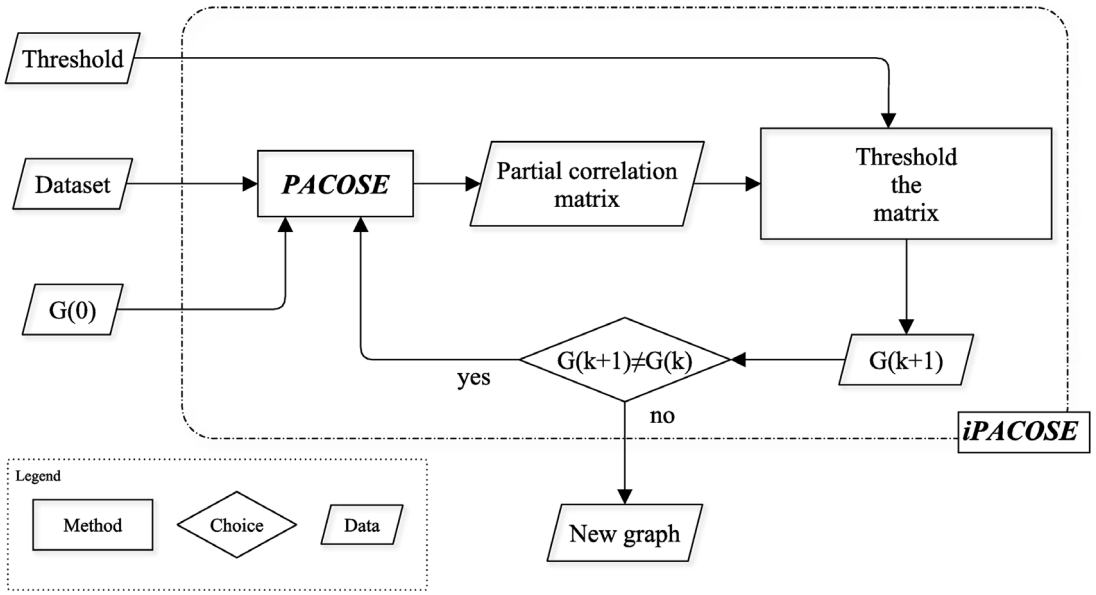

Figure 1. Flowchart representation of the iPACOSE algorithm, representing how it iteratively uses PACOSE to estimate an independence graph from a dataset.
doi:10.1371/journal.pone.0060536.g001

organisation/mitarbeiter/020_professuren/boulesteix/pacose2012/(Accessed 2013 March 13).

Methods

Context

The estimation of networks is a burning issue in bioinformatics. Gaussian graphical models (GGMs) [5,6] have been widely used for this purpose in the last few years [3,7]. In the context of systems biology, the estimation of GGMs is very often characterized by a lower number of individuals (n) or measures than the number of variables (p). In this $n \ll p$ situation, regularization techniques are mandatory to enable the estimation of GGMs.

The core method of the present work is designed to estimate a partial correlation matrix under the constraint that some known coefficients are equal to zero. It is intimately related to so-called *covariance selection* methods, which can themselves be seen as methods able to estimate the covariance matrix or its inverse, the so-called precision matrix, (i) under the constraint that some coefficients in the precision matrix are null [4] or (ii) under the constraint that a certain amount of coefficients are equal to zero in the precision matrix [8,9]. To avoid any confusion with these sensibly different definitions, we chose an acronym closely related to the parameters that we want to estimate: the partial correlations, hence the name of this core method: PACOSE, "PArtial COrrelation SElection". The theory behind PACOSE is further described in the section "PACOSE".

We propose to embed PACOSE into an iterative algorithm designed to estimate independence graphs. The algorithm – called iPACOSE (standing for *iterative* PACOSE) – takes a dataset and a significance level for the partial correlation coefficients as inputs. PACOSE is then applied iteratively to the dataset to estimate an independence graph extracted from the previous iteration's partial correlation matrix by thresholding it. The iPACOSE algorithm is schematically represented in Figure 1. iPACOSE is described in more details in the section "iPACOSE".

Partial correlation and Gaussian Graphical Models

This section briefly reviews the basics of GGM theory used in this paper. Let X denote a p-variate random vector $X=(X_1,\ldots,X_p)^{\mathrm{T}}$ such that variables $X_1,\ldots,X_p$ all have a mean and a variance. $\mathcal{G}$ denotes the graph describing the conditional independencies between the p variables: $\mathcal{G}$ is thus an undirected graph with p nodes. The covariance matrix of X, denoted by Σ, is supposed to be invertible. Its inverse $\Omega=\Sigma^{-1}$ is from now on referred to as the *precision matrix*.

The partial correlation coefficient ρ_{ij} of X_i and X_j given all the other variables $\{X_1,\ldots,X_p\}\backslash\{X_i,X_j\}$ can be estimated as

$$\widehat{\rho}_{ij}=\frac{\widehat{cov}\left(X_i-\widehat{X}_i,X_j-\widehat{X}_j\right)}{\sqrt{\widehat{var}\left(X_i-\widehat{X}_i\right)\widehat{var}\left(X_j-\widehat{X}_j\right)}}, \quad (1)$$

where $\widehat{cov}$ and $\widehat{var}$ denote the empirical covariance and variance, respectively, and $\widehat{X}_i$ stands for the fitted value of X_i in a linear regression model including all other variables except X_j as covariates. In a few words, $\widehat{\rho}_{ij}$ is the correlation of the residuals of the linear models regressing X_i against all variables except X_j and vice-versa.

Another method to compute $\widehat{\rho}_{ij}$ based on linear regressions results from the following property [6]:

$$\widehat{\rho}_{ij}=sign(\widehat{\beta}_{ij})\sqrt{\widehat{\beta}_{ij}\widehat{\beta}_{ji}}, \quad (2)$$

where $\widehat{\beta}_{ij}$ is the estimated coefficient of variable X_j in the linear model regressing X_i against all the other variables. Note that both formulations (1) and (2) implicitly assume that the considered linear regression models can be estimated, which is for instance not the case in high-dimensional data with $n<p$. This issue will be discussed later. Moreover, it can also be shown [6] that the partial

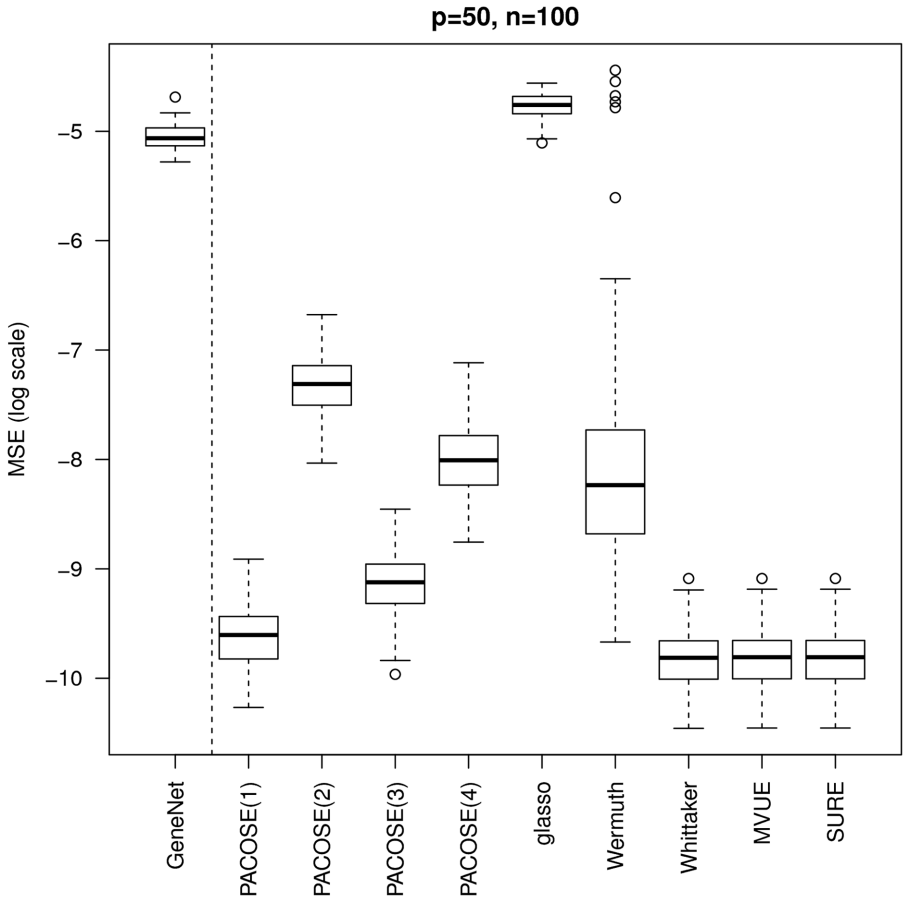

Figure 2. MSE of the partial correlation matrix estimates. $p=50$ and $n=100$, when the graphs are decomposable.
doi:10.1371/journal.pone.0060536.g002

correlation coefficient ρ_{ij} is related to the precision matrix $\Omega=[\omega_{ij}]=\Sigma^{-1}$ as follows:

$$\rho_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}}\sqrt{\omega_{jj}}}, \text{ for } i \neq j. \quad (3)$$

If $X_1,\ldots,X_p$ are Gaussian, the following important property can be shown for $i,j,k \in \{1,\ldots,p\}$ $(k \neq i,j)$, , see for instance [10]:

$$X_i \perp\!\!\!\perp X_j | X_k \Leftrightarrow \rho_{ij}=0, \quad (4)$$

which means that two variables are conditionally independent if and only if their partial correlation equals zero.

The formulation (4) is exploited by numerous methods to estimate gene regulatory networks from high-dimensional microarray gene expression data [7,11,12]. Note, however, that these data often have much more variables (genes) than observations (arrays), hence the term high-dimensional data''. A regularized regression technique has then to be used to estimate β_{ij} and β_{ji}, since least squares regression cannot be performed with $n<p$ data. Another popular approach [3] to estimate GGMs from high-dimensional data consists in applying Eq. (3) using a regularized (invertible) estimator of Σ.

All these methods yield an estimate of the partial correlation matrix. Some methods are essentially sparse, *i.e.* yield a matrix with many zeros [11]. In this case, the graph is simply derived from the partial correlation matrix by connecting pairs of variables with non-zero partial correlations. For other methods [3,7], however, a threshold has to be applied to decide which variables have to be connected.

PACOSE

The concepts briefly reviewed in the above section are important for understanding our novel method – PACOSE -, whose main idea is to combine formulation (2) along with the information given in an a priori independence graph $\mathcal{G}$ between the variables. This is done by setting β_{ij} and β_{ji} to 0 if X_i and X_j are not connected in the graph $\mathcal{G}$. It immediately results from Eq. (2) that $\widehat{\rho}_{ij}=0$.

Setting β_{ij} to 0 impacts the whole linear model

$$X_i = \sum_k \beta_{ik} X_k,$$

since it essentially removes one covariate in the regression model. As a consequence, the estimation of other partial correlation coefficients ρ_{ik} involving X_i and any other variable $X_k, k \neq j$ is also affected.

More precisely, our graph-constrained estimator of the partial correlation between X_i and X_j is given as

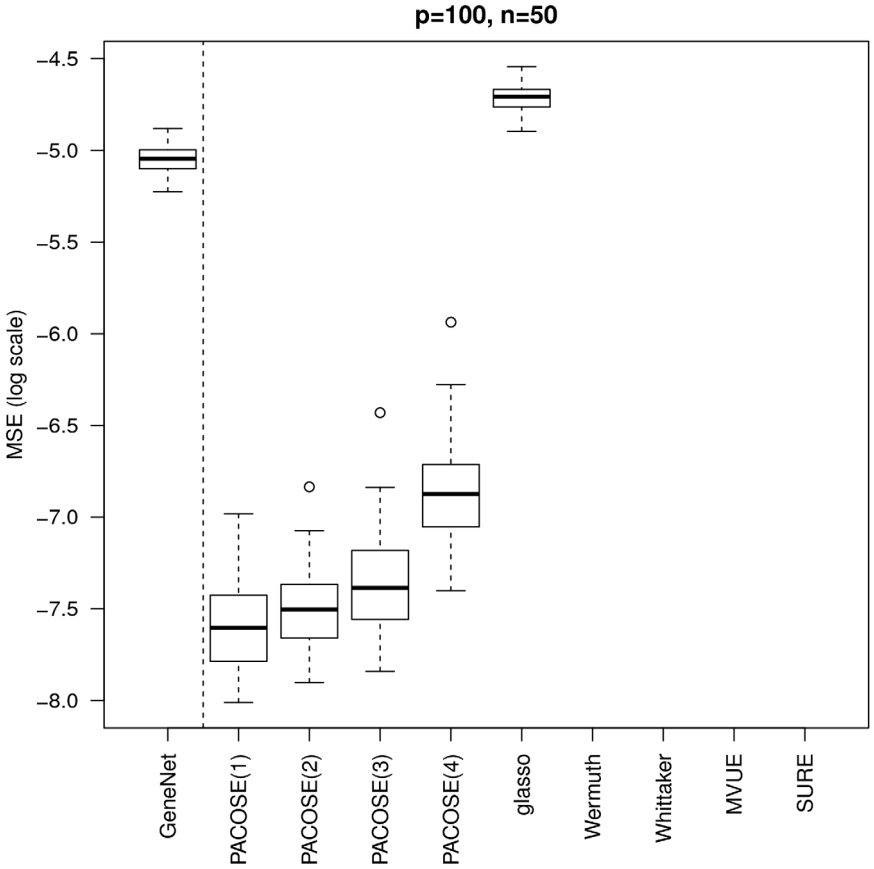

Figure 3. MSE of the partial correlation matrix estimates. $p=100$ and $n=50$, when the graphs are not decomposable. Since the graphs are not decomposable, the estimators MVUE and SURE are not applicable. Wermuth's algorithm does not converge, and the implementation of Whittaker's method requires a decomposition of the graph into cliques.
doi:10.1371/journal.pone.0060536.g003

$$\widehat{\rho}_{ij}^{\mathcal{G}} = sign(\widehat{\beta}_{ij}^{\mathcal{G}})\sqrt{\widehat{\beta}_{ij}^{\mathcal{G}}\widehat{\beta}_{ji}^{\mathcal{G}}}, \tag{5}$$

where

- $\widehat{\beta}_{ij}^{\mathcal{G}}=0$ if X_i and X_j are not connected in $\mathcal{G}$,
- $\widehat{\beta}_{ij}^{\mathcal{G}}$ is the estimated regression coefficient of X_j in the regression of X_i against its connected variables if X_i and X_j are connected, *i.e.* the estimate of coefficient $\beta_{ij}^{\mathcal{G}}$ in the linear regression model

$$X_i = \beta_{i0}^{\mathcal{G}} + \sum_{k:\, k\sim i} \beta_{ik}^{\mathcal{G}} X_k + \epsilon_i, \tag{6}$$

where $k\sim i$ means that variables k and i are connected in $\mathcal{G}$.

This definition implicitly assumes that the estimates of the regression coefficients exist, which may not be the case in high-dimensional settings. This problem is addressed in the next section.

Table 1. Prediction nomenclature in the context of graph inference.

	$i\sim j$	$i\not\sim j$
$p_{ij}\neq 0$	TP	FP
$p_{ij}\neq 0$	FN	TN

The definitions of true and false positives (resp. TP and FP), true and false negatives (resp. TN and FN) in the context of graph inference.
doi:10.1371/journal.pone.0060536.t001

High dimensional settings

When the number of variables connected to i is greater than the number of observations, the estimation of the coefficients of the linear regression model (6) cannot be performed by ordinary least squares. Unfortunately, it is likely to sometimes occur in practical analyses with high-dimensional data. That is why we suggest to replace least squares regression by one of its regularized versions: ridge regression [13], PLS regression [14,15], Lasso [16] or adaptive Lasso [17]. The regularization parameters are estimated by k-fold cross-validation (CV). Once the partial correlation coefficients are estimated, an estimator of the partial correlation matrix Π is obtained via Eq. (2).

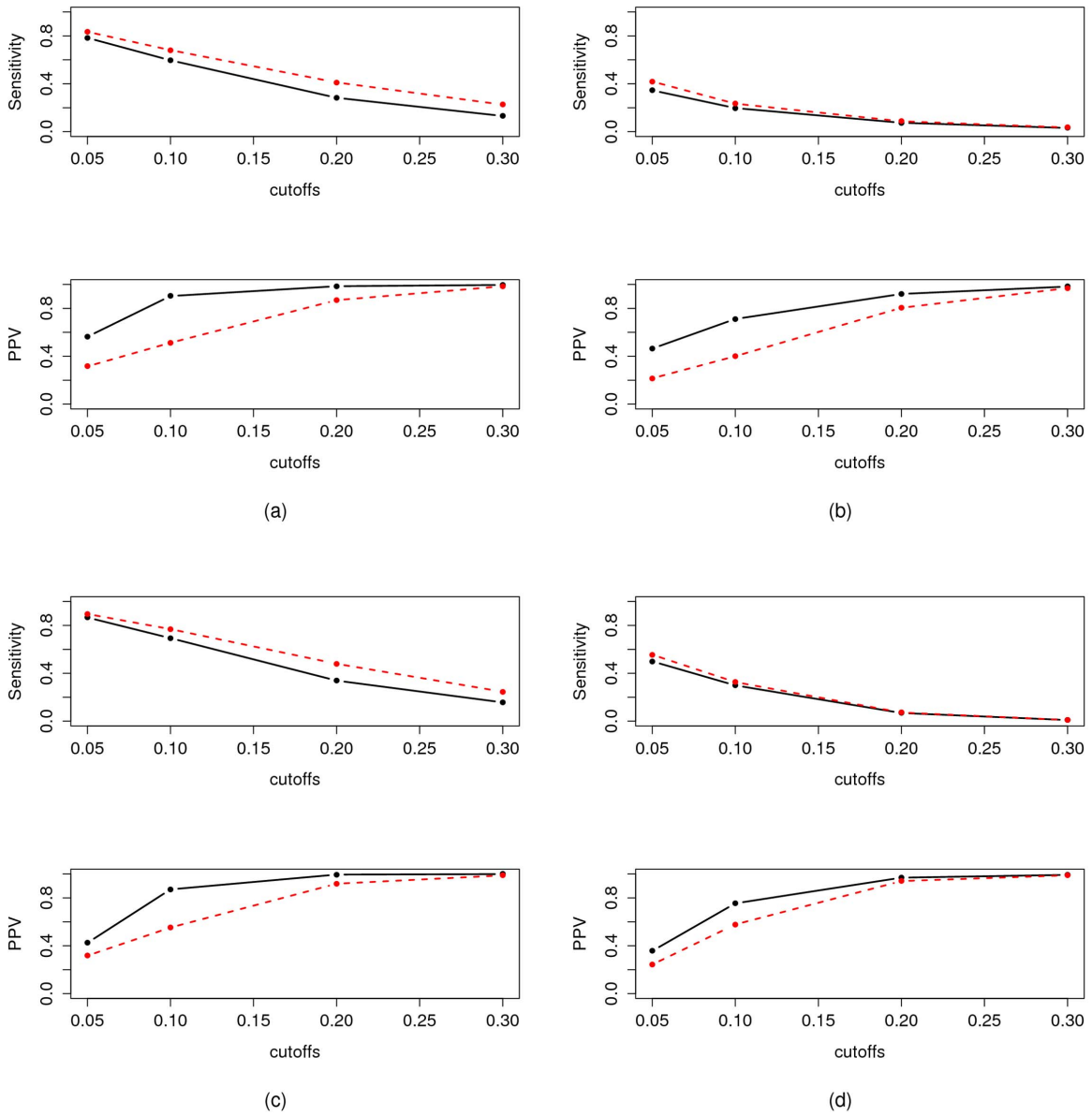

Figure 4. Performance of iPACOSE (black straight lines) when compared to its regression based GGM estimate counterpart (red dashed lines). *(a) and (b)*: performance of the PLS version of iPACOSE. *sen* and *ppv* for $p=100$ and $n=50$ (a) and for $p=50$ and $n=100$ (b). Thresholds: 0.05, 0.1, 0.2 and 0.3. The results of iPACOSE are represented by the black line and the results of the pls.net function with the red dashed line. *UPPER FIGURE*: sensitivity as a function of the threshold, *LOWER FIGURE*: PPV as a function of the threshold. *(c) and (d)*: performance of the Ridge version of iPACOSE. *sen* and *ppv* for $p=100$ and $n=50$ (c) and for $p=50$ and $n=100$ (d). Thresholds: 0.05, 0.1, 0.2 and 0.3. The results of iPACOSE are represented by the black line and the results of the ridge.net function with the red dashed line. *UPPER FIGURE*: sensitivity as a function of the threshold, *LOWER FIGURE*: PPV as a function of the threshold.
doi:10.1371/journal.pone.0060536.g004

Competing approaches

To our knowledge, there is no method in the literature allowing to compute directly the partial correlation matrix with the knowledge of an undirected graph. But there are numerous methods dedicated to the estimation of the inverse covariance matrix knowing a given graph. The literature refers to these methods as covariance selection algorithms. These algorithms are usually used to estimate the covariance matrix, but they can also be used to estimate the precision matrix.

When the graph is decomposable, the covariance matrix can be estimated by maximum likelihood. Alternative methods have been proposed such as the shrinkage estimator designed by Wiesel *et al.* [10]. However, these methods are not able to cope with a non-decomposable graph. This is a major drawback in practice because most of the graphs relevant to bioinformatics are non-decomposable. One thus has to turn to iterative methods [6,18] or methods such as "glasso" [11] based on the optimization of a criterion independently from the nature of the graph.

All the covariance selection methods we refer to in this section compute directly the precision matrix, and not the partial correlation matrix as PACOSE does. In order to compare PACOSE to them, we use Eq. (3) to transform any estimated precision matrix into a partial correlation matrix.

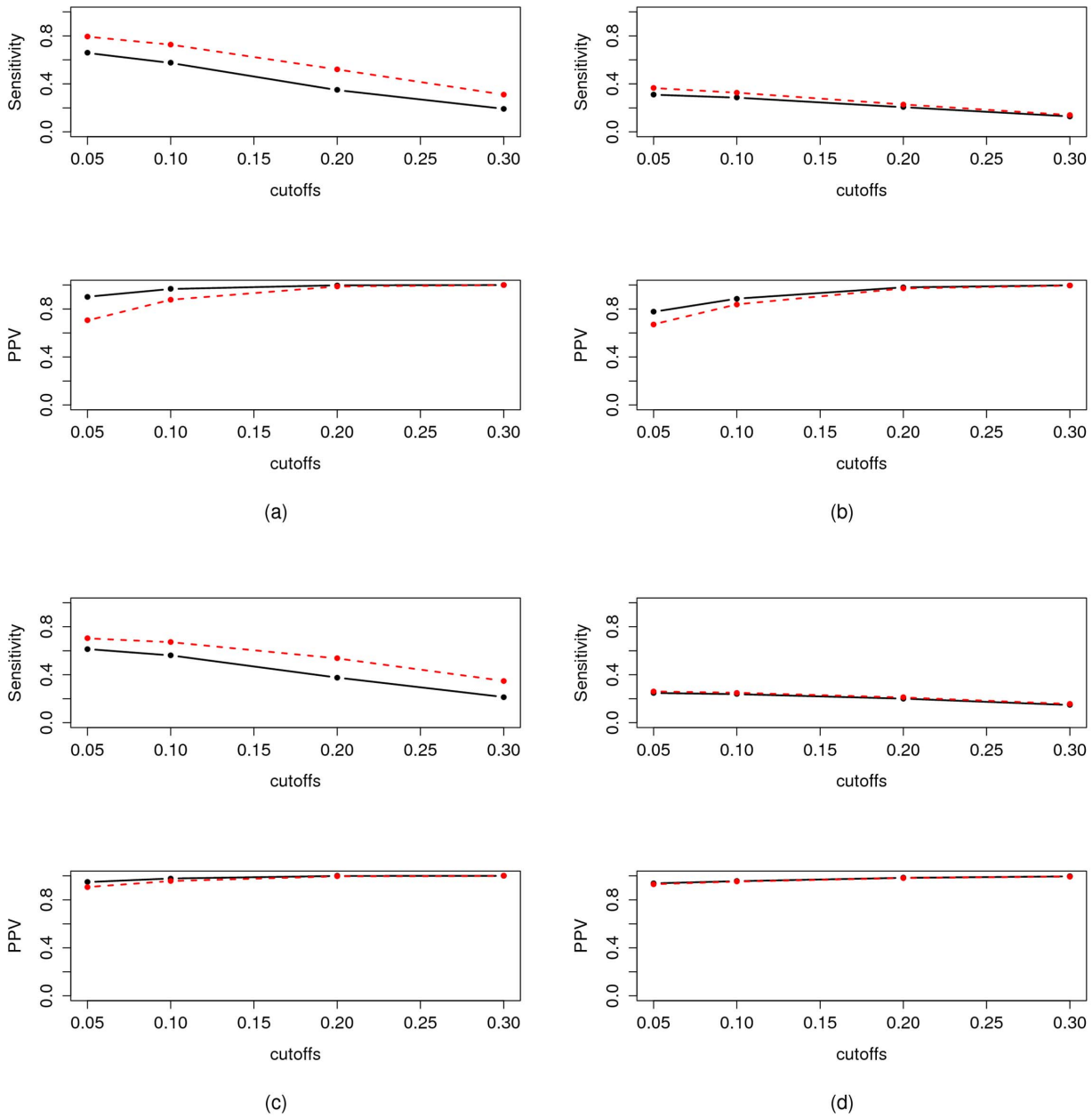

Figure 5. Performance of iPACOSE (black straight lines) when compared to its regression based GGM estimate counterpart (red dashed lines). *(a) and (b)*: performance of the LASSO version of iPACOSE. *sen* and *ppv* for $p=100$ and $n=50$ (a) and for $p=50$ and $n=100$ (b). Thresholds: 0.05, 0.1, 0.2 and 0.3. The results of iPACOSE are represented by the black line and the results of the adalasso.net function with the red dashed line. *UPPER FIGURE*: sensitivity as a function of the threshold, *LOWER FIGURE*: PPV as a function of the threshold. *(c) and (d)*: performance of the adaptive LASSO version of iPACOSE. *sen* and *ppv* for $p=100$ and $n=50$ (c) and for $p=50$ and $n=100$ (d). Thresholds: 0.05, 0.1, 0.2 and 0.3. The results of iPACOSE are represented by the black line and the results of the adalasso.net function with the red dashed line. *UPPER FIGURE*: sensitivity as a function of the threshold, *LOWER FIGURE*: PPV as a function of the threshold.
doi:10.1371/journal.pone.0060536.g005

iPACOSE

When estimating an independence graph from raw data with partial correlation matrices, one usually first estimates the partial correlation matrix and then applies to it a certain threshold, allowing to eliminate small coefficients. The obtained sparse matrix is then considered as the adjacency matrix of the underlying graph, following the principle of GGM.

The idea of the iPACOSE (as in "iterated PArtial COrrelation SElection") algorithm is the following: rather than stopping after this first estimation of the underlying graph, we use this graph as an input for PACOSE, then allowing a re-estimation of the partial correlation coefficients. Since the newly estimated coefficients are likely to become smaller than the given threshold, a new graph can be estimated from this new partial correlation matrix by thresholding it, and so on. With this iterated process, we aim to estimate the coefficients close to the threshold more accurately and then eliminate as many false positive edges as possible.

More precisely, our algorithm iPACOSE takes a data matrix, a threshold and a graph (called $\mathcal{G}^{(0)}$) as inputs and operates as follows:

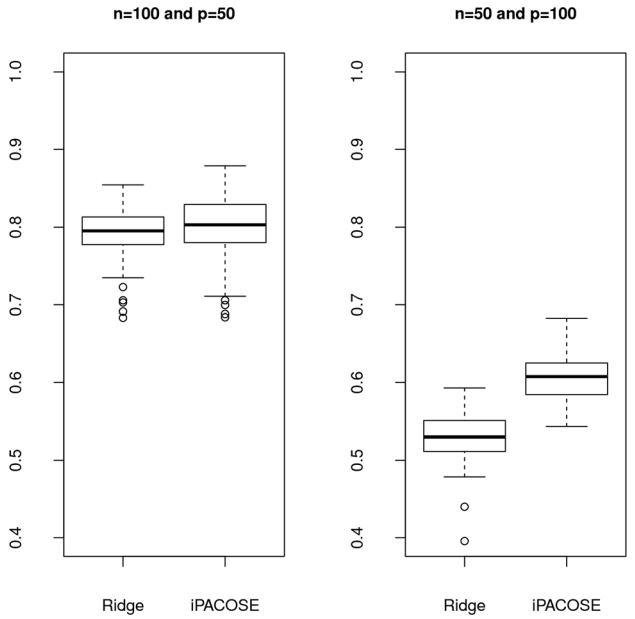

Figure 6. Measure of the stability with Fleiss' κ for the methods ridge.net and the Ridge version of iPACOSE. *LEFT FIGURE*: $n=100$ and $p=50$. *RIGHT FIGURE*: $n=50$ and $p=100$. The regularization parameter of the ridge regression is determined analytically [26] for both methods.
doi:10.1371/journal.pone.0060536.g006

1. Apply PACOSE with the dataset and $\mathcal{G}^{(0)}$ as arguments.
2. Transform the estimated partial correlation matrix into a graph by applying the threshold.
3. Apply PACOSE to the dataset with the graph derived in 2 in order to estimate a new partial correlation matrix.
4. Iterate steps 2 and 3 until the graph does not change anymore.

$\mathcal{G}^{(0)}$ can be estimated with any existing method, such as pcor.shrink from the R package GeneNet [3] or ridge.net, pls.net, adalasso.net or lasso.net from the R package parcor [7].

Results

In this section we present a simulation study for the evaluation of

(a) PACOSE as an estimation procedure for the partial correlation matrix given a fixed undirected graph,

(b) iPACOSE as a procedure for graph estimation, in combination with standard GGM estimation procedures.

When one wants to simulate data knowing a given graph of independence, there is the possibility of using the theory of GGM, more particularly through the constraint (4). Furthermore, the randomly generated precision matrix has to be positive definite. One could see this problem as a so-called "positive definite completion matrix" issue [19]. But the work on this specific issue is once again mainly focused on decomposable graphs. We adopt a more empirical method, which in practice gives a very satisfying range of partial correlation coefficients, and at the end of the algorithm, the fulfillment of constraint (4).

Simulated data

We use simulated data to compare our method to the methods presented in the literature. Erdös-Rényi [20] or Barabasi [21] graphs are used to model the interactions between genes, which allows loops, hubs, and multiple connected components. We use the following algorithm:

(i) Compute a first random Erdös-Rényi [20] (if we want a non-decomposable graph) or Barabasi [21] (if we want a decomposable graph) graph $\mathcal{G}^{(\text{init})}$,

(ii) Get the "upper triangular" adjacency matrix $A^{(\text{init})}$ of this graph and replace any non null coefficient by a random realization of a uniform variable (*e.g.* $\mathcal{U}(]-1,-0.8]\cup[0.8,1[)$, but any interval is possible), which then allows to define an upper triangular weight matrix $W^{(\text{init})}$,

(iii) Compute the following matrix $M=(W^{(\text{init})}+I)^T(W^{(\text{init})}+I)$, where I is the identity matrix, defining a new graph $\mathcal{G}$ slightly different from the

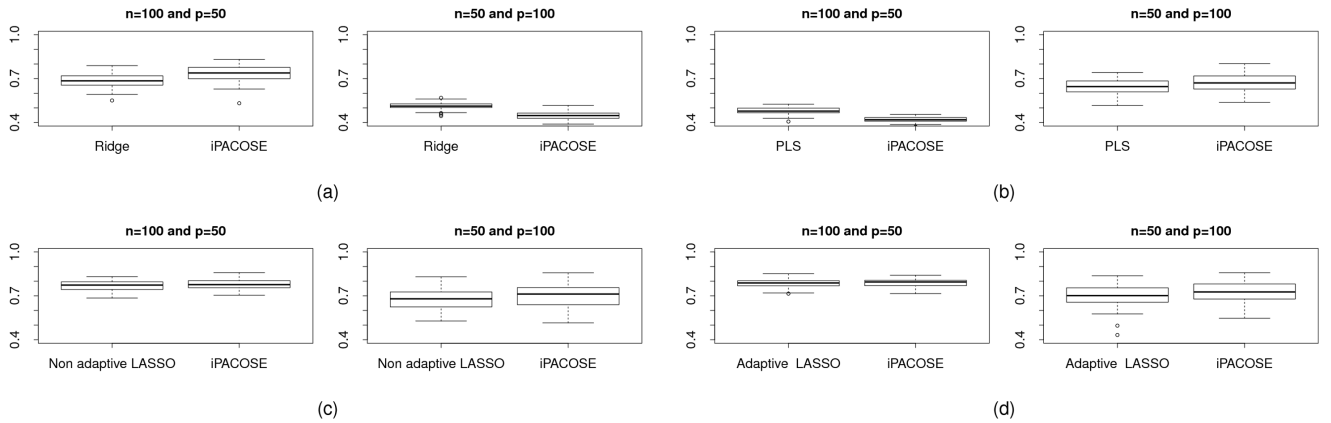

Figure 7. Measure of the stability with Fleiss' κ for (a) ridge.net and the Ridge version of iPACOSE, (b) pls.net and the PLS version of iPACOSE, (c) the non adaptive version of the method adalasso.net and the LASSO version of iPACOSE and (d) the adaptive version of the method adalasso.net and the adaptive LASSO version of iPACOSE. For each one of the couples of figures, the *LEFT FIGURE* corresponds to $n=100$ and $p=50$ and the *RIGHT FIGURE* to $n=50$ and $p=100$. The regularization parameters of the Ridge, PLS, LASSO and adaptive LASSO regressions are determined analytically via a 5-fold cross-validation.
doi:10.1371/journal.pone.0060536.g007

initial graph, but above all defining a sparse positive definite matrix M,

(iv) Normalize this matrix to get a partial correlation matrix $\Pi = M^{\star}$,

(v) Generate the dataset from the multivariate Gaussian distribution $X \sim \mathcal{N}(0, \Sigma = \Pi^{-1})$.

We prefer this algorithm to *e.g.* the algorithm presented in Verzelen *et al.* [22] and Krämer *et al.* [7] because the latter produces partial correlation coefficients often very close to 0 when p is greater than a few dozens. The drawback of this method is that it alters the degree distribution of the initial graph structure – in a drastic way for Erdös-Rényi graphs, and in a very moderate way for Barabasi graphs.

We implemented the covariance selection algorithm presented in [18] in R and C, and the minimum variance unbiased estimator (MVUE) and the Stein unbiased risk estimator (SURE) [10] in R. Whittaker's method [6] is implemented in the R package ggm, and Friedman's *et al.* method [11] in the package glasso.

Estimation of the partial correlation matrix with PACOSE

We compare PACOSE to the competing methods presented above based on the mean square error (MSE) between the estimated partial correlation matrix (denoted $\widehat{\Pi}$) and the real one (denoted Π), as defined by

$$MSE(\widehat{\Pi}) = \frac{1}{N} \sum_{ij} \left(\widehat{\Pi}_{ij} - \Pi_{ij} \right)^2,$$

with $N = p(p-1)/2$.

The following notations are used for the competing approaches: GeneNet [3], glasso [11], MVUE, SURE [10], Whittaker [6], wermuth [18]. It has to be noted that GeneNet does not take into account the information in the given graph: it is considered in our results as a reference method giving an upper bound on the MSE.

These methods are compared to the PACOSE algorithm, where four different regularized regression methods are used to estimate the coefficients in Eq. (2):

- Ridge regression (PACOSE(1)),
- PLS regression (PACOSE(2)),
- LASSO regression (PACOSE(3)),
- adaptive LASSO regression (PACOSE(4)).

All the regularization parameters are estimated with 10-fold cross-validation. The graph used within PACOSE is the real independence graph, which is known since we work on simulated data.

When there are more individuals than variables, and when the considered graphs are decomposable, we can see on Figure 2 that the SURE estimator performs better than all the others methods.

When the setting is less favorable, *i.e.* when there are less individuals than variables and the graphs are not decomposable, the results show a better performance of our estimator, both in terms of stability and accuracy, see Figure 3, especially for the PLS and Ridge regressions. This is a very promising result for PACOSE, since in reality the considered graphs are very unlikely to be decomposable, and the number of variables is generally bigger than the number of individuals. In both Figures 2 and 3, method GeneNet performs poorly, which is due to the fact that it does not consider the underlying graph. This method acts as a baseline representing the methods estimating the partial correlation matrix without any prior knowledge.

The underlying graphs of independence are not precisely known for biological data, estimating them being even a burning issue in bioinformatics. We show in the following that PACOSE can be advantageously integrated into the estimation of GGMs, yielding potential improvements in terms of estimation accuracy.

Estimation of independence graphs with iPACOSE

In this section, we apply iPACOSE to simulated datasets in order to recover partial independence graphs. Our goal is to compare the four different network inference methods: ridge.net, pls.net, the non-adaptive version of adalasso.net and the adaptive version of adalasso.net based on the estimation of the partial correlation matrix, to their iterative versions iPACOSE(1), (2), (3), (4), respectively.

To compare the estimated graphs with the real graph, we use the positive predictive value (PPV, denoted *ppv*) and the sensitivity (denoted *sen*):

$$ppv = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \text{ and } sen = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

where TP, FP and FN are defined in Table 1. Biological networks are indeed often described as sparse, and indicators based on the number of edges are more suitable in this case [23].

The sensitivity and the PPV of the estimated graphs as a function of the threshold are represented on Figures 4 and 5. Two different settings are considered for these simulations: $p = 100$ and $n = 50$ for Figures 4(a), 4(c), 5(a), 5(c), and $p = 50$ and $n = 100$ for Figures 4(b), 4(d), 5(b), 5(d). The key chacteristic of iPACOSE is that it allows to estimate networks with less edges without eliminating too many correct interactions. We can indeed observe on Figures 4(a)–(d) that the PPV, *i.e.* the capacity to estimate sparse networks, is improved when compared to ridge.net or pls.net. On the other hand, when applied to adalasso.net, a method estimating particularly sparse networks, there is no detectable improvement in PPV – see Figures 5(a)–(d).

In other words, Figures 4 and 5 compare the sensitivity and PPV of $\mathcal{G}^{(0)}$ for a given threshold to the sensitivity and PPV of iPACOSE with the same threshold: iPACOSE has a real interest when there is room for improvement in $\mathcal{G}^{(0)}$'s PPV.

Stability

In practical data analyses, the true network is almost always unknown, which makes the evaluation of graph inference methods so difficult on real data. For our particular application, we choose not to assess the performance of iPACOSE by comparing the obtained networks with interactions found in publicly available databases, but rather to evaluate its stability. A stable algorithm is robust against small perturbations of the dataset, see the work of Krämer *et al.* [7] or Varoquaux *et al.* [24] for an example in brain imaging. In our study, the considered datasets are split into 10 groups and GGMs are inferred based on datasets obtained by excluding each of the 10 groups successively. The 10 obtained networks are compared using Fleiss' κ, following the procedure described in Krämer *et al.* [7]. Fleiss' κ is originally designed to measure the degree of agreement between more than two raters. Each rater attributes a grade to an individual: in our case, a rater is a network inference method and a grade is 0 or 1, meaning that an interaction is considered as significant or not. The resulting statistic is always lower than 1 and, the closer the 10 networks, the closer it gets to 1. For a short description of this measure of agreement, see [25] (pp. 256–258).

We first measure the stability of iPACOSE, and compare it to the stability of ridge.net on simulated data. According to Krämer *et al.* [7], ridge.net and the other methods presented in this paper do not show good stability performance. In order to stabilize the method, we replace the determination of the optimal ridge regularization parameter through a cross-validation approach by an analytic determination [26]. The results are shown on Figure 6. We observe that iPACOSE stabilizes the inference of the network. Figures 7(a), 7(b), 7(c) and 7(d) show the same type of stability results, except for the determination of the regularization parameters, which is done with a 5-fold cross validation. Stability is not improved with iPACOSE when it is low with the original inference method. However, when the stability is at a high level, it either remains at the same level or is improved.

Application to a real dataset

This first very positive result still holds for the comparison of the stability of ridge.net and iPACOSE on a real dataset. For this application, we use the real data presented in [27] and further described and used in [28] consisting in $n=310$ amino acid sequences on which were measured $p=104$ different physical properties. The regularization parameters are determined analytically [26] for both methods. Fleiss's κ is computed on this dataset in a 10-fold fashion and is equal to 0.70 for the 10 networks obtained with ridge.net and to 0.85 for the 10 networks obtained with iPACOSE, which is an even higher improvement than in simulated data.

Replacing the 10-fold approach by a 5-fold does not essentially change the results (data not shown), which conforts us in the fact that our results are not depending too strongly on the number of parts the dataset is split into.

Discussion

In this article, we presented PACOSE, a simple method to estimate a partial correlation matrix under the constraint that some known coefficients are null. We also presented iPACOSE, an original procedure to apply PACOSE iteratively within the estimation of independence graphs in combination with any GGM estimation method.

Our results on simulated data suggest that PACOSE's performance is very promising when the known graph describing the sparse structure of the partial correlation matrix is non-decomposable and $n<p$. Since those two characteristics are met when dealing with biological data, our method is all the more interesting.

Having in mind the field of biological data as an application, we designed iPACOSE, an application of PACOSE to the estimation of independence graphs. iPACOSE is a method designed to improve the performance of the graph estimation algorithms based on the estimation of the partial correlation matrix. Results on simulated data show that iPACOSE manages to increase the positive predictive value of the inferred graphs while still showing good sensitivity. Moreover, results on simulated data and confirmed on real world data show that iPACOSE has very interesting stability properties. As a perspective of this work, iPACOSE would provide candidate interactions to work on more elaborate models, such as *e.g.* non linear ordinary differential equations applied to transcriptomic data [29] or used in cancer studies [30]. Such models would both help the discussion with the biologist or the phycisian by providing more elaborate interaction models between genes, and help in the design of "on the bench" experiments for the validation of the interactions found by iPACOSE.

Acknowledgments

We thank the reviewers for their comments and suggestions.

Author Contributions

Analyzed the data: VG AB. Wrote the paper: VG ALB.

References

1. Witten DM, Tibshirani R (2009) Covariance-regularized regression and classification for highdimensional problems. Journal of the Royal Statistics Society Series B 71: 615–636.
2. Friedman J (1989) Regularized discriminant analysis. Journal of the American Statististical Association 84: 165–175.
3. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4: Issue 1, Article 32.
4. Dempster AP (1972) Covariance selection. Biometrics 28: 157–175.
5. Speed TP, Kiiveri HT (1986) Gaussian Markov distributions over finite graphs. The Annals of Statistics 14: 138–150.
6. Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley.
7. Krämer N, Schäfer J, Boulesteix AL (2009) Regularized estimation of large scale gene association networks using gaussian graphical models. BMC Bioinformatics 10: 384.
8. D'Aspremont A, Banerjee O, El Ghaoui L (2008) First-order methods for sparse covariance selection. SIAM Journal on Matrix Analysis and Applications 30: 56–66.
9. Krishnamurthy V, d'Aspremont A (2009) A pathwise algorithm for covariance selection. In: OPT 2009: 2nd NIPS Workshop on Optimization for Machine Learning. MIT Press.
10. Wiesel A, Eldar YC, Hero AO (2010) Covariance estimation in decomposable gaussian graphical models. IEEE Transactions on Signal Processing 58: 1482–1492.
11. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432–441.
12. Tenenhaus A, Guillemot V, Gidrol X, Frouin V (2010) Gene association networks from microarray data using a regularized estimation of partial correlation based on PLS regression. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7: 251–262.
13. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12: 55–77.
14. Wold H (1975) Path models with latent variables: the NIPALS approach, in: H. M. Blalock (Ed.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building. New york: Academic Press.
15. Wold S, Ruhe A, Wold H, Dunn WJ, III (1984) The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing 5: 735–743.
16. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistics Society Series B 58: 267–288.
17. Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101: 1418–1429.
18. Wermuth N, Scheidt E (1977) Fitting a covariance selection model to a matrix, algorithm 105. Journal of the Royal Statistical Society C 26: 88–92.
19. Grone R, Johnson CR, Sá EM, Wolkowicz H (1984) Positive definite completions of partial hermitian matrices. Linear Algebra and its Applications 58: 109–124.
20. Erdős P, Rényi A (1959) On random graphs. Publicationes Mathematicæ 6: 290–297.
21. Barabasi Albert (1999) Emergence of scaling in random networks. Science 286: 509–512.
22. Verzelen N, Villers F (2009) Tests for gaussian graphical models. Computational Statistics & Data Analysis 53: 1894–1905.
23. Vert JP (2008) Reconstruction of biological networks by supervised machine learning approaches. Technical report, Mines ParisTech, Centre for Computational Biology.
24. Varoquaux G, Sadaghiani S, Pinel P, Kleinschmidt A, Poline JB, et al. (2010) A group model for stable multi-subject ICA on fMRI datasets. Neuroimage 51: 288–299.
25. Fisher LD, van Belle G (1993) Biostatistics: A Methodology For the Health Sciences. John Wiley & sons, Inc.
26. Hoerl AE, Kannard RW, Baldwin KF (1975) Ridge regression: some simulations. Communications in Statistics – Simulation and Computation 4: 105–123.
27. Segal MR, Cummings MP, Hubbard AE (2001) Relating amino acid sequence to phenotype: analysis of peptide-binding data. Biometrics 57: 632–642.

28. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9: 307.
29. Quach M, Brunel N, d'Alch Buc F (2007) Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. Bioinformatics 23: 3209–3216.
30. Bianca C, Chiacchio F, Pappalardo F, Pennisi M (2012) Mathematical modeling of the immune system recognition to mammary carcinoma antigen. BMC Bioinformatics 13 Suppl 17: S21.