Avagyan, V., Alonso, A. M. & Nogales, F. J. (2018). D-trace estimation of a precision matrix using adaptive Lasso penalties. *Advances in Data Analysis and Classification, 12*(2), pp. 425–447.

DOI: 10.1007/s11634-016-0272-8

D-trace Precision Matrix Estimation Using Adaptive Lasso Penalties

Vahe AVAGYAN, Andrés M. ALONSO, and Francisco J. NOGALES

Abstract

An accurate estimation of a precision matrix has a crucial role in the current age of high-dimensional data explosion. To deal with this problem, one of the prominent and commonly used techniques is the ℓ_1 norm (Lasso) penalization for a given loss function. This approach guarantees the sparsity of the precision matrix estimator for properly selected penalty parameters. However, the ℓ_1 norm penalization often fails to control the bias of the obtained estimator because of its overestimation behavior. In this paper, we introduce two adaptive extensions of the recently proposed ℓ_1 norm penalized D-trace loss minimization method. The proposed approaches intend to diminish the produced bias in the estimator. Extensive numerical results, using both simulated and real datasets, show the advantage of our proposed estimators.

Key Words: Adaptive Thresholding; D-trace loss; Gaussian Graphical Model; Gene expression data ; High-dimensionality

---

E-mail addresses: vahe.avagyan@uc3m.es (Vahe Avagyan), andres.alonso@uc3m.es (Andrés M. Alonso), fcojavier.nogales@uc3m.es (Francisco J. Nogales).

1 Introduction

The estimation of inverse covariance matrix (also known as precision matrix) is an important problem in various research fields and methodologies. In the recent decade, the high-dimensional precision matrix has attracted a growing interest due to the massive flow of large datasets spanning several scientific areas. An accurate estimate of a precision matrix has a fundamental role in discriminant analysis, forecasting and several other statistical methodologies (Mardia et al. 1979; McLachlan 2004). One of the applications involving a proper and stable precision matrix estimate is the computation of the optimal portfolios for a large number of assets (Frahm and Memmel 2010; Goto and Xu 2013).

The exceeding attractiveness of the precision matrix estimation emerges under the assumption of multivariate normality of data. This statement is initially formulated by Dempster (1972). It is well known that when the data follow a Gaussian distribution, the zero entries (i, j) of the precision matrix indicate the conditional independence between the variables i and j, given all the other variables (Lauritzen 1996). More specifically, under the normality assumption, the precision matrix represents the statistical dependency among the variables. Therefore, the precision matrix is closely related to the Gaussian Graphical Models (GGM), which is a prominent framework for representing the structure of the dependencies between normally distributed variables (Whittaker 1990). There are several applications involving a sparse precision matrix such as the estimation of genetic interaction networks through high-dimensional gene expression data (Stifanelli et al. 2013; Yin and Li 2013),

brain connectivity networks through neuroimaging techniques (Huang et al. 2010; Ryali et al. 2012), climate networks (Zerenner et al. 2014), etc.

Without loss of generality, we assume that X is a $n \times p$ mean-centered sample data matrix. Each row $X_i = (X_{i1}, ..., X_{ip})$ is a realization of a p-variate random vector, independent and identically distributed for $i = 1, ..., n$, and has an unknown $p \times p$ covariance matrix Σ with the corresponding precision matrix $\Omega = \Sigma^{-1}$.

Substantial research exists related to the precision matrix estimation. The most ordinary and classical precision matrix estimator is the inverse of the sample covariance matrix S. Although the sample covariance matrix is an unbiased estimator of the covariance matrix, its inverse S^{-1} contains a considerable bias.[1] Moreover, when $p/n > 1$, the matrix S becomes singular and, therefore, the classical estimator S^{-1} does not exist.

A straightforward approach is to invert a well-estimated covariance matrix, which is known as a two-step or indirect estimation. In this way, several estimators of the covariance and the correlation matrices have been provided with good practical and theoretical properties. Among the most popular ones are the shrinkage estimators (e.g., Ledoit and Wolf 2004; Schafer and Strimmer 2005; Warton 2008; Touloumis 2015), robust estimators (e.g., Nguyen and Welsch 2010), estimators based

[1] When $n > p$, it is known that $E(S^{-1}) - \Omega = \frac{p+2}{n-p-2}\Omega$. Therefore, the inverse of the sample covariance matrix is highly unstable when the ratio $\frac{p}{n}$ increases. For instance, when $p = n/2 - 2$, then $E(S^{-1}) - \Omega = \Omega$, therefore, the bias of the classical estimator S^{-1} has the same magnitude as Ω. On the other hand, in high-dimensional settings, the eigenvalues of the sample covariance matrix are widespread and the largest eigenvalues reach to extreme values, which makes the condition number of S very large (Johnstone 2001).

on thresholding, banding or tapering procedures (e.g., Bickel and Levina 2008; El Karoui 2008; Cai and Yuan 2012; Wang and Daniels 2014) and those based on convex optimization frameworks (e.g., Rothman 2012; Xue et al. 2012; Deng and Tsui 2013; Cui et al. 2014). However, the two-step estimators may not be optimal (Ledoit and Wolf 2012) and, moreover, the two-step approach does not guarantee the sparsity of the precision matrix estimator.

In essence, shrinkage techniques can also be applied for estimating the precision matrix, i.e., different linear combinations between the matrix S^{-1} and a selected target matrix (see, for instance, Haff 1980; Frahm and Memmel 2010; Kourtis et al. 2012). However, as explained above, these approaches can be used only when $p << n$.

To overcome the computational challenges and to deal with the situation of $p > n$, prior research proposed several precision matrix estimators based on a convex optimization framework. To address the sparsity requirement of the matrix and to attain an accurate precision estimator, the Lasso or ℓ_1 regularization can be applied. Originally, Tibshirani (1996) introduced this framework in the regression framework. Banerjee et al. (2008) proposed the ℓ_1 penalized log-likelihood function maximization approach which is one of the remarkable estimations and known in the literature as Graphical Lasso or, simply, glasso method. Prior work studied the ℓ_1 penalized log-likelihood function maximization approach (e.g., Yuan and Lin 2007; d'Aspremont et al. 2008; Banerjee et al. 2008; Rothman et al. 2008; Yin and Li 2013) and several algorithms have been developed to solve the regularization problem efficiently (e.g., Friedman et al. 2008; Duchi et al. 2008; Scheinberg et al. 2010). Moreover,

some scholars proposed approaches to improve the performance of the glasso method through adaptive Lasso and SCAD (Smoothly Clipped Absolute Deviation) penalties (see Fan et al. 2009) or through additional trace norm penalty (see Maurya 2014). Others proposed procedures that efficiently speed-up the algorithms for solving the glasso problem (Witten et al. 2011). More recently, Banerjee and Ghosal (2015) proposed a Bayesian approach to the glasso method. Finally, several authors studied non-likelihood precision estimation methods (see, for instance, Yuan 2010; Cai et al. 2011, among others).

As a consequence, the glasso method has become a state-of-the-art estimator for the precision matrix and one of the most applied approaches for covariance selection. It is worth noting that the loss function of the glasso is the negative log-likelihood function of the Gaussian model. Although the Gaussian assumption of data is quite restrictive, the glasso framework still provides a consistent estimator for non-Gaussian data (Ravikumar et al. 2011). However, the log-likelihood function may not be a comprehensible loss function because of its complex nature. Recently, Zhang and Zou (2014) introduced a so-called D-trace loss which has a much simpler structure. Through numerical simulations, they show that the ℓ_1 norm penalized D-trace loss minimization approach outperforms the glasso estimator in terms of different performance measures.

In this paper, we focus on the ℓ_1 norm penalized D-trace loss minimization method. It is well known that ℓ_1 penalty produces significant biases because of its overestimation feature (see, for instance, Zou 2006; Fan et al. 2009). The contribution of this paper aimed to mitigate those biases. Based on the adaptive

framework, we propose two re-weighted versions of the ℓ_1 norm penalized D-trace loss minimization approach. We employ adaptive thresholding operators in our proposed extensions. Previously, the adaptive framework has been applied in other context, such as variable selection (see Zou 2006), precision matrix estimation (see Fan et al. 2009) and covariance matrix estimation (see Rothman et al. 2009). The advantage of the adaptive Lasso framework in high-dimensional settings is that it provides a stable and sparse estimator, simultaneously corrects the bias and, moreover, it does not augment the computational time.

Through extensive numerical simulations we show that the methods based on the proposed extensions outperform the original D-trace method. In particular, for the simulation study we consider different models, including those used in the simulation experiments by Zhang and Zou (2014). To measure the statistical performance of the methods, we use the entropy loss, the Frobenius norm loss, the operator norm loss and the matrix ℓ_1 norm loss. Furthermore, we use the percentages of correctly estimated zeros and non-zeros, accuracy and Matthews Correlation Coefficient (MCC) to measure the GGM prediction performance. Finally, we investigate the performance of the estimators in discriminant analysis using real datasets.

The rest of the article is organized as follows. In Section 2, after introducing some notations, we describe two extensions of the D-trace precision matrix estimation based on the adaptive Lasso framework. We consider the statistical loss and GGM prediction performance of the proposed estimators in Section 3 through exhaustive numerical simulations. We compare our proposed estimators with the D-trace and glasso estimators. In Section 4, we apply the proposed methodologies to two real-

world applications: the prediction of breast cancer state and the prediction of the colon cancer state. We provide the conclusions in Section 5. Finally, we provide the simulation results in Appendix A.

2 Proposed Methodologies

Before proposing the adaptive extensions of the D-trace method, we introduce the following notations. For any vector $\mathbf{a} = (a_1, ..., a_p)^T \in \mathbb{R}^p$, we define the vector ℓ_2 or Euclidean norm $||\mathbf{a}||_2 = \sqrt{\sum_{j=1}^{p} a_j^2}$. For any symmetric $p \times p$ matrix $\mathbf{A} = [a_{ij}]_{1 \leq i,j \leq p} \in R^{p \times p}$, we denote the Frobenius norm by $||\mathbf{A}||_2 = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}^2}$, the matrix ℓ_∞ norm by $||\mathbf{A}||_\infty = \max_{1 \leq i,j \leq p} |a_{ij}|$, the matrix ℓ_1 norm by $||\mathbf{A}||_{\ell_1} = \max_{1 \leq j \leq p} \sum_{i=1}^{p} |a_{ij}|$, and the spectral or operator norm by $||\mathbf{A}||_{\text{spec}} = \sup_{||x||_2 \leq 1} ||Ax||_2$. We also denote the componentwise ℓ_1 norm by $||\mathbf{A}||_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |a_{ij}|$ and the off-diagonal componentwise ℓ_1 norm by $||\mathbf{A}||_{1,\text{off}} = \sum_{i=1}^{p} \sum_{j=1, j \neq i}^{p} |a_{ij}|$. For any two symmetric $p \times p$ matrices A and B, we write $A \succeq B$ or $A \succ B$ if the matrix $A - B$ is positive semidefinite or positive definite, respectively. We denote the smallest eigenvalue of the matrix A by $\lambda_{min}(A)$. We set diag(A) a diagonal matrix, which has the diagonal entries of A. Finally, we assume that $\mathbf{X}$ is a centered sample data matrix with dimension $n \times p$, where each row $X_i = (X_{i1}, ..., X_{ip})$ is a realization of a p-variate normal random vector that is independent and identically distributed for $i = 1, ..., n$, with covariance matrix Σ and precision matrix $\Omega = \Sigma^{-1}$.

The glasso estimator is the solution of the ℓ_1 penalized log-likelihood problem,

defined as follows:

$$\widehat{\Omega}_{\text{glasso}} = \arg\max_{\Omega} \ \log\det\Omega - \text{trace}(S\Omega) - \nu||\Omega||_{1,\text{off}}, \tag{1}$$

where $S = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ is the sample covariance matrix and $\nu > 0$ is the penalty parameter. Note that in the original definition of the glasso estimator, the norm $||\Omega||_1$ is used in (1) rather than $||\Omega||_{1,\text{off}}$. Although this term varies among different studies, in this particular paper we choose the off-diagonal penalization. This enables us to achieve fair comparison with the proposed method and with the results obtained by the previous scholars.

Zhang and Zou (2014) have proposed the D-trace loss function, which has the following definition:

$$f_{DT}(\Omega, \Sigma) = \frac{1}{2}\text{trace}(\Omega^2\Sigma) - \text{trace}(\Omega). \tag{2}$$

The function $f_{DT}(\Omega, \Sigma)$ is convex in Ω, has a positive-definite Hessian matrix, and a unique minimizer at Σ^{-1} (see Zhang and Zou 2014).

Zhang and Zou (2014) regularize the D-trace loss function through a ℓ_1 norm, thus, proposing the penalized D-trace loss minimization estimator (hereafter, D-trace or DT estimator) as the solution of the following optimization problem:

$$\widehat{\Omega}_{\text{DT}} = \arg\min_{\Omega \succeq \epsilon I} \ \frac{1}{2}\text{trace}(\Omega^2 S) - \text{trace}(\Omega) + \tau||\Omega||_{1,\text{off}}, \tag{3}$$

where $\tau > 0$ is the associated penalty parameter and ϵ is a small positive value. In problem (3), we have selected the off-diagonal $||\Omega||_{1,\text{off}}$ penalty term to be consistent with the original article. Similar to the glasso estimator, the $||\Omega||_1$ penalty can also be used for the D-trace estimator.

To solve the problem (3), Zhang and Zou (2014) developed an algorithm based on the alternating direction method. Previously, other authors have applied the alternating direction method approach for solving optimization problems (see, for instance, Scheinberg et al. 2010; Xue et al. 2012; Cui et al. 2014).

One of the important steps in the algorithm, where the Lasso penalty appears, is the following optimization problem:

$$\min_{\Omega=\Omega^T} \frac{1}{2}\text{trace}(\Omega^2) - \text{trace}(\Omega A) + \tau||\Omega||_{1,\text{off}}, \tag{4}$$

where the matrix A is defined in the algorithm process. One can show that the optimization problem (4) is strongly related to the soft thresholding operator. The solution $\hat{\Omega} = [\hat{\omega}_{ij}]_{1\leq i,j\leq p}$ of problem (4) can be written as:

$$\hat{\Omega} = T(A, \tau), \tag{5}$$

where T is the soft thresholding operator defined by:

$$[T(A,\tau)]_{ij} = \text{sign}(A_{ij})\max(|A_{ij}|-\tau, 0)I_{i\neq j} + A_{ij}I_{i=j} = \begin{cases} A_{ij}, & \text{if } i = j, \\ A_{ij} - \tau, & \text{if } i \neq j, A_{ij} > \tau, \\ A_{ij} + \tau, & \text{if } i \neq j, A_{ij} < -\tau, \\ 0, & \text{if } i \neq j, -\tau \leq A_{ij} \leq \tau \end{cases} \tag{6}$$

for $1 \leq i, j \leq p$.

As discussed in the Section 1, this paper addresses the bias problem of the Lasso. From the regularization point of view, the ℓ_1 penalty may not be the best choice because of this issue. In order to reduce the bias of the D-trace estimator, produced through the Lasso regularization in (4) (or through the soft thresholding operator (6)), we propose two adaptive extensions of the D-trace estimator.

We propose our first adaptive approach, motivated by the idea of the adaptive glasso method provided by Fan et al. (2009). First, for a specific weight matrix $W = [w_{ij}]_{1\leq i,j\leq p}$, we define the Weighted Adaptive Thresholding operator as:

$$[WAT(A,\tau)]_{ij} = \text{sign}(A_{ij})\max(|A_{ij}| - \frac{\tau}{|w_{ij}|}, 0)I_{i\neq j} + A_{ij}I_{i=j}, \tag{7}$$

for $1 \leq i, j \leq p$. One can straightforwardly verify the following property of the weighted adaptive thresholding operator (7):

$$w_{ij} = 0 \implies [WAT(A,\tau)]_{ij} = 0, \tag{8}$$

for $1 \leq i, j \leq p$. The small w_{ij} weights imply large penalties for the (i, j) entries, whereas the large w_{ij} weights imply small penalties for the (i, j) entries.

Next, we can write the weighted adaptive thresholding operator (7) as the solution of the following convex optimization problem:

$$\min_{\Omega=\Omega^T} \frac{1}{2}\text{trace}(\Omega^2 I) - \text{trace}(\Omega A) + \tau \sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} \frac{|\omega_{ij}|}{|w_{ij}|}. \tag{9}$$

Finally, by replacing the problem in (4) with the problem in (9), we derive our proposed *Weighted Adaptive D-trace estimator*, defined as:

$$\widehat{\Omega}_{\text{WADT}} = \arg\min_{\Omega\succeq\epsilon I} \frac{1}{2}\text{trace}(\Omega^2 S) - \text{trace}(\Omega) + \tau \sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} \frac{|\omega_{ij}|}{|w_{ij}|}. \tag{10}$$

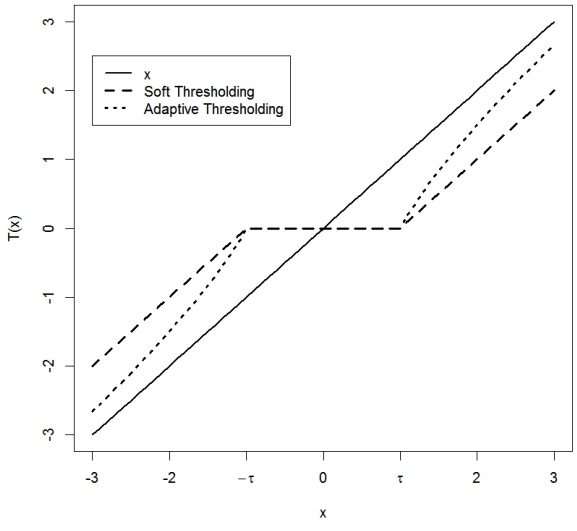
Essentially, the matrix W is a prior information about the precision matrix or any consistent, computationally cheap estimator (e.g., a well-defined two-step estimator) and, therefore, should be chosen properly.

Our second adaptive approach is motivated by Rothman et al. (2009), where we use the Adaptive Thresholding operator, defined as:

$$[AT(A,\tau)]_{ij} = \text{sign}(A_{ij})\max(|A_{ij}| - \frac{\tau}{|A_{ij}|}, 0)I_{i\neq j} + A_{ij}I_{i=j}, \tag{11}$$

for $1 \leq i, j \leq p$. The operator (11) can be considered as a special case of the operator (7), when $w_{ij} = A_{ij},\ 1 \leq i, j \leq p$. To illustrate the idea, Figure 1 depicts the soft and the adaptive thresholding operators for $\tau = 1$.

Figure 1. Soft and Adaptive thresholding functions, for $\tau = 1$.



The main advantage of the operator (11) is the absence of a weight matrix. Through the Adaptive Thresholding operator (11), the large entries A_{ij} are penalized less and the small entries are penalized more. In other words, the operator (11) overestimates less than the soft threshodling operator (6) since many smaller values will be discarded. Hence, the operator (11) provides smaller bias than the operator (6) (i.e., the Lasso penalization). As with the Weighted Adaptive D-trace estimator, one can derive formulations similar to (9) and (10) for the Adaptive Thresholding

operator. However, in these formulations the weight matrix W can not be defined directly, since the matrix A appears in the solver and is not fixed. We can obtain the D-trace estimator through the Adaptive Thresholding operator (11) by simply replacing the soft thresholding operator (6) with the operator (11) in the algorithm (see the Algorithm 1 bellow for more details). We call the estimator obtained through the operator (11) the *Adaptive D-trace estimator* $\hat{\Omega}_{ADT}$.

For completeness, we present the algorithm for solving the DT method and the necessary modifications for solving WADT and ADT methods.

We first provide the definitions of some functions employed in the algorithm. Assume that $A = UVU^T$ is the eigen-decomposition of any $p \times p$ symmetric matrix $A \succ 0$ and $v_1 \geq ... \geq v_p$ are its eigenvalues. For any $p \times p$ matrix B, define

$$G(A, B) = U\{(U^T BU) \circ C\}U^T, \tag{12}$$

where $C_{i,j} = \dfrac{2}{v_i + v_j}$ for $1 \leq i, j \leq p$ and $\circ$ denotes the Hadamard product of matrices.

Consider any symmetric matrix A and let $A = UVU^T = U\text{diag}(v_1, ..., v_p)U^T$ is its eigen-decomposition. For any $\epsilon > 0$, define

$$[A]_+ = U\text{diag}\{\max(v_1, \epsilon), ..., \max(v_p, \epsilon)\}U^T. \tag{13}$$

Algorithm 1 provides the necessary steps for solving our proposed estimation methods:

It is important to note that we can significantly reduce the computational time of the Algorithm 1 by discarding the constraint $\Omega \succeq \epsilon I$ in the initial optimization problem (DT, WADT or ADT). This enables us to omit the function (13) from the

Algorithm 1 Alternating direction method

Step 1. Initialization: $k = 0$, $\Lambda_0^0 = \Lambda_1^0$, $\Theta_0^0 = \Theta_1^0$.

Step 2. Repeat the following sub-steps until convergence:

(a) Set k=k+1.

(b) Compute the matrix $\Theta^{k+1} = G(S + 2\rho I, I + \rho\Theta_0^k + \rho\Theta_1^k - \Lambda_0^k - \Lambda_1^k)$, where function G is defined in (12).

(c) Set $\Theta_1^{k+1} = [\Theta^{k+1} + \Lambda_1^k/\rho]_+$. Compute $\Theta_0^{k+1} = T(\Theta^{k+1} + \Lambda_0^k/\rho, \tau/\rho)$ in case of DT estimator, $\Theta_0^{k+1} = WADT(\Theta^{k+1} + \Lambda_0^k/\rho, \tau/\rho)$ in case of WADT estimator and $\Theta_0^{k+1} = ADT(\Theta^{k+1} + \Lambda_0^k/\rho, \tau/\rho)$ in case of ADT estimator. The thresholding functions T, WADT and ADT are defined in (6), (7) and (11), respectively.

(d) Set $\Lambda_0^{k+1} = \Lambda_0^k + \rho(\Theta^{k+1} - \Theta_0^{k+1})$ and $\Lambda_1^{k+1} = \Lambda_1^k + \rho(\Theta^{k+1} - \Theta_1^{k+1})$.

step 2c, which is the most computationally expensive part of the algorithm. We call the optimization problem without the constraint $\Omega \succeq \epsilon I$ the secondary problem, defined as:

$$\tilde{\Omega} = \arg\min_{\Omega^T=\Omega} \frac{1}{2}\text{trace}(\Omega^2 S) - \text{trace}(\Omega) + \tau\text{PEN}(\Omega), \tag{14}$$

where PEN(Ω) term is defined according to the estimation method (DT, ADT or WADT). Following Zhang and Zou (2014), we also present the simplified version of the Algorithm 1.

The algorithm stops if the following two conditions are satisfied:

$$\frac{||\Theta^{k+1} - \Theta^k||_2}{\max(1, ||\Theta^k||_2, ||\Theta^{k+1}||_2)} < 10^{-7}, \qquad \frac{||\Theta_0^{k+1} - \Theta_0^k||_2}{\max(1, ||\Theta_0^k||_2, ||\Theta_0^{k+1}||_2)} < 10^{-7},$$

Algorithm 2 Alternating direction method (simplified)

Step 1. Initialization: $k = 0$, Λ^0, $\Theta_0^0 = \text{diag}(S)^{-1}$.

Step 2. Repeat the following sub-steps until convergence:

(a) Set k=k+1.

(b) Compute the matrix $\Theta^{k+1} = G(S + 2\rho I, I + \rho\Theta_0^k - \Lambda^k)$.

Compute $\Theta_0^{k+1} = T(\Theta^{k+1} + \Lambda^k/\rho, \tau/\rho)$ in case of DT estimator, $\Theta_0^{k+1} = WADT(\Theta^{k+1} + \Lambda^k/\rho, \tau/\rho)$ in case of WADT estimator and $\Theta_0^{k+1} = ADT(\Theta^{k+1} + \Lambda^k/\rho, \tau/\rho)$ in case of ADT estimator.

(d) Set $\Lambda^{k+1} = \Lambda_0^k + \rho(\Theta^{k+1} - \Theta_0^{k+1})$.

Step 3. Consider the converged Θ^k as the solution of the secondary problem (14).

Step 4. If $\lambda_{\min}(\tilde{\Theta}) > \epsilon$, report $\tilde{\Theta}$ as the solution of the initial problem. Otherwise, use Algorithm 1 with $\tilde{\Theta}$ as the starting value for Θ_0^0 and Θ_1^0.

Finally, in the algorithm we use $\rho = 1$ and $\epsilon = 10^{-8}$. For more details we refer to Zhang and Zou (2014).

3 Simulation Study

In this section, we implement a simulation study to show the goodness of the proposed WADT and ADT estimators and to compare their associated performance with those of the DT estimator and the state-of-the-art estimator glasso. Particularly,

in subsection 3.1, we introduce the models considered for the true precision matrix Ω. In subsection 3.2, we describe the performance evaluation. In subsection 3.3, we provide the discussion of the obtained results.

3.1 Considered Models

We perform an exhaustive numerical simulation study through eight different sparsity configurations for the precision matrix, including random and fixed patterns. The considered models for the true precision matrix Ω are the following:

- *Model* 1. AR(2) structure: $\omega_{i,i} = 1$, $\omega_{i,j} = 0.2$ for $1 \leq |i - j| \leq 2$, and zero otherwise.
- *Model* 2. AR(4) structure: $\omega_{i,i} = 1$, $\omega_{i,j} = 0.2$ for $1 \leq |i - j| \leq 4$, and zero otherwise.
- *Model* 3. A matrix with $\omega_{i,i} = 1$, $\omega_{i,i+1} = 0.2$ for $\text{mod}(i, p^{1/2}) \neq 0$, $\omega_{i,i+p^{1/2}} = 0.2$, and zero otherwise.
- *Model* 4. AR(1) structure: $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.45$, and zero otherwise.
- *Model* 5. (Modified) AR(1) structure with different entries: $\Omega = G^{1/2}\Omega_{AR(1)}G^{1/2}$, where G is a diagonal matrix with entries $G_{ii} = \dfrac{4i + p - 5}{5(p - 1)}$ and $\Omega_{AR(1)}$ is a matrix with a structure defined in the model 4.
- *Model* 6. Decay structure: $\omega_{ij} = 0.6^{|i-j|}$.
- *Model* 7. A random positive-definite matrix, containing 5% of non-zero entries.

- *Model* 8. A random positive-definite matrix, containing 10% of non-zero. entries[2]

Our choice of these models is motivated as follows. To compare our proposed methods with Zhang and Zou (2014), we consider the models employed in their study (models 1, 2 and 3). In addition, we consider other models commonly used in the prior literature, such as AR(1) structure model (model 4 - Yuan and Lin (2007), Friedman et al. (2008)), its modified version (model 5) and decay structure model (model 5 - Cai et al. (2011), Fan et al. (2009)). Note that models 1-6 have deterministic patterns. We study the performance of the considered methods also using models with random patterns (models 7 and 8). This allows us to obtain more robust evaluation and to have better insight about the performance of the estimation methods.

Consistent with Zhang and Zou (2014), we simulate multivariate normal random samples with zero mean and sample size $n = 400$, for each of the models. For the number of variables, we choose $p = 484$ for model 3 and $p = 500$ for the other models.[3] These values allow us to examine the performance of the proposed estimators in high-dimensional settings and, especially, when $p > n$. Finally, we repeat this procedure 100 times.

3.2 Performance Evaluation

Similar to Zhang and Zou (2014), to evaluate the statistical performance of a

[2]Models 7 and 8 are generated using the Matlab command *sprandsym*.

[3]For model 3 the value of $p^{1/2}$ is required to be an integer

given estimator $\widehat{\Omega}$, we consider the Frobenius norm ℓ_2, the spectral norm ℓ_{spec} and the matrix ℓ_1 norm, defined respectively as:

$$\ell_2(\widehat{\Omega}, \Omega) = ||\widehat{\Omega} - \Omega||_2, \tag{15}$$

$$\ell_{\text{spec}}(\widehat{\Omega}, \Omega) = ||\widehat{\Omega} - \Omega||_{\text{spec}}, \tag{16}$$

and

$$\ell_1(\widehat{\Omega}, \Omega) = ||\widehat{\Omega} - \Omega||_{\ell_1}. \tag{17}$$

Next, we consider the entropy loss function, also known as Kullback-Leibler (KL) loss function, consistent with its widespread application in the prior literature (see, for instance, Yuan and Lin 2007; Rothman et al. 2008; Fan et al. 2009; Yin and Li 2013). This function is defined as:

$$\text{KLL}(\widehat{\Omega}, \Omega) = \text{trace}(\Omega^{-1}\widehat{\Omega}) - \log\det(\Omega^{-1}\widehat{\Omega}) - p. \tag{18}$$

In order to evaluate the sparsity pattern or GGM estimation performance, we compute the percentages of correctly estimated non-zeros and zeros (also known as sensitivity and specificity, respectively) and the accuracy of classification, defined respectively as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100, \tag{19}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100, \tag{20}$$

and

$$\text{Accuracy} = \frac{\text{TN+TP}}{p^2} \times 100. \tag{21}$$

Here TP, TN, FP and FN are the numbers of true positives (i.e., the number of correctly estimated non-zero entries), true negatives (i.e., the number of correctly estimated zero entries), false positives (i.e., the number of erroneously estimated non-zero entries) and false negatives (i.e., the number of erroneously estimated zero entries), respectively. It is worth noting that FP and FN refer to Type I and Type II errors, respectively. We also compute the Matthews Correlation Coefficient (MCC), which is commonly used to measure the performance of binary classifiers. The MCC was introduced by Matthews (1975) and is defined as:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{22}$$

In order to select the penalty parameters ν and τ, in line with Zhang and Zou (2014), we use five-fold cross-validation technique. For the WADT estimator, as a weight matrix we choose the inverse of the Ledoit-Wolf shrinkage covariance estimator $W = \hat{\Sigma}_{LW}^{-1}$. We use the Matlab code of Zhang and Zou (2014) to implement the algorithm for the DT method and the modification of their code for the WADT and ADT estimators.

3.3 Discussion of Results

We provide the simulation results in Appendix A to conserve space. Tables 5-12 report the averages of the corresponding losses and measurements over 100 replications. The standard deviations (SD) are given in parentheses. Tables 9, 10

and 12 provide the measurements in percentages. We organize the discussion of our results as follows. We first compare our proposed estimators ADT and WADT with the DT estimator. We then compare our proposed estimators ADT and WADT with the glasso estimator. We finally compare the DT estimator with the glasso estimator.

We report the statistical losses in Tables 5-8. We observe that for most of the models either the ADT or the WADT estimator provides the lowest losses versus the other methods (DT and glasso). More specifically, the ADT estimator provides the lowest KLL for models 1, 2, 6, the lowest Frobenius norm and spectral norm for models 2, 6 and the lowest matrix ℓ_1 norm for models 1, 3, 6. On the other hand, the WADT estimator provides the lowest KLL for models 3, 4, 5, 7, 8, the lowest Frobenius norm and spectral norm for models 1, 3, 4, 5, 7, 8 and the lowest matrix ℓ_1 norm for models 4, 5, 7, 8. The only exception when the ADT estimator fails to outperform the DT estimator is for models 2, 8 in terms of matrix ℓ_1 norm and for models 1, 3, 7 in terms of spectral norm. The only exception when the ADT estimator fails to outperform the glasso method is for model 3 in terms of KLL. The only exception when the WADT estimator fails to outperform the DT estimator is for models 1, 2 (only in terms of matrix ℓ_1 norm) and for model 6, which is precisely a dense model. The WADT method outperforms glasso method in all the models.

The comparison of the performances of DT versus glasso yields the following insights. In line with Zhang and Zou (2014), we find that DT outperforms glasso for all the models in terms of Frobenius norm, spectral norm, and ℓ_1 norm. However, in their work, Zhang and Zou (2014) did not compare DT and glasso in terms of KLL. We find mixed results in comparative performance of DT versus glasso. We

observe that DT outperforms glasso for models 1, 2, 5, 6, 7, 8 in terms of the KLL. In contrast to Zhang and Zou (2014), we find that DT fails to outperform glasso for models 3 and 4 in terms of KLL.

We report the GGM prediction performance in Tables 9-12.[4] We observe that for most of the models either the ADT or the WADT estimator provides better GGM prediction performance than the other methods (DT and glasso). More specifically, the ADT estimator provides the highest specificity for models 1, 3, the highest sensitivity for model 2 and the highest MCC and accuracy for models 1, 2, 3. On the other hand, the WADT estimator providesc the highest specificity for models 2, 4, 5, 7, 8 and the highest MCC and accuracy for models 4, 5, 7, 8. All the estimators provide the same sensitivity for models 4 and 5. The only exception when our proposed estimators (ADT and WADT) fail to outperform the DT estimator is for models 1, 3, 6, 7, 8 only in terms of sensitivity. However, for these models the DT estimator fails to outperform the estimators ADT and WADT in terms of the overall GGM prediction measures MCC and accuracy. In addition, the only exception when our proposed estimators fail to outperform the glasso estimator is for models 3, 6, 7, 8 in terms of sensitivity. However, the glasso estimator fails to outperform the proposed estimators in terms of the overall GGM prediction measures MCC and accuracy for those models. Comparing the DT estimator with the glasso estimator our findings show that the later outperforms the DT estimator for models 3, 6, 7, 8 in terms of sensitivity and for model 3 in terms of specificity. In terms of the overall GGM

[4]The specificity, MCC and accuracy are excluded for model 6 because these measures are defined only for sparse models.

prediction measures the DT estimator outperforms the glasso estimator for all the models except for model 3, where the glasso provides slightly higher accuracy and MCC than DT.

As a summary, our proposed adaptive approaches ADT and WADT outperform the estimators DT and glasso for overwhelming majority of the considered models. In spite of few exceptions, the proposed methods provide better performance in terms of the statistical losses and GGM prediction measures, than the competitive methods. In addition, our findings show that the WADT method provides relatively better results than the ADT method when the required weight matrix is the inverse of an estimated covariance matrix.

4 Real Data Applications

In this section, we perform an empirical analysis of the proposed adaptive approaches through real-data examples. In particular, we use breast cancer and colon cancer datasets to predict the tumour behaviour using Linear Discriminant Analysis (LDA). All applied datasets are available in the web site of the National Center for Biotechnology Information.[5]

4.1 Breast Cancer Data

In the first application, we focus on the problem of predicting breast cancer patients (subjects) with pathological complete response (pCR). This is an important issue because after the neoadjuvant chemotherapy, according to Kuerer et al. (1999),

[5]Available at `http://www.ncbi.nlm.nih.gov/`.

the pCR indicates a cancer-free life with high probability. For this application we use a dataset (for the description of the dataset we refer to Shi et al. 2010) containing gene expression levels of subjects with different stages of breast cancer. The dataset consists of 22,283 gene expression levels of 271 subjects. There are 58 subjects with pCR and 213 subjects with residual disease (RD).

First, we divide the data into a training set and a testing set with sizes 227 (almost 5/6 of the observations) and 44 (almost 1/6 of the observations), respectively, and repeat this process 100 times. For the testing set, we randomly select 9 subjects with pCR and 35 subjects with RD (roughly proportional to the number of the subjects in each group). The training set contains the remaining subjects. Second, based on the training set we perform two sample t-tests between the two groups in order to select the most significant 100 genes with the smallest p-values. Third, using the training set, we estimate the precision matrix Ω with the DT, ADT, WADT and glasso methods. We obtain the penalty parameters for these methods using five-fold cross-validation technique. Finally, we use the estimated precision matrix in the LDA score, defined as follows:

$$\delta_t(Y) = Y^T \widehat{\Omega} \widehat{\mu}_t - \frac{1}{2} \widehat{\mu}_t^T \widehat{\Omega} \widehat{\mu}_t, \tag{23}$$

where $t = 1, 2$ ($t = 1$ for pCR and $t = 2$ for RD) and $\widehat{\mu}_t = \dfrac{1}{n_t} \sum_{i \in class_t} x_i$ is the within group average, calculated using the training data. We use the LDA score $\delta_t(Y)$ to classify the subject Y from the testing set. The rule for the classification is $\widehat{t} = \arg\max \delta_t(Y)$ $(t = 1, 2)$. To measure the prediction accuracy for all the methods, we use the specificity, sensitivity and Matthews Correlation Coefficient (MCC), as defined in Section 3.2. We consider TP and TN as the number of correctly predicted

RD and pCR, respectively, and FP and FN as the number of erroneously predicted RD and pCR, respectively. We report the average measurements over 100 replications in Table 1.

Table 1. Average pCR/RD classification measurements over 100 replications for $p = 100$ genes.

Method	Specificity	Sensitivity	MCC
GLASSO	0.4800	0.7751	0.2333
DT	0.6556	0.7537	0.3572
ADT	0.6989	0.7409	0.3782
WADT	0.7211	0.7334	0.3889

Our findings show that the glasso provides the highest sensitivity, but it attains the lowest specificity and MCC. On the other hand, the adaptive approach WADT provides the highest specificity and dominates all the other estimators in terms of MCC. Furthermore, the ADT and WADT estimators show similar results, the latter being slightly better.

To check the robustness of the obtained results, we repeat the same application by considering the most significant 200 genes instead of 100. Table 2 reports the results. Our findings show that the results are roughly similar to those obtained with 100 genes. The adaptive methods ADT and WADT outperform DT and glasso methods in terms of the overall measurement MCC.

Table 2. Average pCR/RD classification measurements over 100 replications for $p = 200$ genes.

Method	Specificity	Sensitivity	MCC
GLASSO	0.4600	0.7891	0.2310
DT	0.6333	0.7620	0.3459
ADT	0.7033	0.7394	0.3793
WADT	0.7089	0.7414	0.3860

4.2 Colon Cancer Data

In the second application, we consider the problem of classifying the colorectal cancer patients with Microsatellite Stability (MSS) state and Microsatellite Instability (MSI) state. The dataset (for the description of the dataset we refer to Jorissen et al. 2008) contains the expression levels of 54,675 genes for 155 colorectal cancer samples. There are 77 MSS and 78 MSI specimens in the dataset.

As with the first application, we divide the data into a training set and a testing set with sizes 130 (almost 5/6 of the observations) and 25 (almost 1/6 of the observations), respectively, and repeat this process 100 times. We randomly select 12 MSS and 13 MSI specimens (roughly proportional to the number of the subjects in each group), respectively, for the testing set and the training set contains the remaining subjects. Again, we select the 100 most-significant genes and estimate the precision matrix Ω with the DT, ADT, WADT and glasso methods. We obtain the penalty parameters for these methods using five-fold cross-validation technique. Finally, we use the estimated precision matrix in the LDA score (23), where $t = 1$ is for MSS

specimens and $t = 2$ is for MSI specimens.

Table 3 shows the average performance measures over the 100 replicates. We observe that glasso provides the lowest performance measures while the WADT estimator provides the highest ones. The DT and ADT estimators provide relatively similar results.

Table 3. Average MSI/MSS classification measurements over 100 replications for $p = 100$ genes.

Method	Specificity	Sensitivity	MCC
GLASSO	0.9258	0.8961	0.8262
DT	1	0.8977	0.9020
ADT	1	0.8915	0.8966
WADT	1	0.9208	0.9235

We repeat the same application by considering the most significant 200 genes instead of 100. Table 4 provides the results. We observe that the results are similar to those obtained using 100 genes.

In sum, our findings show that in the considered applications the proposed WADT and ADT methods are able to provide better classification performance than DT and glasso estimators.

5 Conclusions

In this article, we develop two novel approaches for estimating the precision matrix, based on the adaptive ℓ_1 regularization framework. We extend the recently

Table 4. Average MSI/MSS classification measurements over 100 replications for $p = 200$ genes.

Method	Specificity	Sensitivity	MCC
GLASSO	0.8558	0.8330	0.6956
DT	1	0.9015	0.9050
ADT	1	0.9054	0.9086
WADT	1	0.9238	0.9258

introduced D-trace estimator to Weighted Adaptive D-trace (WADT) and Adaptive D-trace (ADT) estimators to correct the bias of the estimated precision matrix produced by the ℓ_1 penalty. In our proposed methodologies we use the adaptive thresholding operators. We conduct an extensive numerical analysis, applying both simulated and real data sets. For the WADT estimator we use the two-step precision matrix estimator as a weight matrix. Our findings show that it is a practical choice. We use different loss functions and prediction performance measures for the evaluation. The results show that the proposed estimators outperform the DT and glasso estimators. In particular, the WADT and ADT estimators provide lower statistical losses and higher GGM prediction measures than those for the DT and glasso methods.

A Appendix: Numerical Results

Table 5. Average KL losses (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	21.680 (1.388)	38.497 (0.342)	16.792 (0.436)	18.201 (0.685)
DT	20.598 (0.384)	37.335 (0.331)	19.317 (0.494)	19.381 (0.454)
ADT	19.171 (0.483)	34.318 (0.434)	17.511 (0.595)	6.727 (0.285)
WADT	19.536 (0.970)	35.512 (0.496)	12.860 (0.513)	4.652 (0.214)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	23.336 (0.389)	53.028 (0.270)	41.149 (0.333)	47.480 (0.333)
DT	18.518 (0.517)	30.733 (0.433)	29.248 (0.423)	33.168 (0.407)
ADT	10.642 (0.416)	21.219 (0.344)	28.342 (0.390)	31.611 (0.916)
WADT	4.9425 (0.247)	48.439 (0.337)	21.880 (0.475)	26.031 (0.483)

Table 6. Average Frobenius norm losses (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	7.402 (0.314)	12.042 (0.030)	5.398 (0.060)	6.931 (0.314)
DT	6.953 (0.066)	11.467 (0.041)	4.898 (0.063)	4.890 (0.082)
ADT	6.685 (0.094)	10.681 (0.068)	4.803 (0.081)	2.741 (0.076)
WADT	6.563 (0.396)	10.949 (0.123)	4.068 (0.080)	2.366 (0.066)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	5.512 (0.042)	20.782 (0.032)	3.307 (0.013)	3.774 (0.011)
DT	2.668 (0.070)	16.057 (0.080)	2.296 (0.027)	2.765 (0.025)
ADT	1.938 (0.063)	13.478 (0.119)	2.205 (0.030)	2.627 (0.068)
WADT	1.562 (0.058)	18.106 (0.066)	1.960 (0.024)	2.307 (0.026)

Table 7. Average operator norm losses (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	0.774 (0.028)	1.630 (0.007)	0.589 (0.013)	0.663 (0.026)
DT	0.741 (0.018)	1.556 (0.012)	0.535 (0.016)	0.539 (0.024)
ADT	0.750 (0.023)	1.454 (0.020)	0.544 (0.022)	0.388 (0.029)
WADT	0.704 (0.054)	1.474 (0.024)	0.451 (0.021)	0.348 (0.038)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	0.797 (0.022)	2.980 (0.005)	0.613 (0.009)	0.736 (0.007)
DT	0.412 (0.030)	2.474 (0.017)	0.544 (0.009)	0.644 (0.009)
ADT	0.317 (0.031)	2.198 (0.032)	0.551 (0.009)	0.644 (0.012)
WADT	0.292 (0.032)	2.691 (0.012)	0.509 (0.010)	0.593 (0.012)

Table 8. Average matrix ℓ_1 norm losses (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	1.329 (0.138)	2.032 (0.042)	0.992 (0.050)	0.970 (0.038)
DT	1.109 (0.047)	1.939 (0.034)	0.924 (0.043)	0.680 (0.034)
ADT	1.051 (0.045)	1.953 (0.052)	0.840 (0.045)	0.505 (0.038)
WADT	1.138 (0.121)	1.955 (0.052)	0.846 (0.053)	0.477 (0.048)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	0.923 (0.030)	3.390 (0.039)	1.242 (0.014)	1.659 (0.015)
DT	0.590 (0.045)	2.900 (0.042)	1.077 (0.033)	1.571 (0.028)
ADT	0.426 (0.047)	2.612 (0.054)	1.077 (0.034)	1.575 (0.039)
WADT	0.385 (0.046)	2.916 (0.026)	0.997 (0.044)	1.522 (0.038)

Table 9. Average specificity (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	97.01 (1.21)	98.14 (0.05)	98.18 (0.05)	96.30 (0.72)
DT	98.26 (0.03)	98.18 (0.04)	98.03 (0.04)	99.33 (0.02)
ADT	99.49 (0.02)	98.68 (0.03)	99.73 (0.01)	99.63 (0.01)
WADT	98.96 (0.68)	98.87 (0.15)	99.05 (0.08)	99.66 (0.02)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	95.31 (0.07)	NA (NA)	94.78 (0.07)	94.73 (0.07)
DT	97.40 (0.05)	NA (NA)	97.86 (0.05)	98.02 (0.04)
ADT	99.17 (0.02)	NA (NA)	98.45 (0.03)	98.38 (0.38)
WADT	99.70 (0.02)	NA (NA)	99.63 (0.02)	99.46 (0.02)

Table 10. Average sensitivity (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	88.94 (5.04)	61.23 (0.77)	99.62 (0.15)	100 (0)
DT	91.17 (0.82)	67.53 (0.81)	99.54 (0.20)	100 (0)
ADT	84.13 (1.26)	68.18 (0.91)	97.20 (0.53)	100 (0)
WADT	84.55 (5.20)	62.01 (1.62)	98.97 (0.33)	100 (0)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	100 (0)	4.86 (0.08)	20.25 (0.30)	12.92 (0.19)
DT	100 (0)	3.88 (0.04)	19.37 (0.26)	11.36 (0.15)
ADT	100 (0)	1.77 (0.02)	17.87 (0.24)	10.72 (0.67)
WADT	100 (0)	0.68 (0.09)	16.65 (0.23)	9.80 (0.13)

Table 11. Average MCC (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	0.467 (0.082)	0.467 (0.005)	0.589 (0.006)	0.372 (0.036)
DT	0.555 (0.005)	0.511 (0.005)	0.573 (0.004)	0.685 (0.006)
ADT	0.720 (0.009)	0.565 (0.006)	0.872 (0.005)	0.785 (0.007)
WADT	0.639 (0.080)	0.549 (0.010)	0.708 (0.016)	0.800 (0.009)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	0.329 (0.002)	NA (NA)	0.136 (0.002)	0.094 (0.002)
DT	0.427 (0.003)	NA (NA)	0.216 (0.003)	0.164 (0.002)
ADT	0.646 (0.006)	NA (NA)	0.230 (0.003)	0.172 (0.007)
WADT	0.816 (0.010)	NA (NA)	0.325 (0.004)	0.229 (0.003)

Table 12. Average accuracy (with standard deviations) over 100 replications.

Methods	Model 1	Model 2	Model 3	Model 4
Glasso	96.93 (1.15)	97.48 (0.04)	98.19 (0.05)	96.33 (0.71)
DT	98.19 (0.03)	97.63 (0.04)	98.05 (0.03)	99.33 (0.02)
ADT	99.33 (0.02)	98.13 (0.03)	99.71 (0.01)	99.63 (0.01)
WADT	98.82 (0.62)	98.21 (0.12)	99.05 (0.08)	99.67 (0.02)
Methods	Model 5	Model 6	Model 7	Model 8
Glasso	95.34 (0.07)	NA (NA)	91.21 (0.06)	86.96 (0.06)
DT	97.41 (0.05)	NA (NA)	94.10 (0.04)	89.79 (0.04)
ADT	99.18 (0.02)	NA (NA)	94.59 (0.03)	90.05 (0.28)
WADT	99.70 (0.02)	NA (NA)	95.65 (0.02)	90.94 (0.02)

Acknowledgements

We express our gratitude to the authors Teng Zhang and Hui Zou for sharing their Matlab code that solves the ℓ_1 norm penalized D-trace loss minimization problem. Andrés M. Alonso gratefully acknowledges financial support from CICYT (Spain) Grants ECO2011-25706 and ECO2012-38442. Francisco J. Nogales and Vahe Avagyan are supported by the Spanish Government through project MTM2013-44902-P.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *The Journal of Multivariate Analysis*, 136:147–162.

Bickel, P., J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042.

Cui, Y., Leng, C., and Sun, D. (2014). Sparse estimation of high-dimensional correlation matrices. *Computational Statistics and Data Analysis*, Preprint avilable at http://dx.doi.org/10.1016/j.csda.2014.10.001.

d'Aspremont, A., Banerjee, O., and Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal Appl.*, 30:56–66.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Deng, X. and Tsui, K. (2013). Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics*, 22(2):494–512.

Duchi, J., Gould, S., and Koller, D. (2008). Projected subgradient methods for learning sparse gaussians. In *Proceeding of the 24th Conference on Uncertainty in Artificial Intelligence*.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Applied Statistics*, 36(6):2717–2756.

Fan, J., Feng, J., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Frahm, G. and Memmel, C. (2010). Dominating estimator for minimum-variance portfolios. *Journal of Econometrics*, 159:289–302.

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

Goto, S. and Xu, Y. (2013). Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantitative Analysis, (Forthcoming)*.

Haff, L. R. (1980). Estimation of the inverse covariance matrix: Random mixtures of the inverse wishart matrix and the identity. *The Annals of Statistics*, 8(3):586–597.

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., and Reiman, E. (2010). Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50:935–949.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics*, 29(3):295–327.

Jorissen, R. N., Lipton, L., Gibbs, P., Chapman, M., Desai, J., Jones, I. T., Yeatman, T. J., East, P., Tomlinson, I. P., Verspaget, H. W., Aaltonen, L. A., Kruhøffer, M., Orntoft, T. F., Andersen, C. L., and Sieber, O. M. (2008). DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clinical Cancer Research*, 14(24):8061–8069.

Kourtis, A., Dotsis, G., and Markellos, N. (2012). Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix. *Journal of Banking & Finance*, 36:2522–2531.

Kuerer, H. M., Newman, L. A., Smith, T. L., Ames, F. C., Hunt, K. K., Dhingra, K., Theriault, R. L., Singh, G., Binkley, S. M., Sneige, N., Buchholz, T. A., Ross, M. I., McNeese, M. D., Buzdar, A. U., Hortobagyi, G. N., and Singletary, S. E. (1999). Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. *Journal of Clinical Oncology*, 17(2):460–469.

Lauritzen, S. (1996). *Graphical Models.* Clarendon Press. Oxford.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, New York.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451.

Maurya, A. (2014). A joint convex penalty for inverse covariance matrix estimation. *Computational Statistics and Data Analysis*, 75:15–27.

McLachlan, S. (2004). *Discriminant Analysis and Statistical Pattern Recognition.* Willey Interscience.

Nguyen, T. D. and Welsch, R. E. (2010). Outlier detection and robust covariance estimation using mathematical programming. *Advances in Data Analysis and Classification*, 4(4):301–334.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Rothman, A., Bickel, P., and Levina, E. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

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

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99(2):733–740.

Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32.

Scheinberg, K., Ma, S., and Goldfarb, D. (2010). Sparse inverse covariance selec-

tion via alternating linearization methods. In *Advances in Neural Information Processing Systems*.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., deLongueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A. Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Scherf, U., Thierry-Mieg, J., Wang, C., Wilson, M., and Wolber, P. K. (2010). The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827–838.

Stifanelli, P. F., Creanza, T. M., Anglani, R., Liuzzi, V. C., Mukherjee, S., Schena, F. P., and Ancona, N. (2013). A comparative study of covariance selection models for the inference of gene regulatory networks. *Journal of Biomedical Informatics*, 46:894–904.

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

Touloumis, A. (2015). Nonparametric Stein-type shrnikage covariance matrix estimators in high-dimensional settings. *Computational Statistics and Data Analysis*, 83:251–261.

Wang, Y. and Daniels, M. J. (2014). Computationally efficient banding of large covariance matrices for ordered data and connections to banding the inverse Cholesky factor. *Journal of Multivariate Analysis*, 130:21–26.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103(481):340–349.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491.

Yin, J. and Li, J. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of Multivariate Analysis*, 116:365–381.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286.

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

Zerenner, T., Friederichs, P., Lehnertz, K., and Hense, A. (2014). A gaussian graphical model approach to climate networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(2):023103.

Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, 88:1–18.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.