Radboud Repository
sharing science

**An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics**

Engel, J.; Buydens, L.; Blanchet, L.

2017, Article / Letter to editor (Journal of Chemometrics, 31, 4, (2017), pp. 1-19, article e2880)

Doi link to publisher: https://doi.org/10.1002/cem.2880

Version of the following full text: Publisher's version

Published under the terms of article 25fa of the Dutch copyright act. Please follow this link for the

Terms of Use: https://repository.ubn.ru.nl/page/termsofuse

Downloaded from: https://hdl.handle.net/2066/174399

Download date: 2025-04-12

**Note:**

To cite this publication please use the final published version (if applicable).

Received: 31 August 2016 | Revised: 12 December 2016 | Accepted: 14 December 2016

DOI: 10.1002/cem.2880

WILEY Journal of CHEMOMETRICS

SPECIAL ISSUE ARTICLE

# An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics

**Jasper Engel**[1] | **Lutgarde Buydens**[2] | **Lionel Blanchet**[3]

[1] NERC Biomolecular Analysis Facility – Metabolomics Node (NBAF-B), School of Biosciences, University of Birmingham, Birmingham, West Midlands, UK

[2] Radboud University Nijmegen, Institute for Molecules and Materials, Nijmegen, Gelderland, the Netherlands

[3] Department of Pharmacology and Toxicology, School of Nutrition, Toxicology and Translational Research in Metabolism (NUTRIM), Maastricht University Medical Centre, Maastricht, Limburg, the Netherlands

**Correspondence**
Jasper Engel, NERC Biomolecular Analysis Facility – Metabolomics Node (NBAF-B), School of Biosciences, University of Birmingham, B15 2TT Birmingham, UK.
Email: j.engel@science.ru.nl

The covariance matrix (or its inverse, the precision matrix) is central to many chemometric techniques. Traditional sample estimators perform poorly for high-dimensional data such as metabolomics data. Because of this, many traditional inference techniques break down or produce unreliable results. In this paper, we selectively review several modern estimators of the covariance and precision matrix that improve upon the traditional sample estimator. We focus on 3 general techniques: eigenvalue-shrinkage estimation, ridge-type estimation, and structured estimation. These methods rely on different assumptions regarding the structure of the covariance or precision matrix. Various examples, in particular using metabolomics data, are used to compare these techniques and to demonstrate that in concert with, eg, principal component analysis, multivariate analysis of variance, and Gaussian graphical models, better results are obtained.

**KEYWORDS**

eigenvalue shrinkage, metabolomics, ridge-type estimation, sparse covariance matrix, sparse precision matrix

## 1 | INTRODUCTION

Many contemporary experiments in chemistry and related fields give rise to an ever-increasing amount of measurement data, originating from multiple advanced analytical technologies. Examples include infrared spectroscopic data for food authentication,[1] Raman data in industrial process monitoring,[2] nuclear magnetic resonance (NMR) or liquid chromatography–mass spectrometry (LC-MS)–based metabolomics data for biomarker discovery experiments, or a combination of data types in personalized health care.[3,4]

The availability of such high-dimensional data has reshaped statistical thinking and data analysis resulting in, for example, the emergence of data-driven research.[5–7] An absolute necessity in this respect is the use of powerful multivariate statistical techniques such as principal component analysis (PCA), multivariate analysis of variance (MANOVA), linear discriminant analysis (LDA), or Gaussian graphical Models (GGM) to make sense of the data.[8–10] These methods, however, were originally developed for traditional data sets where the number of samples ($n$) is (much) larger than the number of variables ($p$). Nowadays, the opposite is often encountered ($n << p$), and many multivariate techniques break down or their estimates are not reliable.

The poor performance of many traditional statistical methods for high-dimensional data can be readily understood when considering the sample covariance matrix. This matrix, or its inverse (the precision matrix), is an essential element of many multivariate techniques such as PCA, LDA, and MANOVA.[8,9] Typically, the sample estimator is used in these models. However, it is well known that as soon as $n < p$, the sample covariance matrix loses full rank as a growing number of eigenvalues become 0, amongst other issues.[11,12] Because of this, the poor performance of statistical techniques based on the covariance matrix is no surprise.

For high-dimensional data, the sample covariance matrix is not a good estimator of the population covariance matrix.[11,12] A (naïve) strategy to obtain a more efficient estimator can be to consider an estimator with a lot of structure imposed. For example, one could assume that all variables are uncorrelated, giving rise to a diagonal covariance matrix.[8] Clearly, relevant information might be discarded from the data when correlations are ignored. Another (perhaps more often used) strategy in chemometrics

 Copyright © 2017 John Wiley & Sons, Ltd. 

to obtain more efficient estimates is to assume that the covariance matrix has a low-dimensional structure.[9] This means that a few principal components (PCs) explain a large percentage of variance, and the analysis is restricted to this low-dimensional space. Again, relevant information might be contained in the discarded PCs. Larger and noisier data, such as encountered in –omics, are less likely to completely respect this assumption of low dimensionality. Additionally, the estimate of the explained variance for each PC is based on the sample eigenvectors of the covariance matrix. It is well known that these are usually grossly overestimated in high-dimensional data meaning that the percentage of explained variance of the corresponding component is overoptimistic.[13,14] This further hampers selection of the relevant PCs.

In recent years, much effort in fields such as statistics, machine learning, and econometrics has focused on developing improved (regularized) estimators of covariance or precision for high-dimensional data.[12,15,16] This paper provides a selective overview of 3 families of regularization approaches (see Figure 1). The term regularization refers to the fact that these methods effectively constrain the solution space thereby reducing the risk of overfitting. These reviewed methods offer alternative (or complementary) strategies to deal with high-dimensional data in chemometrics. Many of the discussed estimators are computationally inexpensive and can be (or are) easily combined with well-known multivariate statistical techniques in high-dimensional problems to potentially obtain more interpretable and/or improved (in some relevant sense) models. This is demonstrated for analysis of 3 metabolomics data sets.

The paper is structured as follows. Section 2 first briefly reviews the properties of the sample covariance matrix in the high-dimensional setting. Next, an overview of the regularization strategies discussed in this paper is presented. Sections 3 to 6 subsequently present an overview of eigenvalue-shrinkage estimators, ridge-type estimators, sparse estimators of the covariance matrix, and sparse estimators of the precision matrix, respectively. In each section, an application of the estimators for the analysis of metabolomics data is presented. Section 7 discusses software availability. Finally, a summary and some suggestions for further research are presented in Section 8.

## 2 | THE SAMPLE COVARIANCE MATRIX

The covariance matrix ($\mathbf{\Sigma}$) and its inverse, the precision matrix ($\mathbf{\Sigma}^{-1}$), play a central role in many chemometric techniques.[8–10] Examples include canonical correlation analysis, GGM, hierarchical clustering, ordinary least squares regression, LDA, (generalized) partial least squares regression, PCA, MANOVA, mixture modelling, quadratic discriminant analysis (QDA), soft independent modelling of class analogy (SIMCA), and multivariate statistical process control (MSPC) techniques.[8–10] Usually, the sample estimate of the covariance matrix is used as an estimate of the population covariance matrix, which is given by

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c, \tag{1}$$

where $\mathbf{X}_c$ corresponds to the mean-centred data matrix (samples in rows and variables in columns) and $n$ indicates the total number of samples.[10,12] This estimator has a number of desirable properties in the low-dimensional setting.[10,12] For example, it is unbiased, closely related to the maximum likelihood estimator (see below), and its eigenvalues and eigenvectors are closely related to their population counterparts. Additionally, $\mathbf{S}$ is invertible, and therefore, the sample estimate of the precision matrix ($\mathbf{S}^{-1}$) can easily be obtained.

### 2.1 | The sample covariance estimator is a poor estimator for high-dimensional data

Estimation of the covariance matrix requires the determination of $(p^2+p)/2$ parameters ($p$ is the number of variables in the data). Therefore, the sample estimator suffers from significant drawbacks in the high-dimensional and/or small sample setting where the number of samples is similar to, or much smaller, than the number of variables.[12,17] This is also true for the precision matrix. Usually, the so-called

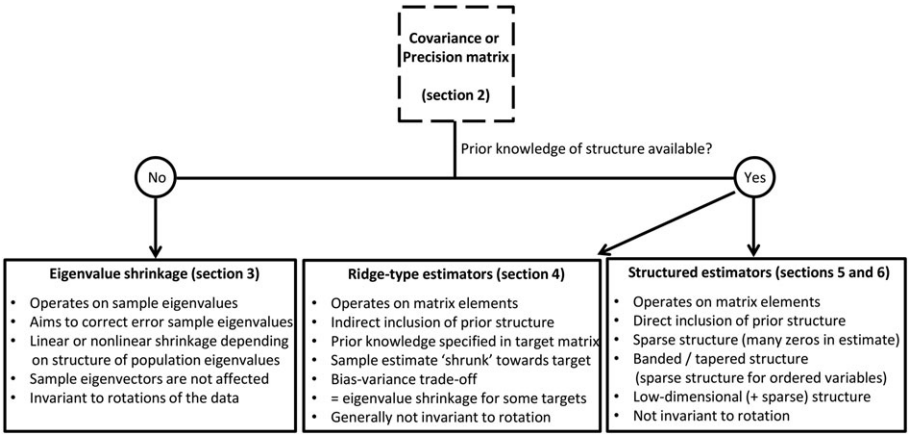


**FIGURE 1** An overview of the 3 families of covariance and precision matrix estimators that are discussed in this paper

concentration ($c = n/p$) is used to characterize the difficulty of covariance (and precision) estimation, since the effects of sample size and dimensionality are interdependent. Small values for $c$ indicate small sample sizes and/or high dimensionality.

For small values of $c$, the covariance matrix is potentially estimated with large error.[12,17,18] For example, consider the eigenvalues of **S** in Figure 2.[18] A systematic error can be observed even when $c = 10$ (the data contain 10 times more samples than variables): the large eigenvalues are overestimated and the small eigenvalues are underestimated. This effect is generally obtained in high-dimensional data and becomes more severe when $c$ becomes smaller.[17–20] When $c$ is not much larger than 1, the sample estimate is numerically ill conditioned, ie, inverting it to estimate the precision matrix will amplify estimation error. Additionally, when $c < 1$, matrix **S** loses full rank. This can be seen in Figure 2 as a growing number of eigenvalues are 0. This has several undesirable consequences. First, **S** is not positive definite anymore, and second, it cannot be inverted as **S** becomes singular.[18]

The systematic error of the sample eigenvalues has a negative impact on many chemometric techniques. For example, estimates of percentage of explained variance in PCA (and therefore principal component selection rules such as screeplots) or estimates of within-group variance in MSPC, LDA, and MANOVA might be misleading since they are based on the sample eigenvalues.[10] Moreover, techniques that use the precision matrix such as MSPC, LDA, and MANOVA are not applicable at all when $c < 1$ since the sample covariance matrix cannot be inverted.[8–10] Not only the eigenvalues of **S** are estimated with error for low concentrations $c$ but also the direction of the sample eigenvectors can differ greatly from their population counterparts.[21,22] This is clearly problematic in PCA where the sample eigenvectors correspond to the loadings of the PC. Another example can be found when considering the matrix elements $s_{ij}$ (indices $i$ and $j$ indicate the $i$th row and $j$th column in **S**) or $s_{ij}^{-1}$ themselves. These can be very noisy (estimated with large error), which can, for example, be problematic when constructing network models based on correlations or partial correlations.[22]

## 2.2 | Approaches to covariance and precision matrix estimation

In recent years, a wide variety of methods for improving (in some relevant sense) upon the sample covariance estimator (or the precision matrix estimator) have been proposed in the literature.[12,15,16] As shown in Figure 1, this paper selectively covers the families of eigenvalue-shrinkage estimators, ridge-type estimators, and structured estimators. The methods are relatively straightforward to understand, applicable to high-dimensional data, computationally inexpensive, and easy to combine with many chemometric techniques. An overview of other approaches such as Bayesian approaches and methods based on the generalized linear model can be found in the following articles and references therein.[12,15,16,23]

## 2.3 | Eigenvalue-shrinkage estimators

As shown earlier in Figure 2, the sample eigenvalues of the covariance matrix tend to be overdispersed for high-dimensional data. Shrinkage approaches retain the sample eigenvectors but aim to adjust (shrink) the sample eigenvalues in a data-dependant manner to correct for their overdispersion (bias).[11,24] This results in an improved covariance matrix (and precision matrix) estimate. Shrinkage estimators are rotationally invariant: rotating the data in some way leads to the same rotation being applied to the shrinkage estimate. This property is preferable when no a priori information of the structure of the covariance matrix is available. However, the methods rely on the (debatable) premise that the eigenvector basis is estimated correctly.

## 2.4 | Ridge-type estimators

Often, it is reasonable to assume that the covariance (or precision) matrix has a particular structure that can be taken advantage of. In this scenario, it seems natural to apply regularization directly to the elements of the matrix. Ridge-type estimators take a "weighted average" of the sample estimate with a target matrix, ie, the elements of the sample matrix are shrunk towards the values in the target.[25] The target is a

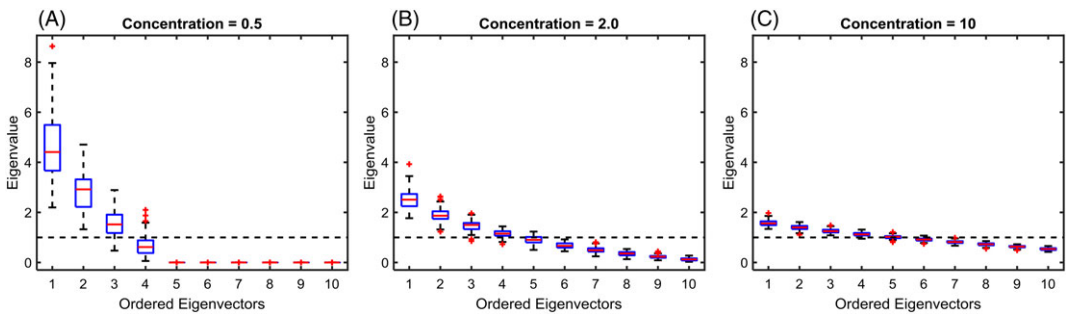


**FIGURE 2** Boxplots of the eigenvalues of the sample covariance matrix for simulated data with 10 variables and, A, 5 samples, B, 20 samples and, C, 100 samples over 100 simulations. The dashed line indicates the population eigenvalues. The simulated samples were drawn from a multivariate Gaussian distribution with zero mean and the identity matrix as covariance matrix

low-variance high-bias estimator of the covariance matrix. Therefore, ridge-type regularization can be seen as optimization of the bias-variance trade-off.[18] An example of a target matrix is the identity matrix. In this case, ridge-type regularization is closely related to shrinkage of eigenvalues.[11] Ridge-type estimators, however, allow for other targets and therefore allow for improving the estimates of the eigenvectors as well.

### 2.5 | Structured estimators

An alternative way to come up with improved estimators is to directly incorporate prior knowledge of structure in the estimation process. This way, the space of possible solutions is reduced mitigating the risk of overfitting. Structured estimators can assume different kinds of structures such as sparseness, bandedness, or a low-dimensional structure.[12,26,27] It is hoped that the true underlying matrices indeed conform to these structures. Especially, the development of sparse estimators has received a lot of attention in the literature. Here, it is assumed that many elements in the population covariance or precision matrix are zero.[12,26,27] In the context of multivariate Gaussian data, a zero in the covariance matrix corresponds to a pair of variables that are marginally (linearly) independent, while a zero in the precision matrix corresponds to a pair of variables that are linearly independent conditional on the other variables.[28] Therefore, sparse estimators offer a useful tool for exploring the variable dependence structure in the data (see Section 6.5 for more details). Note that although ridge-type estimators may shrink towards a sparse target, they do not generally produce sparse estimates. In contrast to ridge-type estimators, however, not all structured estimators are guaranteed to provide well-conditioned estimates, which can be problematic in practical applications. Finally, the sparsity notion is not adaptable to strongly correlated data sets.

### 2.6 | Penalized estimation of the covariance and precision matrix

Improved estimators for the covariance or precision matrix for high-dimensional data can be derived in different ways. One popular approach is to specify the estimator as a constrained (penalized) minimization or maximization problem.[12,16] The penalty effectively constrains the solution space, thereby mitigating the risk of overfitting. By using different penalties, estimators from each of the families of methods shown in Figure 1 can be obtained. For example, minimization of the following log likelihood (the loss function) leads to the maximum likelihood estimate of the covariance matrix ($\widehat{\mathbf{\Sigma}}_{ML} = \frac{n-1}{n}\mathbf{S}$):

$$\widehat{\mathbf{\Sigma}} = \arg\min_{\mathbf{\Sigma}} \mathcal{L}(\mathbf{\Sigma}, \mathbf{S}) = \arg\min_{\mathbf{\Sigma}} \log|\mathbf{\Sigma}| + tr\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right), \quad (2)$$

where $\mathcal{L}(\mathbf{\Sigma}, \mathbf{S})$ indicates the kernel of the log likelihood function. Here, the log likelihood function is based on the assumed Gaussian distribution of the data.[10,29] An improved estimator can be obtained by adding a penalty function:

$$\widehat{\mathbf{\Sigma}}(\delta) = \arg\min_{\mathbf{\Sigma}} \mathcal{L}(\mathbf{\Sigma}, \mathbf{S}) + p_\delta(\mathbf{\Sigma}), \quad (3)$$

where $p_\delta(\mathbf{\Sigma})$ is a penalty and $\delta$ is a tuning parameter that controls the "strength" of the penalty. For example, Equation 3 with penalty $p_\delta(\mathbf{\Sigma}) = \delta tr(\mathbf{\Sigma}^{-1})$ essentially corresponds to an eigenvalue-shrinkage estimator of the covariance matrix (see Sections 3.1 and 4).[30] Larger values for $\delta$ lead to a larger adjustment of the sample eigenvalues. In contrast, the use of an L1-norm penalty ($p_\delta(\mathbf{\Sigma}) = \delta\|\mathbf{\Sigma}\|_1 = \sum_{i,j}|\sigma_{ij}|$) leads to a sparse estimator for the covariance matrix.[31] Larger values for $\delta$ lead to sparser estimates. Increasing values for $\delta$ lead to estimators with lower variance (better behaved in high-dimensional data), but at the same time, relevant structure in the matrix might be lost. Many options for choosing an optimal value for the tuning parameter are available such as cross-validation (eg, with respect to the original loss in Equation 2), or a model-selection approach such as the Bayesian information criterion. Clearly, other criteria such as cross-validated classification accuracy can be used as well when the regularized estimator is used in combination with a specific chemometric technique such as LDA.

A disadvantage of likelihood-based approaches is that the estimators tend to perform worse if the data do not meet the distributional assumption, typically a multivariate Gaussian distribution.[12] Additionally, the penalized likelihood function is not always convex/concave (for example, Expression 3 with $\mathcal{L}(\mathbf{\Sigma}, \mathbf{S}) = \log|\mathbf{\Sigma}| + tr\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right)$ in combination with an L1-norm penalty), which increases the computational complexity of the estimator.[31] The Frobenius loss function is a popular alternative to the likelihood loss function to deal with these problems[12]:

$$\widehat{\mathbf{\Sigma}}(\delta) = \arg\min_{\mathbf{\Sigma}} \frac{1}{2}\|\mathbf{S}-\mathbf{\Sigma}\|_F^2 + p_\delta(\mathbf{\Sigma}). \quad (4)$$

Here, $\|\bullet\|_F^2$ corresponds to the squared Frobenius norm of its matrix argument. Frobenius loss has a number of attractive properties such as being convex (easy to solve with low computational complexity), nonparametric (no knowledge of distributional assumptions required), and its ease of implementation (quite often an analytical solution to the penalized problem is available).

## 3 | SHRINKAGE APPROACHES

Shrinkage of eigenvalues is the oldest approach to regularization of the covariance matrix.[32] No specific assumptions regarding the structure of the covariance matrix are made by these techniques. Instead, shrinkage approaches aim to correct the distorted eigenvalues structure of $\mathbf{S}$ (and $\mathbf{S}^{-1}$).[11] The term shrinkage refers to these methods essentially pulling (shrinking) the highest sample eigenvalues

downwards and the lowest eigenvalues upwards. As mentioned above, the sample eigenvectors are not changed. Because of this, shrinkage estimators are invariant to rotations of the data.[11]

Consider the spectral decomposition of the sample covariance matrix given by[12]

$$\widehat{\mathbf{\Sigma}} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathbf{T}} = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i^T, \quad (5)$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues $\lambda_i$, and $\mathbf{P}$ is an orthogonal matrix of normalized eigenvectors $\mathbf{e}_i$ (column vectors). This decomposition is familiar from techniques such as PCA where $\boldsymbol{e}_i$ are the loadings of the PCs and $\lambda_i$ is the explained variance. Shrinkage approaches aim to regularize the covariance matrix by applying a linear or nonlinear function $\vartheta(.)$ to the sample eigenvalues such that their estimation error is reduced (see Figure 2). Relatively little work has been performed on estimating the precision matrix directly, and in practical application, often the inverse of the shrunken covariance estimate is used.[18,33]

### 3.1 | Linear shrinkage

One of the first (and most popular) shrinkage methods for high-dimensional data was introduced by Ledoit and Wolf.[11,34] The Ledoit–Wolf linear shrinker (LW-LIN) shrinks the sample eigenvalues towards a central value in a linear way:

$$\vartheta(\lambda_i, \delta) = \delta\bar{\lambda} + (1-\delta)\lambda_i, \quad (6)$$

where $\bar{\lambda}$ indicates the average sample eigenvalue. Note that this approach pulls the highest sample eigenvalues downwards and the lowest ones upwards. In Figure 3, it can be seen that the shrunken eigenvalues are much closer to the "true" values compared to those of the sample covariance matrix. The tuning parameter $\delta$ $(0 \leq \delta \leq 1)$ controls the amount of shrinkage that is applied. When the bias in the sample eigenvalues increases, we expect $\delta$ to increase. When $\delta = 0$, the sample eigenvalues are used, and when $\delta = 1$, all eigenvalues are set to the average eigenvalue. The (data dependent) optimal amount of shrinkage is most likely somewhere in between these extremes and can, for example, be determined by cross-validation.[25,35] Unfortunately, such an approach is relatively computationally intensive.

Ledoit and Wolf showed that an analytical expression for the optimal amount of shrinkage can be obtained using a loss function very similar to the Frobenius loss shown in Equation 4.[11,34] This way, the optimal amount of shrinkage can easily be determined in a data-driven manner. Also, the resulting estimate is guaranteed to be positive definite.[11] The LW approach, however, requires information on the unknown population covariance matrix $\mathbf{\Sigma}$.[11] Schafer and Strimmer show that in practice, the unknown information from $\mathbf{\Sigma}$ can well be approximated from the sample estimate ($\mathbf{S}$).[18]

The linear eigenvalue-shrinkage estimator (Expression 6) has 2 alternative interpretations (see Section 4): (1) It can also be seen as a bias-variance trade-off between using the sample covariance matrix as covariance estimate or a multiple of the identity matrix (an estimate that assumes that all variables are uncorrelated)[11] and (2) it can also be viewed as a penalized maximum likelihood estimate.[30] The likelihood-based framework allows for likelihood-based approaches to optimize the shrinkage parameter such as cross-validation (with respect to the original likelihood, see Equation 2).[30] Using the penalized likelihood framework, van Wieringen et al show that linear shrinkage of eigenvalues not only improves upon the sample covariance estimator but also greatly improves estimation of the precision matrix.[25]

### 3.2 | Nonlinear shrinkage

Ledoit and Wolf showed that linear shrinkage captures almost all possible improvement over $\mathbf{S}$ when the concentration $(c = n/p)$ is small and/or when the population eigenvalues are similar in magnitude. However, when the population eigenvalues are dispersed (eg, a few PCs describe most variation in the data, or the eigenvalues have a staircase structure), linear shrinkage only improves upon the sample covariance matrix only slightly.[21,24] Random matrix theory shows that eigenvalue shrinkage is a fundamentally nonlinear

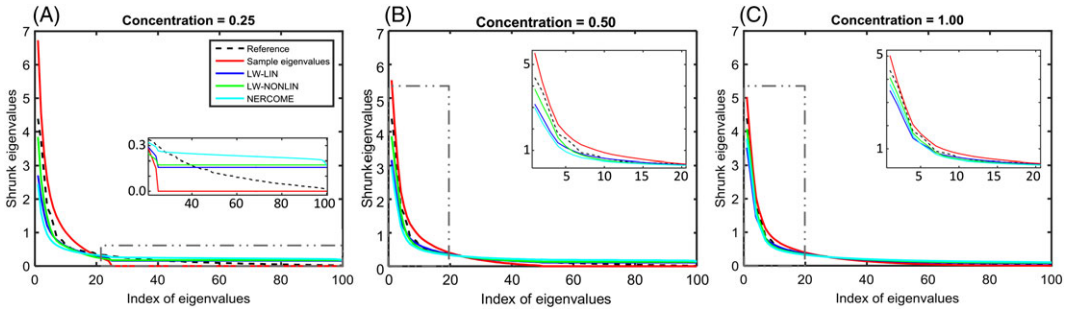


**FIGURE 3** Comparison of the average eigenvalues of the sample, linear (Ledoit–Wolf linear shrinker), and nonlinear (Ledoit–Wolf nonlinear shrinker and nonparametric eigenvalue-regularized covariance matrix estimator) covariance estimators for UHPLC-MS data with 100 peaks and, A, 25 samples, B, 50 samples and, C, 100 samples over 100 repetitions in which the samples were randomly selected from the data matrix. The "reference" eigenvalues were estimated using all 1189 samples. Note that LW-LIN was used in combination with the Ledoit–Wolf analytical expression for the optimal amount of shrinkage (see Section 3.1) UHPLC-MS indicates Ultra-High-Performance Liquid Chromatography - Mass Spectrometry

problem of which linear shrinkage is only an approximation.[24] Therefore, nonlinear shrinkage of the sample eigenvalues offers a great route to further improve the shrinkage estimator.[21,24] Whereas linear shrinkage estimators shrink all eigenvalues uniformly, nonlinear approaches potentially shrink each eigenvalue differently.

Won et al consider nonlinear eigenvalue shrinkage from the penalized likelihood perspective (see Equations 2 and 3).[36] A condition number constraint (penalty) is used. The condition number is the ratio between the largest and smallest eigenvalue of the estimate. Won et al show that their penalized likelihood condition number regularized (CNR) estimator is also a nonlinear shrinkage estimator. More specifically, the CNR estimate can be obtained by using Equation 5 and truncating the sample eigenvalues that are larger than $\kappa_{max}\tau^*$ to $\kappa_{max}\tau^*$ and those smaller than $\tau^*$ to $\tau^*$. The scalar $\kappa_{max}$ (maximum allowed condition number) is determined by cross-validation, and $\tau^*$ (minimum allowed eigenvalue) is determined directly from the data given $\kappa_{max}$. The CNR estimator gives most improvement compared to linear shrinkage when a few population eigenvalues are much larger than the others (ie, most of the variation in the data can be explained by a few PCs).[36] These effects diminish when this is not the case. Recently, the CNR estimator was extended to obtain a covariance (or precision) matrix estimate that is well-conditioned and sparse.[37]

To handle scenarios where CNR loses its competitive edge over linear LW shrinkage, Chi et al propose the so-called covariance estimate regularized by nuclear norms (CERNN) method.[38] They penalize the log likelihood (Equation 3) with a nuclear norm penalty, which is essentially a penalty on sums of eigenvalues. This penalty effectively steers the eigenvalues of the estimate away from the extremes 0 and $\infty$. The following nonlinear eigenvalue shrinker is obtained:

$$\vartheta(\lambda_i,\delta)=\frac{-n+\sqrt{n^2+4\delta\alpha[n\lambda_i+\delta(1-\alpha)]}}{2\delta\alpha}, \qquad (7)$$

where the scalar $\delta$ (optimized with cross-validation) controls the amount of shrinkage, and scalar $\alpha$, which is estimated directly from the data, ensures that the eigenvalues are shrunk towards the average sample eigenvalue. The CERNN method shrinks extreme sample eigenvalues in a similar manner as CNR but less drastically, and shrinks intermediate eigenvalues similarly to the LW-LIN linear shrinkage estimator. By simulation, Chi et al show that CERNN outperforms CNR when there is a need to shrink all eigenvalues (similar to LW-LIN), but with extra emphasis on extreme eigenvalues (unlike LW-LIN).[38] In their simulation, CNR only outperformed CERNN when the majority of the variation in the data was explained by a single PC.

An alternative approach to nonlinear eigenvalue shrinkage is taken by Ledoit and Wolf[21,24] and Lam.[39] They use tools from random matrix theory, specifically the Marcenko–Pastur equation, which describes the complex (and nonlinear) relationship between the population and sample eigenvalues. More specifically, Ledoit and Wolf developed the QuEST function (quantized eigenvalues sampling transform), which uses the Marcenko–Pastur relationship and thereby allows for estimation of the population eigenvalue spectrum based on the sample eigenvalues.[21,24] This estimate can subsequently be combined with the sample eigenvectors to obtain an estimate of the covariance or precision matrix. We will refer to this approach as Ledoit–Wolf nonlinear shrinker (LW-NONLIN). The QuEST solver requires nonconvex optimization, which makes the LW-NONLIN method computationally quite intensive.[40] Recently, Lam proposed the nonparametric eigenvalue-regularized covariance matrix estimator (NERCOME) and showed that its eigenvalues asymptotically approach those found by LW-NONLIN (it is questionable how useful this asymptotical property is in practice given the low sample sizes encountered in chemometrics).[39] The estimator is nonparametric and, compared to LW-NONLIN and the associated QuEST function, has a particularly simple implementation. Nonparametric eigenvalue-regularized covariance matrix estimator is based on splitting the data. The first split is used to estimate the eigenvectors of the covariance estimate, and the second split is used to estimate the eigenvalues:

$$\widehat{\mathbf{\Sigma}}=\mathbf{P}_1 diag\left(\mathbf{P}_1^T\mathbf{S}_2\mathbf{P}_1\right)\mathbf{P}_1^T, \qquad (8)$$

where $\mathbf{P}_1$ contains the eigenvectors of the sample covariance matrix ($\mathbf{S}_1$) of the first split, and $\mathbf{S}_2$ corresponds to the sample covariance estimate of the second split. Note that $diag\left(\mathbf{P}_1^T\mathbf{S}_2\mathbf{P}_1\right)$ is an estimate of the eigenvalues based on the eigenvectors of the first split and the sample covariance matrix of the second split. The NERCOME method makes use of $\mathbf{P}_1$ and $\mathbf{S}_2$ that are independent to regularize the eigenvalues. Lam argued that a limited number of samples should be used to estimate $\mathbf{S}_2$ (eg, $\max\left(30,\frac{n}{10}\right)$ samples). The remaining samples should be used to estimate $\mathbf{P}_1$. Additionally, they showed that the estimator can be further improved by averaging the covariance estimate over multiple splits in the data. By simulation, Lam compared NERCOME to NONLIN and principal orthogonal complement thresholding (POET; a method that will be discussed in Section 5.5).[39] The 3 approaches performed similarly when the population eigenvalues had a staircase-like structure where 40% of the population eigenvalues were small, 20% had intermediate values, and 40% were large. Since Ledoit and Wolf showed that LW-NONLIN outperformed the linear LW shrinker, it is expected that NERCOME does so as well.

Interestingly, Lam showed that high-dimensional data with an intrinsic low dimensional structure with only a few large eigenvalues (which is often assumed in chemometrics) can render the LW-NONLIN shrinker incorrect.[39] In contrast, NERCOME was able to deal with such data.[39] Unfortunately, no comparison between NERCOME and CNR (which, as mentioned above, performs best for data with a low-

dimensional structure) has been performed in the literature. It would be interesting to perform such a comparison, including the S-POET estimator (discussed in Section 5.5) and the nonlinear shrinkers that were proposed for such data by Donoho.[41] With respect to computational complexity, NERCOME is outperformed by LW-LIN, CNR, and CERNN since these approaches only require a single eigendecomposition of the sample covariance matrix.

Since the optimal type of eigenvalue shrinkage depends on the structure of the population eigenvalues (which is unobservable), we suggest to use nonlinear shrinkage approaches in practice. However, the computational complexity of linear shrinkers is much lower, which can make them the preferred option. Typically, linear shrinkage estimates already greatly improved upon using the sample covariance estimate.[24,42]

### 3.3 | Example: shrinkage estimators, PCA, and MANOVA

We demonstrate the application of the linear and nonlinear eigenvalue shrinkage estimators to Ultra-High-Performance Liquid Chromatography - Mass Spectrometry (UHPLC-MS) metabolomics data. The data originate from the HUSERMET study, which aimed to quantify the normal metabolic variation in serum of healthy humans.[43]

First, XCMS[44] was applied for peak deconvolution and alignment, followed by sample filtering (only peaks that were present in at least 80% of the samples were retained), probabilistic quotient normalization, missing value imputation using kNN (k = 5), and a generalized logarithmic transformation.[45] The final data matrix contained 1189 rows (samples) and 2178 columns (peaks). A subset of the 100 peaks with the highest variance was considered for further analysis. This way, a "traditional" data matrix with a (much) larger number of samples than variables was obtained ($c = 12$). The eigenvalues of the sample covariance estimate of this matrix were be used as "reference" values since no knowledge of the population eigenvalues was available.

Figure 3 compares the sample covariance estimator to shrinkage estimators for different sample sizes (as indicated by the concentration value). The overdispersion (bias) of the sample eigenvalues is clearly more pronounced for low concentrations (on average, the sample eigenvalues differ more greatly from the reference values). Additionally, as expected, the variance of the sample eigenvalue estimator is larger for low sample sizes as well (not shown). Note that that sample covariance matrix was singular for $c < 1$ (panels A and B) since a number of sample eigenvalues were zero. In contrast, the covariance estimates obtained by the shrinkage estimators were nonsingular for any value of $c$. Moreover, it can be seen that the sample eigenvalue dispersion was corrected by both the linear and nonlinear shrinkage approaches. Especially, the eigenvalues obtained by the LW-NONLIN estimator matched the reference eigenvalues quite closely (note that this comparison is not completely fair since the reference eigenvalues are not the population eigenvalues). For high sample sizes (eg, 400 samples) the LW-NONLIN and NERCOME estimators provided nearly equivalent results (not shown). However, for lower sample sizes as in Figure 3, the results of NERCOME were generally in between those of LW-NONLIN and those of the linear shrinker. This is attributed to the fact that the sample size was too low in this case to be able to reliably split the data in 2 halves (see Section 3.2). This can be seen as a drawback of the NERCOME approach. Given that LW-LIN is only an approximation to a nonlinear problem, it is no surprise that the eigenvalues obtained by LW-LIN matched the reference eigenvalues the least of all shrinkers. However, considerable improvement compared to the sample estimate could still be observed.

#### 3.3.1 | Principal component analysis

Next, we consider PCA. As mentioned earlier, estimates of percentages of explained variance in PCA might be misleading since they are based on the sample eigenvalues, which are biased (see Figure 3). Typically, the variance of the first PCs is overestimated (the eigenvalues are biased upwards) while the variance of the last PCs is underestimated (eigenvalues are biased downwards). Consequently, with most rules to decide upon the number of PCs to retain, eg, rules based on the cumulative percentage of explained variance, typically too few components are selected.[21]

Interestingly, even if the population eigenvalues were known, they could not be used for PC selection because they are obtained from the population eigenvectors.[21] Ideally, since in practice the population eigenvectors are unknown, the so-called out-of-sample variance estimate based on the sample eigenvectors should be used, which is given by $\lambda_i^{OOS} = \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{e}_i$.[21] This corresponds to the variance in the population that is explained by the sample eigenvectors (loadings). Estimation of the out-of-sample variance, however, requires the population covariance matrix, which is unknown. Interestingly, as shown in Figure 4, the estimates of explained variance by the shrinkage approaches approached the out-of-sample variances quite well. This was also true for lower sample sizes of, eg, 25. In contrast, the sample estimates and the reference eigenvalue (a proxy for the population eigenvalues) were biased upwards and selected too few PCs. This effect became more pronounced for lower sample sizes. Given the larger number of components that would be selected by, eg, a 70% explained variance rule, it is questionable whether total variance explained is the right criterion for selection here. Many other PC selection rules, however, are also based on the eigenvalues (explained variance) and are expected to suffer from the same pitfalls.

#### 3.3.2 | Multivariate analysis of variance

Finally, we consider a supervised problem. For this purpose, 2 groups of 25 samples were randomly selected from the

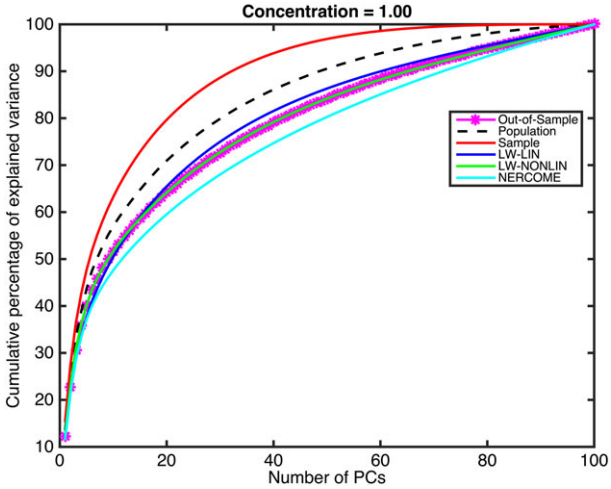


**FIGURE 4** Percentage of variance explained by the top $k$ principal components based on the eigenvalues of different covariance estimates for UHPLC-MS data with 100 peaks and 100 samples. The average is based on 100 repetitions in which the samples were randomly selected from the data matrix. The "reference" and "out-of-sample" estimates were based on all 1189 samples. PCs indicates principal components; UHPLC-MS, Ultra-High-Performance Liquid Chromatography - Mass Spectrometry; LW-LIN, Ledoit-Wolf linear shrinker; LW-NONLIN, Ledoit-Wolf nonlinear shrinker; NERCOME, nonparametric eigenvalue-regularized covariance matrix estimator

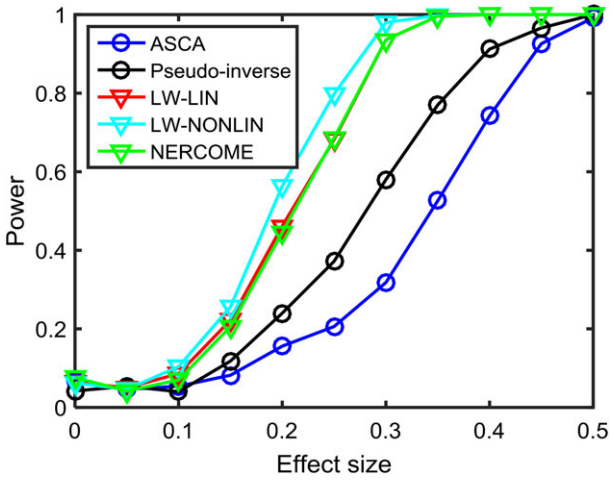


**FIGURE 5** Analysis of the UHPLC-MS data by multivariate analysis of variance using various regularization approaches. The introduced group difference (effect size) is plotted against the percentage of cases for which a significant difference was observed (power). For each method, a permutation test in combination with the Wilk's lambda test-statistic was used to determine the statistical significance ($\alpha = 0.05$). LW-LIN indicates Ledoit–Wolf linear shrinker; LW-NONLIN, Ledoit–Wolf nonlinear shrinker; NERCOME, nonparametric eigenvalue-regularized covariance matrix estimator; UHPLC-MS, Ultra-High-Performance Liquid Chromatography - Mass Spectrometry; ASCA, ANOVA simultaneous component analysis

reduced UHPLC-MS data table (with 100 peaks) discussed above. A constant value ranging from 0 to 0.5 was added to the intensities of the first 20 peaks in group 1. This process was repeated 500 times. In each repetition, MANOVA was used to test if the groups were significantly different.

Test statistics in MANOVA are based on the eigenvalues of matrix $\mathbf{R} = \mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{B}$ is the between-group scatter matrix and $\mathbf{W}$ the within-group scatter matrix. See this 1 study[46] for more details. Note that the sample estimate of matrix $\mathbf{W}$ is (up to a constant) a sample covariance estimate (the one discussed in Figure 3). This matrix is singular for high-dimensional data, and MANOVA cannot be applied. In Figure 5, we replaced the sample estimate of $\mathbf{W}$ by an eigenvalue-shrinkage estimator to remedy this issue. Note that the inverse of the shrinkage estimate was used to estimate matrix $\mathbf{R}$. This approach using (linear) shrinkage is known as regularized MANOVA (rMANOVA), see this 1 study[47] for more details. To the best of our knowledge, the combination of nonlinear shrinkage with MANOVA is novel. As shown in the figure, rMANOVA using linear or nonlinear eigenvalue shrinkage approaches greatly outperformed popular alternatives such as ANOVA simultaneous component analysis (ASCA)[48] or using the Moore–Penrose pseudoinverse (of $\mathbf{W}$)[49] to estimate $\mathbf{W}^{-1}$. Regularized MANOVA with LW-NONLIN shrinkage had slightly higher power compared to LW-LIN and NERCOME. However, the relative improvement greatly depended on the direction of group separation, and often, the methods performed quite similarly (results not shown). It is questionable whether the use of LW-NONLIN offered a practical advantage given its (much) higher computational costs. Ridge-type estimators (see Section 4.1) might offer a computationally less intensive route to improve upon rMANOVA with linear shrinkage.

## 4 | RIDGE-TYPE ESTIMATORS

Ridge-type estimators reduce the effective number of degrees of freedom by taking a weighted average of the sample estimate with a target matrix, ie, the elements of the sample matrix are shrunk towards the values in the target.[11,18,25,30] This way, a better estimator for high-dimensional data is obtained. As will be seen below, these estimators share many similarities with the rather ad hoc approach to regularization that is taken in ridge regression ($\widehat{\mathbf{\Sigma}} = \mathbf{S} + \kappa\mathbf{I}$), hence, the name ridge-type estimators.[25] Ridge-type estimators are appealing in practice because of their simplicity and ease of implementation. Most work has focused on estimation of the covariance matrix. An estimate for the precision matrix is often obtained by taking the inverse of the ridge covariance estimate.

Typically, the target matrix ($\mathbf{T}$) is a low-variance high-bias estimator of the covariance matrix, ie, a matrix with a "simple" (low dimensional) structure.[18] In contrast, the sample covariance matrix estimator is unbiased but has high variance. Therefore, ridge-type estimators can be motivated from a bias-variance trade-off as they seek to balance matrices $\mathbf{T}$ and $\mathbf{S}$. For example, consider a multiple of the identity matrix as a target[11,18]:

$$\widehat{\mathbf{\Sigma}}(\delta) = \delta\frac{tr(\mathbf{S})}{p}\mathbf{I} + (1-\delta)\mathbf{S}. \quad (9)$$

The parameter $\delta$ controls the amount of regularization that is applied. The choice of penalty value is crucial. When $\delta$ is close to zero, the estimate is close to $\mathbf{S}$ and might therefore be ill conditioned. In contrast, by choosing $\delta$ too large (close to 1), the estimate is essentially the simple target and relevant information might be lost. In Equation 9, by increasing $\delta$, the sample covariance matrix is slowly shrunk towards the identity matrix. This target assumes uncorrelated variables (all off-diagonal elements in $\mathbf{I}$ in zero; all pairwise correlations between variables are zero). Correlation information is potentially lost by shrinking $\mathbf{S}$ towards $\mathbf{I}$, but at the same time, a much more stable estimate is obtained.[11] Many options for choosing $\delta$ are available. Ledoit and Wolf showed that under Frobenius loss, an analytical solution for the optimal value for $\delta$ can be obtained.[11,18] This is the same approach as was discussed in Section 3.1. This offers a computationally inexpensive approach to control the amount of shrinkage compared to, for example, cross-validation.

Besides the scaled identity matrix in Equation 9, other, possibly more realistic, targets ($\mathbf{T}$) can be used as well[18,50–52]:

$$\widehat{\mathbf{\Sigma}}(\delta) = \delta\mathbf{T} + (1-\delta)\mathbf{S}. \tag{10}$$

If the target is positive definite, the resulting ridge estimate (with optimal value for $\delta$) will be so as well.[52] Different targets have been proposed in the literature such as the scalar multiple of the identity matrix,[11] diag($\mathbf{S}$), a common correlation matrix (the same correlation between all pairs of variables),[18] a tapered matrix (see Section 5.4),[51] and a sparse matrix (see Section 5.1).[50] All these targets have analytical solutions (under Frobenius-like loss) that are based on the work of Ledoit and Wolf (see Section 3.1).[11,18,33,50–54] The choice of target should be guided by the presumed structure of the population covariance matrix. Any low-variance target will lead to an improved estimator (with respect to a relevant risk [expected loss] function) upon $\mathbf{S}$, although only a minor one in case of a misspecified target.[18,55,56] Often, it is difficult to identify a sensible target, and the ridge-type estimate may be misspecified. Because every target has a different bias-variance trade-off with respect to the unknown population covariance matrix, there is often no single ideal target.[55] Recently, Lancewicki et al and Bartz et al proposed the multitarget shrinkage estimator that allows for shrinkage of $\mathbf{S}$ towards multiple targets simultaneously.[55,56] Multitarget shrinkage leads to equal or improved estimates compared to single target shrinkage when multiple (sensible) targets are available, such as those listed above (identity matrix, common correlation, etc).[55,56] Although some of the targets need to be sensible, the approach is robust against misspecification of one (or multiple) of the targets.[55,56] In this sense, the method somewhat alleviates the problem of choosing a single sensible target and possible misspecification of this target.

Above, the ridge-type estimator has been proposed from a bias-variance point of view. In case of a single target that is a multiple of $\mathbf{I}$, it has been shown that the resulting ridge-type estimate is equal to the linear eigenvalue shrinker discussed in Section 3.1 (the estimator only changes upon the sample eigenvalues, but not the eigenvectors).[11] This offers a second interpretation of the estimator. Not all ridge-type estimators can be interpreted as eigenvalue-shrinkage estimators. For a general target $\mathbf{T}$ (eg, a sparse matrix estimate from Section 5), ridge-type shrinkage might result in (non) linear shrinkage of eigenvalues as well as a change in eigenvectors.[50] This can be an advantage in applications such as PCA where estimation of eigenvectors and eigenvalues is required. A third interpretation of ridge-type estimators is that of the solution to a penalized log-likelihood problem[25,30]:

$$\widehat{\Sigma}(\delta) = \arg\min_{\mathbf{\Sigma}} \log|\mathbf{\Sigma}| + (1-\delta)tr\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right) + \delta tr\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right). \tag{11}$$

Replacing $\mathbf{S}$ by $(1-\delta)\mathbf{S}$ in the original log-likelihood (2) and adding a penalty obtains this expression. The estimate can therefore be seen as a penalized maximum likelihood estimate. The behaviour of the penalty can easily be understood when $\mathbf{T}=\mathbf{I}$. In this case, it can be easily seen that that ill conditioned estimates that approach singularity are penalized more; the diagonal elements of $\widehat{\Sigma^{-1}}$ will approach infinity in this case, so the matrix trace (the penalty) will be large. Recently, van Wieringen et al discussed Equation 11 from the viewpoint of ridge regression and argued that the penalty is not completely equal to a "ridge" penalty.[25] They proposed an alternative ridge-type estimator:

$$\widehat{\Sigma}(\delta) = \left[\delta\mathbf{I} + \frac{1}{4}(\mathbf{S}-\delta\mathbf{T})^2\right]^{\frac{1}{2}} + \frac{1}{2}(\mathbf{S}-\delta\mathbf{T}). \tag{12}$$

A fast cross-validation procedure is used to optimize the tuning parameter $\delta$.[25] An estimate for $\widehat{\mathbf{\Sigma}}^{-1}$ can be obtained without inversion as $\widehat{\mathbf{\Sigma}}^{-1}(\delta) = \frac{1}{\delta}\left[\widehat{\mathbf{\Sigma}}(\delta) - (\mathbf{S}-\delta\mathbf{T})\right]$.[25] van Wieringen et al compared the 2-ridge–type estimators (Equations 11 and 12) in a simulation study.[25] They observed that the alternative ridge estimator compared favourably in several risk functions (eg, expected Frobenius loss) for small to intermediate values of $\delta$ (these are, in practice, usually the most interesting values) when the target adequately represented the population matrix. For large amounts of shrinkage, similar results were obtained by both approaches because they both shrink towards the same target.[25]

### 4.1 | Revisiting the MANOVA example

In example 3.3.2, the LW-LIN was used to improve MANOVA for analysis of high-dimensional data (see Figure 5). More specifically, linear eigenvalue shrinkage was used to improve upon the sample estimate of the within-group

scatter matrix **W**. The method was referred to as rMANOVA.[47] As mentioned above, the linear eigenvalue shrinker is a special case of a ridge-type estimator (Equation 10) using a multiple of the identity matrix as target. Therefore, rMANOVA can also be interpreted from a bias-variance trade-off point of view as it seeks to balance using matrix **W** (equal to [up to a constant] **S**) and the target matrix **T** that specifies a simple within-group scatter structure using Equation 9. Other targets, besides the identity matrix, can also be used in rMANOVA such as *diag*(**W**) or a sparse matrix.[47]

## 5 | SPARSE ESTIMATION OF THE COVARIANCE MATRIX

Sparse covariance matrix estimators assume that the population covariance matrix is (approximately) sparse, meaning that many of its off-diagonal elements are zero or nearly so. This approach effectively reduces the solution space of the covariance matrix estimator, which mitigates the risk of overfitting. Examples of a sparse structure are, for example, often encountered in genomics and metabolomics. Sparse models offer superior interpretability, because zeros in the covariance matrix correspond to pairs of variables that are uncorrelated. In the case of multivariate Gaussian data, this translates to marginal independence between these variables. Note that although ridge-type estimators may shrink towards a sparse target, they do not generally produce sparse estimates. In contrast to ridge-type estimators, however, not all structured estimators are guaranteed to provide well-conditioned estimates, which can be problematic in practical applications. If the population covariance matrix is indeed (moderately) sparse, many of the sparse estimators discussed below return not only improved estimates of the covariance matrix but also improved estimates of its eigenvectors and eigenvalues. This is a useful property for applications of the covariance matrix estimator in, for example, PCA.

### 5.1 | Thresholding

The most popular (and obvious) approach to sparse covariance matrices is thresholding, which sets small-valued off-diagonal elements in the sample covariance matrix to zero (see Figures 6 and 7).[26,57] This way, estimation of small elements is avoided so that noise does not accumulate. Thresholding estimators of the covariance matrix are the solution to the following penalized problem[15,26,58]:

$$\widehat{\boldsymbol{\Sigma}}(\delta) = \arg\min_{\boldsymbol{\Sigma}} \frac{1}{2}\|\mathbf{S}-\boldsymbol{\Sigma}\|_F^2 + p_\delta(\boldsymbol{\Sigma}). \tag{13}$$

In other words, they minimize the Frobenius distance between the sample covariance matrix and a sparse estimate, where the amount of sparsity is controlled by the penalty. Often, the thresholding is applied to the off-diagonal matrix elements only.[16] The penalty function can be expressed element-wise, therefore minimizing Equation 13 is the same as minimizing element-wise[15,26]:

$$\widehat{\sigma}_{ij} = \arg\min_{\sigma_{ij}} \frac{1}{2}\left(s_{ij}-\sigma_{ij}\right)^2 + p_\delta\left(\sigma_{ij}\right). \tag{14}$$

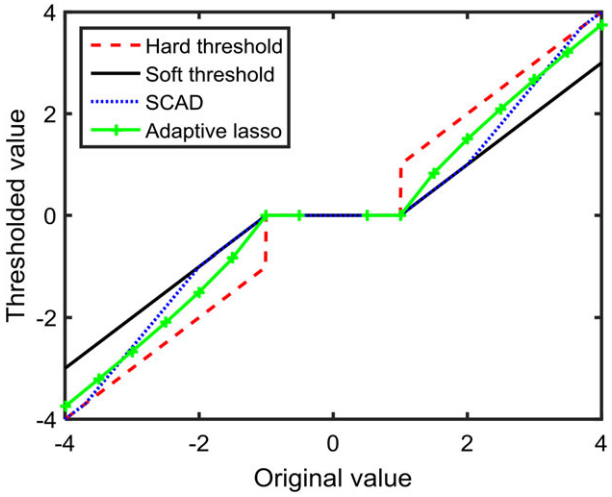


**FIGURE 6** Original values plotted against thresholded values for different thresholding operators SCAD indicates smoothly clipped absolute deviation

Often, penalties are used for which Expression 14 has a closed-form solution. Therefore, thresholding approaches carry almost no computational burden and scale well to data with extremely large dimensions. A list of popular penalties and their analytical solution is provided in Table 1.[15,16,26,59]

In Figure 6, it can be seen that hard thresholding (HT) sets all elements of the covariance matrix with an absolute value below threshold $\delta$ to zero. The thresholding-value $\delta$ should be larger than the expected estimation error of the matrix elements and can, for example, be found by cross-validation or using a false discovery rate approach.[18,57] Soft thresholding (ST) is related to HT, in particular to the choice in which elements to set to zero. However, the difference is that ST also shrinks elements above the threshold towards zero (see Figure 6). Often, HT tends to zero out too many elements, presumably because of its inability to shrink small values (that are just above the threshold).[15,26] In this sense, ST offers a better estimator. The ST estimator, however, introduces biases for the nonzero entries since they are always shrunk independent of their size (clearly, extremely large covariances should not be affected by thresholding).[15] The ST estimator often compensates for this bias by choosing a less sparse alternative.[8] The adaptive lasso (sometimes referred to as adaptive threshold (AT)) is similar to ST, but the threshold is adjusted for each entry in the covariance matrix using a weighting function. The idea is to apply a larger amount of shrinkage (larger weight) to smaller empirical covariances. This results in a procedure that is more similar to HT, where small elements are shrunk, but large ones are not. There are multiple choices for weight, for example, the reciprocal of $s_{ij}$.[31] The smoothly clipped absolute deviation (SCAD) is a linear interpolation

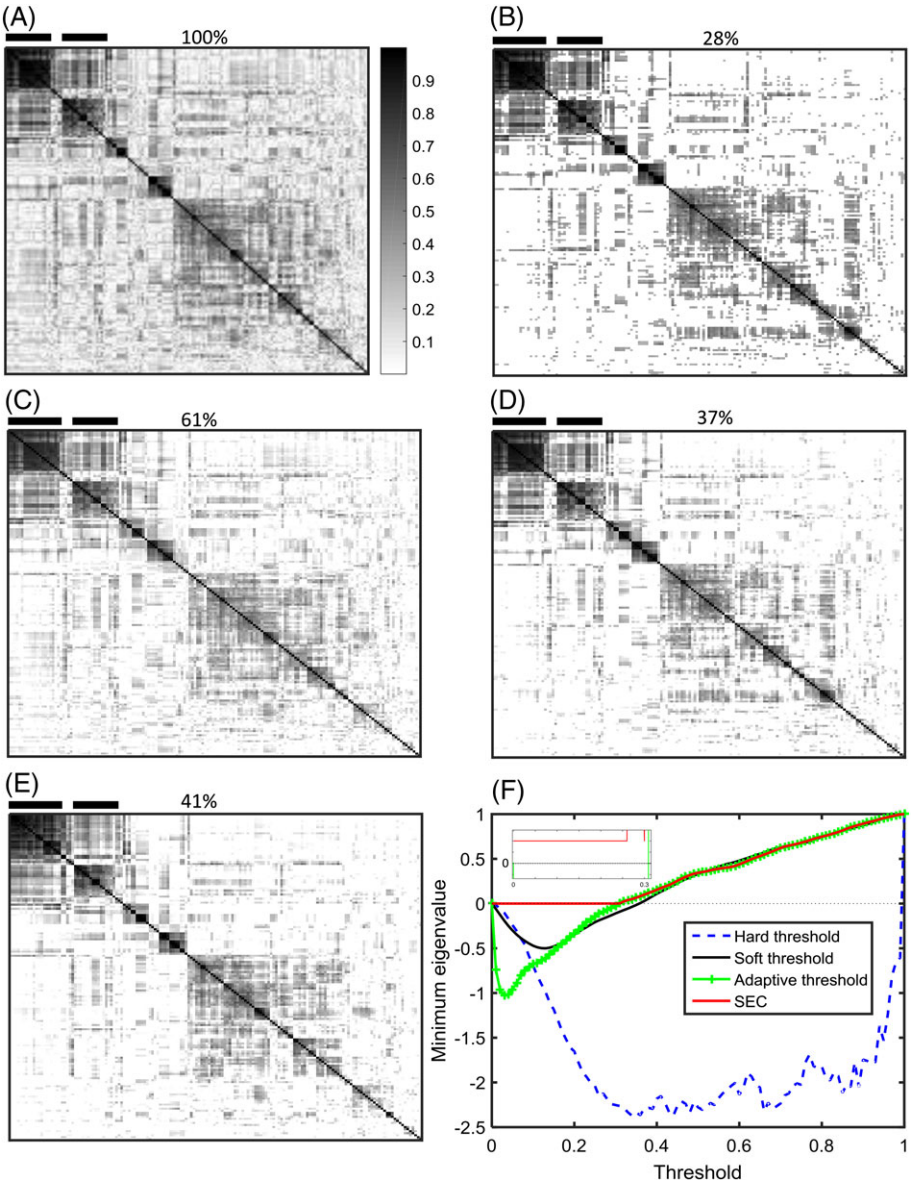


**FIGURE 7** Heat maps of the absolute values of estimated correlation matrices using (A-E) the sample covariance estimator, hard thresholding, soft thresholding, adaptive thresholding (adaptive lasso using the reciprocal of the sample correlation matrix elements as weight), and SEC (adaptive thresholding with positive definite constraint), respectively. The percentages indicate the number of nonzero elements in the matrix. Panel F displays the minimum eigenvalue of the 4 thresholding estimators against the threshold value. The horizontal black bars indicate the majority of the 40 variables with lowest *P* value (based on a *t* test) SEC indicates the "Sparse Estimation of the Correlation Matrix" estimator

**TABLE 1** Overview of popular thresholding penalties and their solution

| Thresholding type | Penalty ($p_\delta(\sigma_{ij})$) | Solution (thresholding operator $T(s_{ij}, \delta)$) |
|---|---|---|
| Hard thresholding | $\delta^2 - (\lvert\sigma_{ij}\rvert - \delta)^2 1(\lvert\sigma_{ij}\rvert < \delta)$ | $s_{ij} 1(\lvert s_{ij}\rvert > \delta)$ |
| Soft thresholding | $\delta\lvert\sigma_{ij}\rvert$ | $\mathrm{sgn}(s_{ij})(\lvert s_{ij}\rvert - \delta)_+$ |
| Smoothly-clipped absolute deviation (SCAD) | $\begin{cases} \delta\lvert\sigma_{ij}\rvert, \text{when } \lvert\sigma_{ij}\rvert \le \delta \\ \dfrac{(2a\delta\lvert\sigma_{ij}\rvert - \sigma_{ij}^2 - \delta^2)}{[2(a-1)]}, \text{when } \delta < \lvert\sigma_{ij}\rvert \le a\delta \\ \dfrac{(a+1)\delta^2}{2}, \text{when } \lvert\sigma_{ij}\rvert > a\delta. \end{cases}$ | $\begin{cases} \mathrm{sgn}(s_{ij})(\lvert s_{ij}\rvert - \delta)_+, & \text{when } \lvert s_{ij}\rvert \le 2\delta \\ \dfrac{\{(a-1)s_{ij} - \mathrm{sgn}(s_{ij})a\delta\}}{a-2}, & \text{when } 2\delta < \lvert s_{ij}\rvert \le a\delta \\ s_{ij}, & \text{when } \lvert s_{ij}\rvert > a\delta \end{cases}$ |
| Adaptive lasso | $\delta w(s_{ij})\lvert\sigma_{ij}\rvert$ | $\mathrm{Sgn}(s_{ij})(\lvert s_{ij}\rvert - \delta w(s_{ij})\lvert s_{ij}\rvert)_+$ |

$1(\bullet)$ is the indicator function, which is 1 if its argument is true and 0 if otherwise; $(x)_+ = 0$ if $x < 0$ ; and $x$ otherwise; $w(\bullet)$ indicates a weighting function that is decreasing with larger absolute values of its argument.

of ST up to $2\delta$ and hard thresholding after $a\delta$.[60] Similar to the adaptive lasso, the advantages of HT and ST are combined this way. Typically, parameter $a$ is set to 3.7 as was recommended by Fan and Lin.[60]

Rothman et al placed the above mentioned operators in a general framework resulting in the class of generalized thresholding estimators.[26] Generalized thresholding has the property that it estimates the true zeros in the covariance matrix with probability tending[26] to 1. Rothman et al showed for simulated data that for truly sparse covariance matrices, thresholding leads to an improved estimate compared to sample covariance estimate and that penalties that combine the advantages of hard and soft thresholding such as SCAD and adaptive lasso tend to perform best.[26] On similar grounds, Fan et al advocate using the SCAD or adaptive lasso.[15,16] However, other authors state that there are no clear theoretical or empirical results to favour a particular thresholding rule in all cases.[58] When the true covariance matrix is not sparse, generalized thresholding performs similarly to just using the sample covariance estimate.[26] An example might be encountered in near infrared spectroscopy where many wavelengths are highly collinear. In such a case, the ridge-type estimators discussed in Section 4.1, or a low-dimensional + sparse estimator (Section 5.5) might be preferable.

### 5.2 | Adaptive thresholding

Most thresholding approaches use universal thresholding rules, ie, they apply the same threshold level to all of the elements of the covariance matrix. However, the entries of the covariance matrix can have very different scales (the variances of some variables are much larger than those of other variables). Therefore, it makes sense to take this into account by applying a unique threshold to each element. This can potentially lead to more precise estimation as shown by Cai et al and can be achieved by adaptive thresholding similar to the adaptive lasso approach discussed above.[61] In this case, however, highly variable covariances should receive a large weight. A simple way to achieve this is to work with autoscaled data and effectively threshold the correlation matrix. This scale-free approach is equivalent to applying the following entry-dependent thresholding to the elements of the covariance matrix[16]:

$$\delta_{ij} = \sqrt{s_{ii}s_{jj}}\delta. \tag{15}$$

Alternatively, Cai et al propose adaptive weighting of the covariance matrix using the standard error of each entry as weight.[61] Note that adaptive thresholding can be used in combination with any of the generalized thresholding rules discussed above (ie, by generalized thresholding of the sample correlation matrix).

### 5.3 | Thresholding with a positive definite constraint

Although the element-wise thresholding methods are simple, they do not guarantee that the estimated covariance matrix is positive definite. The larger the dimension, the less likely the estimator is to be positive definite.[58] This may be problematic if the matrix is used in further analysis such as in an LDA or QDA model.

When the threshold is sufficiently large, the estimated covariance matrix is positive definite with high probability.[16] An example is shown in Figure 7F. Note that the range of thresholds for which the covariance estimate is positive definite is wider for ST than AT (and, although not shown, also for the SCAD estimator) compared to HT. In practice, one could imagine to only investigate high thresholds for which the estimated matrix is positive definite. This approach, however, is not suitable for less sparse problems where a low threshold should be applied. Alternatively, one could use a ridge-type approach with a thresholded target or restrict the estimate to its eigenvectors with nonzero eigenvalue. However, these approaches destroy the sparsity pattern.[16]

To obtain a sparse positive definite matrix directly, Bien et al propose a penalized likelihood approach. The resulting optimization problem is, however, not convex.[31] The routines that have been proposed to solve the problem converge slowly and may not reach the global optimum.[31] An alternative approach is to add a positive definiteness constraint to the thresholding minimization problem defined in Equation 13.[58,62] The resulting problem is convex, and fast algorithms have been developed to solve it[58]:

$$\widehat{\mathbf{\Sigma}}(\delta) = \arg\min_{\mathbf{\Sigma}} \frac{1}{2}\|\mathbf{S}-\mathbf{\Sigma}\|_F^2 + p_\delta(\mathbf{\Sigma}), \text{subject to } \mathbf{\Sigma} \succcurlyeq \epsilon\mathbf{I}. \tag{16}$$

Here, $\mathbf{\Sigma} \succcurlyeq \epsilon\mathbf{I}$ means that $\mathbf{\Sigma} - \epsilon\mathbf{I}$ is semipositive definite for a small positive constant $\epsilon$ (ie, $\epsilon = 10^{-5}$). This guarantees that $\widehat{\mathbf{\Sigma}}$ is positive definite (its minimum eigenvalue $\geq \epsilon$). For simulated and real data, it has been shown that the constrained approaches[58,63,64] receive similar performance with respect to several loss functions as generalized thresholding, while being positive definite.[58,63,64]

### 5.4 | Estimators for ordered variables

The literature on sparse estimators can be loosely divided into 2 categories. The first class of methods is the thresholding approaches discussed above. They are invariant to permutations of the variables, ie, they do not assume a specific ordering of the variables. The other class of methods deals with the situations where the variables do have a natural ordering, as in longitudinal data, spatial data, and some types of spectroscopic data. Specifically taking this structure into account leads to improved estimators compared to using generalized thresholding.[12]

A simple approach to estimate a covariance matrix for such data is banding.[65] The method can be viewed as a HT rule where subdiagonals in the matrix that are too far away from the centre are set to zero.[51,66] The implicit assumption

made is that variables far apart in the ordering have small correlations (covariance), i.e. the covariance matrix is assumed to be subdiagonally sparse (note that this assumption is violated for some types of spectroscopy, e.g. NMR). Unfortunately, banded matrices are not guaranteed to be positive definite. Tapering is a smooth version of banding that guarantees positive definiteness.[65] Instead of a hard threshold, the diagonal entries gradually decay to zero (similar to soft-thresholding). An alternative approach to banding and tapering is to regularize the lower triangular matrix of the Cholesky decomposition of $\mathbf{S}$.[23,67]

### 5.5 | Low-dimensional + sparse estimators

The thresholding approaches that have been discussed above assume that the covariance matrix is sparse. Although this assumption is reasonable for many applications, it is not always appropriate. In such cases, it might be useful to assume that the covariance matrix is sparse conditional on some common source of variation that affects many variables in the data (and makes them correlated), ie, many pairs of variables are weakly correlated after taking out the common variation. As mentioned by Fan et al, an example can be found in biology where genes from the same pathway may be coregulated by a small amount of regulatory factors, which makes their expressions highly correlated.[28]

Fan et al propose the (POET) estimator for data with such a structure[28]:

$$\mathbf{\Sigma} = \sum_{i=1}^{k} \lambda_i \, \mathbf{p}_i^T \mathbf{p}_i + \mathbf{\Sigma}_r. \tag{17}$$

Note that it is assumed that the eigenvalues of the first $k$ eigenvectors are much larger than the remaining ones (ie, the first $k$ PCs describe the majority of the variation in the data). The residual covariance structure is assumed to be sparse and can be estimated using one of the thresholding approaches discussed above. As discussed in Sections 2.1 and 3, the first $k$ eigenvalues are usually estimated with large error. Therefore Fan et al recently proposed the shrinkage POET (S-POET) estimator where a nonlinear shrinkage function is applied to the highest $k$ eigenvalues.[68]

### 5.6 | Metabolite clustering via correlations

We consider the application of thresholding estimators of the correlation matrix for the analysis of NMR metabolomics data. This approach is, for example, used in statistical total correlation spectroscopy to elucidate both intermetabolite and intrametabolite correlations of the data.[69] Here, we investigate the effect of thresholding estimators for clustering of $^1$H-HRMAS-NMR (Proton - High Resolution Magic Angle Spinning - Nuclear Magnetic Resonance spectroscopy) data from a metabolomics experiment on *Caenorhabditis elegans*.[70] The data set contained 2 classes of samples, namely sod-1 (tm 776) mutants and N2 wild-type nematodes. Similar to the original publication, the data were reduced over the chemical shift range 0 to 9 ppm excluding residual water signal (4.6-5.0 ppm).[70] Subsequently, each spectrum was normalized using probabilistic quotient normalization and binned using statistical recoupling of variables.[71] The processed data set had 72 samples and 798 variables (bins).

For illustration, we analyze a subset of the 798 bins to clearly visualize the differences between methods in heat maps. A subset of the variables is studied to obtain a clearer illustration of differences between the thresholding methods. Note that similar differences between the methods as will be shown below were observed when the full data matrix was analyzed (results not shown). Similarly to the microarray example presented by Rothmann et al, we ranked each variable in the data by how much discriminative information it provided using a 2-sample $t$-test.[26] Subsequently, a subset of the top 40 bins (lowest $P$ value) and bottom 120 bins (highest $P$ value) was retained for further analysis. This way, the reduced data set contained informative and noninformative bins. Next, the correlation between the bins was visualized in a heat map. For this purpose, the correlation matrix was estimated and, subsequently, its rows and columns were ordered by average-link hierarchical clustering using the estimated correlation in the dissimilarity measure[9]:

$$d_{ij} = 1 - |r_{ij}|, \tag{18}$$

where $r_{ij}$ is the estimated Pearson's correlation coefficient between bins $i$ and $j$. The heat map obtained using the sample correlation matrix was compared to the map of thresholded correlation matrices. The optimal value for the tuning parameter $\delta$ was determined by cross-validation with a moderately sized training set in each split. Note that clustering of the variables was applied to clearly highlight the expected block structure of the correlation matrices. Clustering should not be applied in applications where the ordering between the bins is important such as statistical total correlation spectroscopy.[69,72]

Figure 7 shows the resulting heat maps of the sample and thresholded estimates. Note that absolute values rather than the correlations themselves were plotted because we were only interested in the strength of the pairwise association between 2 bins (and not its sign). It is clear that the 40 informative bins formed 2 strongly correlated blocks. Some block-structure could also be observed between the uninformative bins. These bins might correspond baseline signal (with very small offset differences between samples) or multiple bins corresponding to an unimportant peak in the spectrum. As shown in Figure 7, all thresholding approaches greatly increased the number of zeros in the estimate (the percentages in the figure indicate the number of nonzero matrix elements). As expected, ST resulted in the least sparse matrix (see Section 5.1). The SCAD (not shown), AT, and SEC ("Sparse Estimation of the Correlation matrix"; AT with positive definite constraint) estimates were in-between those of

ST and HT. HT estimated many more zeros than the other methods.

As mentioned in Section 5.1, HT tends to threshold too many elements, which often results in nearly diagonal estimates and a loss of important correlation information. Although the HT estimate in Figure 7B seems to reveal the block structure in the data moderately well, cross-validation provided some evidence of this effect: Only a 1% improvement (reduction of Frobenius loss) over the sample correlation matrix was observed for HT (for analysis of the full spectrum no improvement was observed), while the other (much) less sparse estimates showed up to 30% improvement. Often, it is desirable to obtain positive definite estimates (minimum eigenvalue >0).[16] Figure 7F plots the minimum eigenvalue of the thresholded correlation matrices as a function of the threshold. It is clearly seen that for each estimator there is a range of thresholds for which a positive definite estimate is obtained. Hard thresholding, however, yielded the narrowest range of thresholds to give positive definiteness corresponding to very sparse matrix estimates (in this case a diagonal matrix).

The examples above show why a combination of thresholding and shrinkage (ST, AT, and SCAD) might be preferred in practice compared to using a hard-thresholding rule.

## 6 | SPARSE PRECISION MATRIX ESTIMATION

For many real life applications, the quantity of interest is the inverse covariance or precision matrix. For example, this matrix is used in techniques such as LDA and QDA.[8] Additionally, whereas the covariance matrix keeps information related to pairwise correlation between variables, the inverse reflects partial correlation. More specifically, a zero in the precision matrix corresponds to a pair of variables that are partially uncorrelated. In case of data with a multivariate Gaussian distribution, this corresponds to variables that are independent conditional on the other (see Section 6.5 for more details).[15,16,22,27] Because of these reasons, a large body of literature has focused on sparse estimation of the precision matrix. Note that sparse covariance estimators are not particularly useful in this respect since their inverse does not have to be sparse.

### 6.1 | The graphical lasso

A natural, and arguably the most popular, approach for estimating sparse precision matrices is to penalize the (Gaussian) log likelihood[27]:

$$\widehat{\Sigma^{-1}} = \arg\min_{\Sigma^{-1}>0} -\log\left|\Sigma^{-1}\right| + tr\left(\mathbf{S}\Sigma^{-1}\right) + p_\delta\left(\Sigma^{-1}\right). \quad (19)$$

Here, $p_\delta(\Sigma^{-1})$ corresponds to a sparsity inducing penalty. The tuning parameter $\delta$ is typically optimized using cross-validation or a model selection approach.[73,74] Note that the solution of the unpenalized problem is the maximum likelihood precision matrix. Several penalties have been proposed in combination with Expression 19 amongst which the lasso, adaptive lasso, and SCAD are the most popular ones (see Figure 6 for their behaviour; the lasso penalty closely resembles ST).[12,16] Sometimes, the diagonal elements of $\Sigma^{-1}$ are not penalized.[75] This way, $\Sigma^{-1}$ is shrunk towards $diag(\Sigma^{-1})$ for large penalties, which can be thought of as a target matrix (see Section 4.1).

Several algorithms have been proposed to solve the problem Equation 19. To date, the most popular approach is the graphical lasso (glasso), where a solution to (Equation 19) is found by solving a series of coupled regression problems in an iterative fashion.[27,75–77] A special property of glasso is that the estimated precision matrix is always positive definite as long as the algorithm is initialized with a positive definite matrix such as a shrinkage estimator.[23] Hsieh et al propose a quadratic approximation to the objective in Equation 19 to reduce the computational load.[78,79] This makes Equation 19 applicable even when the data have millions of variables. The glasso was developed to solve (Equation 19) for the lasso penalty but can also be used to deal with the adaptive lasso (by a weighting of the data) or the SCAD (by using a linear approximation of the penalty that is iteratively solved by glasso).[16]

### 6.2 | Column-by-column approaches

Another approach to sparse precision matrix estimation is to estimate the matrix column-by-column using, for example, penalized regression. Compared to glasso, the column-by-column methods are computational, less complex, and more amendable to theoretical analysis. However, the resulting estimate of the precision matrix is not necessarily positive definite as is the case with glasso.

Consider the lasso linear regression of variable $j$ ($\mathbf{X}_j$) on the other variables in the data set ($\mathbf{X}_{\setminus j}$)[8,80]:

$$\widehat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\left\|\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\boldsymbol{\beta}\right\|_F^2 + \delta\|\boldsymbol{\beta}\|_1. \quad (20)$$

Note that the lasso (L1-norm) penalty $\delta\|\boldsymbol{\beta}\|_1$ forces some coefficients in $\widehat{\boldsymbol{\beta}}_j$ to be exactly zero. If the rows in $\boldsymbol{X}$ follow a multivariate normal distribution, it can be shown that the resulting regression coefficients $\beta_{ij}$ are given by[23,81]:

$$\beta_{jk} = -\frac{\sigma^{jk}}{\sigma^{jj}},\ j \neq k, \quad (21)$$

where $\sigma^{jk}$ indicates the $jk$th element of $\Sigma^{-1}$. In other words, the coefficients in $\widehat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\Sigma}_j^{-1}$ (after its diagonal element $\sigma^{jj}$ is removed) share the same zero coefficients. Therefore, the sparsity pattern of $\Sigma^{-1}$ can be easily recovered by solving Expression 21 for every column in $\Sigma^{-1}$.[16,23] A disadvantage of estimating each column in $\Sigma^{-1}$ separately is the lack of

symmetry of the approach, ie, $\widehat{\sigma}^{jk}$ is generally not equal to $\widehat{\sigma}^{kj}$. Therefore, typically, a postprocessing rule is applied such as replacing the estimates of $\widehat{\sigma}^{jk}$ and $\widehat{\sigma}^{kj}$ by their minimum or their average.[80] An alternative remedy is to take the natural symmetry of the problem into account by merging all the column-by-column regression subproblems into a single problem.[37,82–84]

The column-by-column approach has been adapted by many methods. Notable examples include the graphical Dantzig selector,[81] CLIME,[85] SCIO (sparse column inverse operator),[86] the scaled-Lasso method,[87] and TIGER (Tuning-Insensitive Graph Estimation and Regression).[88] These methods differ from each other mainly by how they solve the sparse regression subproblem. The graphical Dantzig selector and CLIME use the Dantzig selector,[89] SCIO uses the Lasso (similar to Expression 20), while this 1 study[87] uses the scaled-Lasso which has a few similarities to the TIGER. The TIGER solves the sparse regression problem using square root–Lasso regression.[90] Most of the column-by-column methods are justified by some theoretical choices on the tuning parameter $\delta$ that cannot be implemented in practice. For example, in Equation 20, the optimal parameter-value depends on the variance of the residual noise, which is typically unknown.[80] Therefore, in practice approaches such as cross-validation are used to optimize $\delta$. To save computational time, usually, the same tuning parameter is used for each regression subproblem.

A big advantage of the TIGER (and the closely related scaled lasso) compared to other column-by-column methods is that the choice of $\delta$ does not depend on the unknown noise variance.[88] Because of this, only limited effort is required to select the optimal tuning-value for each individual regression subproblem (the method is essentially tuning free). Liu and Wang show that TIGER outperforms glasso and often outperforms the other column-by-column approaches (CLIME) with respect to identification of the true zero and nonzero coefficients[88] in $\mathbf{\Sigma}^{-1}$. Recently, a tuning invariant extension of the CLIME estimator was proposed as well.[91] An advantage of this approach compared to TIGER is that it can also handle data from non-Gaussian, heavily tailed distributions.[16] An advantage of CLIME is that the estimate is positive definite with high probability.[85] Computationally efficient implementations of TIGER and CLIME are discussed in.[92–94]

### 6.3 | Estimators for ordered variables

Regularization of the precision matrix for ordered variables is often based on its Cholesky decomposition.[23] More specifically, the elements of the lower triangular matrix ($\mathbf{L}$) can be obtained by regressing each variable on its predecessors. Therefore, sparse estimates of $\mathbf{L}$ can be obtained by using penalized regression similar to the column-by-column methods above.[95,96] Note that although these estimators incorporate sparsity via matrix $\mathbf{L}$, the estimate of the precision matrix itself is generally not sparse (unless the sparsity pattern in $\mathbf{L}$ has a specific structure). Alternatively, matrix $\mathbf{L}$ can be banded in which case, the estimated precision matrix will also be sparse.[65,97]

### 6.4 | Conditional sparsity

Sparse precision matrix estimators can be combined with the POET framework.[15] Recall from section 5.5 that the POET estimator first extracts common variation from the data by PCA and assumes a sparse structure on the residual covariance matrix. Similarly, one of the sparse precision matrix estimators (eg, TIGER) can be applied to the residual data to obtain a sparse precision estimate conditional on the common variation that was extracted.[15]

### 6.5 | Sparse precision matrix estimation for GGM

Gaussian graphical modelling is a popular approach in, eg, genomics to visualize the (linear) dependencies between a set of variables.[22] A GGM is an undirected graph in which each edge represents conditional dependence of 2 variables (see Figure 8 for an example). As mentioned in Section 6, a zero in the precision matrix corresponds to a pair of variables that are partially uncorrelated. In case of data with a multivariate Gaussian distribution, this corresponds to variables that are conditionally independent. In other words, for such data, a zero in the precision matrix indicates the absence of an edge in the GGM.[22] The GGMs can be easily constructed form sparse estimates of the precision matrix.

For illustration, we analyze a targeted metabolomics data set in which 151 metabolites were quantified in blood serum samples from a large cohort of 1020 individuals, from Krumsiek et al.[98] These authors fitted a GGM to the data by estimation of all pairwise partial correlations in combination with a false discovery rate procedure to identify significant partial correlations. Note that this approach is not applicable when $n << p$. Here, we use the graphical lasso to estimate a sparse precision matrix and the corresponding GGM, which is also applicable in this case. Prior to fitting the glasso using the Quadratic Inverse Covariance algorithm,[79] the data was log-transformed and autoscaled. Inspection of the variables by Quantile-Quantile-plots of the preprocessed data showed that they were approximately normally distributed.

The result of applying the graphical lasso to the data is shown in Figure 8. It can be seen that the estimated precision matrix became sparser for increasing values of the tuning parameter $\delta$ revealing a block-like structure. Consequently, the corresponding GGMs displayed less dense networks with increasing values of $\delta$. Interestingly, a clear modular structure (the GGM indicates local clustering) with respect to the 7 metabolite classes that were measured was revealed this way. The acyl-carnitines and amino acids were clearly separated from the other classes (mainly phospholipids). The 4

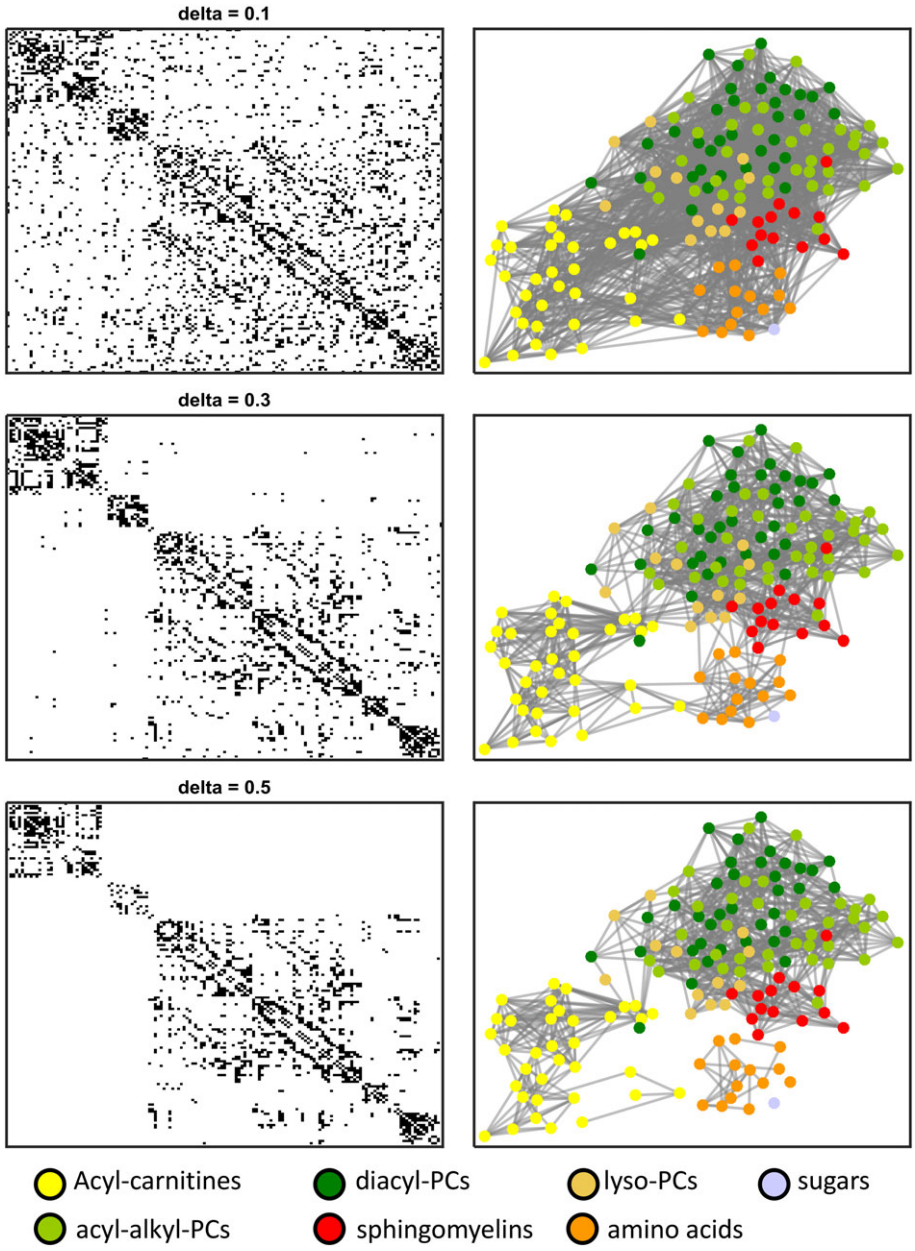


**FIGURE 8** Gaussian graphical models of a targeted metabolomics data set based on gLasso estimates of the precision matrix for different values of the tuning parameter $\delta$. The heat maps in the left-hand column indicate in black the nonzero elements in the precision matrix estimate. The right-hand column displays the corresponding GGM. PCs indicates phosphatidylcholines

groups of phospholipids were more strongly connected in the network, and no clustering could be observed even for much higher values of the sparsity tuning parameter. We observed that the modular structure of the inferred GGM was highly robust to changes in sample size (eg, similar results were obtained when a subset of 100 samples was analyzed).

The sparse estimators of the correlation matrix can also be used to construct a network. This would typically, however, result in denser and less interpretable networks. A GGM (and the associated sparse precision matrix) encodes only direct relationships between variables while a network based on the correlation matrix also displays indirect relationships (conditional versus marginal dependence).

## 7 | ALGORITHMS AND SOFTWARE AVAILABILITY

Most of the optimization problems discussed throughout the text are convex and can be solved using standard tools from convex optimization.[99] In many cases, however, more efficient (problem-specific) algorithms have been developed such as the graphical lasso (Section 6.1). We refer the reader to the relevant references in the text for more details. As shown above, most shrinkage (Section 3), ridge-type (Section 4) and thresholding (Section 5) optimization problems have an analytical solution and can therefore be directly implemented using the equations in this text.

**TABLE 2** A nonexhaustive list of Matlab and/or R implementations of eigenvalue shrinkage, ridge-type, and structured estimators for the covariance and precision matrix

| Category | Subcategory | Method | R package |
|---|---|---|---|
| Eigenvalue shrinkage | Linear shrinkage | LW-LIN | CorpCor, nlshrink |
| | Nonlinear shrinkage | CNR, CERNN, LW-NONLIN | CondReg, cernn, nlshrink |
| Ridge type | | LW-based<br>Van Wieringen-based<br>NOVELIST | CorpCor, rags2ridges, NOVELIST |
| Sparse covariance matrix | Thresholding | Generalized thresholding | FinCovRegularization, POET, PDSCE (posdef constraint) |
| | Low-dimensional + sparse | POET | POET |
| Sparse precision matrix | Graphical lasso | glasso, BIC&QUIC | glasso, QUIC, |
| | Column-by-column | CLIME, TIGER | Huge, flare |

Abbreviations: BIC indicates Bayesian information criterion; CERNN, covariance estimate regularized by nuclear norms; CNR, condition number regularized; LW-LIN, Ledoit–Wolf linear shrinker; LW-NONLIN, Ledoit–Wolf nonlinear shrinker; NOVELIST, novel integration of the sample and thresholded covariance estimators; POET, principal orthogonal complement thresholding; QUIC, quadratic approximation of inverse covariance matrices; TIGER, tuning-insensitive graph estimation and regression.

Software packages are available for a large number of the covariance and precision matrix estimators discussed above. Table 2 presents a nonexhaustive list of R implementations.

## 8 | CONCLUSION AND FUTURE RESEARCH

With the advent of high-throughput analytical techniques, we are able to capture a wealth of information in a single sample. Although the resulting high-dimensional data sets offer great possibilities such as data-driven research, they also pose great challenges to traditional multivariate statistical methods. For example, the sample estimates of the covariance matrix and the precision matrix, which are used in a multitude of techniques, are unreliable.

In this paper, an overview of modern estimators of the covariance and precision matrix was presented. By means of analysis of 3 metabolomics data sets, it was shown that these estimators hold great promise in chemometrics: They are easily combined with existing chemometric techniques such as PCA, MANOVA, and GGM approaches to obtain better results. These are not the only chemometric techniques that could benefit from regularized estimators of the covariance or precision matrix. Other examples include canonical correlation analysis, (generalized) partial least squares regression, ordinary least squares, mixture modelling, LDA, QDA, MSPC techniques and soft independent modelling of class analogy.

The methods covered in this paper can be roughly divided into 3 separate families. These methods rely on different assumptions regarding the structure of the covariance or precision matrix. The eigenvalue-shrinkage approaches aim to correct the overdispersion (bias) of the sample eigenvalues but do not affect the sample eigenvectors. These approaches are useful when no knowledge regarding the structure of the covariance or precision matrix is available. Ridge-type and structured estimators may be preferred when prior knowledge regarding the structure of the population covariance (or precision) matrix is available. Ridge-type estimators shrink the covariance or precision matrix towards a "simple" target and thereby indirectly incorporate knowledge regarding the structure of the population matrix in the estimator. Ridge-type estimators are appealing in practice because of their simplicity and ease of implementation. However, they typically do not result in sparse (interpretable) estimates. Structured estimators aim to directly impose a specific structure (often a sparse structure) on the covariance estimate by constraining the solution space. Sparse estimators offer a useful tool for exploring the variable dependence structure in the data. Currently, estimators are being developed that combine these regularization assumptions, eg, sparse model assumptions plus eigenvalue shrinkage.[37]

There are many remaining challenges in the estimation of covariance and precision matrices. Although a multitude of methods have been proposed, usually somewhat arbitrary criteria (eg, a specific loss function) are used to compare them. This makes it difficult to select 1 specific method over another in practice. Additionally, the application of regularized covariance and precision estimates in multivariate statistical models has been much less studied (not much is known about the functionals of regularized estimates).[68] Clearly, a more thorough evaluation of the different regularization approaches in the context of specific multivariate techniques and a specific data type (eg, a relatively highly collinear near infrared data structure or a more sparse liquid chromatography–mass spectrometry structure) is required. Finally, we would like to remark that direct regularization of the eigenvectors of the covariance matrix was not covered in this paper. This approach is useful, for example, in sparse PCA and sparse LDA.[22] Comparison of this approach to the other regularization approaches and possibly combining them offers another interesting avenue for further research.

## REFERENCES

1. Engel J, Blanchet L, Buydens LM, Downey G. Confirmation of brand identity of a trappist beer by mid-infrared spectroscopy coupled with multivariate data analysis. *Talanta*. 2012;99:426–432.

2. Cooper JB. Chemometric analysis of Raman spectroscopic data for process control applications. *Chemom Intel Lab Syst*. 1999;46(2):231–247.
3. Engel J, Blanchet L, Engelke UF, Wevers RA, Buydens LM. Towards the disease biomarker in an individual patient using statistical health monitoring. *PLoS One*. 2014;9(4):e92452.
4. Chen R, Mias GI, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*. 2012;148(6):1293–1307.
5. Buydens L. Towards tsunami-resistant chemometrics. *Anal Sci*. 2013;401.
6. Donoho D. High-dimensional data analysis: the curse and blessings of dimensionality. In: *American Math Society on Math Challenges of the 21st Century.* Los Angeles; 2000.
7. Fan J. Han F., and Liu H., Challenges of Big Data Analysis. *National Science Review*, 2014.
8. Hastie TTR, Friedman J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.; 2001.
9. Vandeginste BG, Massart DL, de Jong S, et al. *Handbook of Chemometrics and Qualimetrics*. Elsevier; 1998.
10. Mardia KV, Kent JT, Bibby JM. In: Kent JT, Bibby JM, eds. *Multivariate Analysis. Probability and mathematical statistics*. London; New York: Academic Press; 1979.
11. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivariate Anal* 2004;88(2):365–411.
12. Pourahmadi M. In: Balding DJC, Fitzmaurice NAC, Goldstein GM, Johnstone H, Molenberghs IM, Scott G, eds. D.W.*High-Dimensional Covariance Estimation. Wiley Series in Probability and Statistics*. Hoboken, New Jersey, US: John Wiley & Sons; 2013.
13. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* 2001;295–327.
14. Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat Sinica*. 2007;1617–1642.
15. Fan J, Liu H. Statistical analysis of big data on pharmacogenomics. *Adv Drug Deliv Rev*. 2013;65(7):987–1000.
16. Fan J, Liao Y, Liu H. An overview of the estimation of large covariance and precision matrices. *Econom J*. 2016;19(1):C1–C32.
17. Bai Z, Silverstein JW. *Spectral Analysis of Large Dimensional Random Matrices*. 20 Springer; 2010.
18. Schäfer JSK. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*. 2005;4(1).
19. Paul D, Aue A. Random matrix theory in statistics: a review. *J Stat Plan Infer* 2014;150:1–29.
20. Bai Z, Yin Y. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann Probab*. 1993;1275–1294.
21. Ledoit O, Wolf M. Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. *J Multivariate Anal* 2015;139:360–384.
22. Hastie T, Tibshirani R, Wainwright M. *Statistical Learning With Sparsity: The Lasso and Generalizations*. CRC Press; 2015.
23. Pourahmadi M. Covariance estimation: the GLM and regularization perspectives. *Stat Sci*. 2011;26(3):369–387.
24. Ledoit O, Wolf M. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann Stat*. 2012;40(2):1024–1060.
25. van Wieringen W.N. and Peeters C.F., Ridge Estimation of Inverse Covariance Matrices From High-dimensional Data. *Comput. Stat. Data An.*, 2016.
26. Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. *J Am Stat Assoc*. 2009;104(485):177–186.
27. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432–441.
28. Fan J, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. *J Roy Stat Soc B Met* 2013;75(4):603–680.
29. Anderson TW, Olkin I. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl*. 1985;70:147–171.
30. Warton DI. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J Am Stat Assoc*. 2008;103(481):340–349.
31. Bien J, Tibshirani RJ. Sparse estimation of a covariance matrix. *Biometrika*. 2011;98(4):807–820.
32. Stein C. Estimation of a covariance matrix. *Rietz Lecture, 39th Annual Meeting IMS, Atlanta, GA*, 1975.
33. Wang C, Pan G, Tong T, Zhu L. Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Stat Sin*. 2015;25(3):993–1008.
34. Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J Empir Financ*. 2003;10 (5):603–621.
35. Theiler J. *The Incredible Shrinking Covariance Estimator.* Security, and Sensing: *SPIE Defence*; 2012.
36. Won JH, Lim J, Kim SJ, Rajaratnam B. Condition-number-regularized covariance estimation. *J Roy Stat Soc B Met*. 2013;75(3):427–450.
37. Ali A, Khare K, Oh S-Y, Rajaratnam B, Generalized pseudolikelihood methods for inverse covariance estimation. *arXiv preprint arXiv:1606.00033*, 2016.
38. Chi EC, Lange K. Stable estimation of a covariance matrix guided by nuclear norm penalties. *Comp Stat Data An*. 2014;80:117–128.
39. Lam C. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann Stat*. 2016;44(3):928–953.
40. Ledoit, O. and M. Wolf, Numerical implementation of the QuEST function. *arXiv preprint arXiv:1601.05870*, 2016.
41. Donoho DL, Gavish M, Johnstone IM, Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*, 2013.
42. Ledoit O, Wolf M. Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. *J Multivariate Anal.* 2015;139:360–384.
43. Dunn WB, Lin W, Broadhurst D, et al. Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics*. 2015;11 (1):9–26.
44. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*. 2006;78(3):779–787.
45. Parsons HM, Ludwig C, Günther UL, Viant MR. Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*. 2007;8 (1):234.
46. Ståhle L, Wold S. Multivariate analysis of variance (MANOVA). *Chemom Intel Lab Syst*. 1990;9(2):127–141.
47. Engel J, Blanchet L, Bloemen B, et al. Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Anal Chim Acta*. 2015;899:1–12.
48. Smilde AK, Jansen JJ, Hoefsloot HC, et al. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*. 2005;21(13):3043–3048.
49. Srivastava MS. Multivariate theory for analyzing high dimensional data. *J Japan Statist Soc*. 2007;37(1):53–86.
50. Huang N, Fryzlewicz P. *NOVELIST Estimator of Large Correlation and Covariance Matrices and Their Inverses*. London School of Economics and Political Science: Technical report, Department of Statistics; 2015.
51. Chen X, Wang ZJ, McKeown MJ. Shrinkage-to-tapering estimation of large covariance matrices. *IEEE T Signal Proces.* 2012;60(11):5640–5656.
52. Fisher TJ, Sun X. Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comput Stat Data An.* 2011;55(5):1909–1918.
53. Chen Y, Wiesel A, Hero AO. Shrinkage estimation of high dimensional covariance matrices. in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009. IEEE.
54. Touloumis A. Nonparametric stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Comput Stat Data An*. 2015;83:251–261.
55. Lancewicki T, Aladjem M. Multi-target shrinkage estimation for covariance matrices. *IEEE T Signal Proces*. 2014;62(24):6380–6390.

56. Bartz D, Höhne J, Müller K-R, Multi-target shrinkage. *arXiv preprint arXiv:1412.2041*, 2014.
57. Bickel PJ, Levina E. Covariance regularization by thresholding. *Ann Stat.* 2008;2577–2604.
58. Xue L, Ma S, Zou H. Positive-definite $\ell$1-penalized estimation of large covariance matrices. *J Am Stat Assoc*. 2012;107(500):1480–1491.
59. Fan J, Feng Y, Wu Y. Network exploration via the adaptive LASSO and SCAD penalties. *Ann Stat.* 2009;3(2):521.
60. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348–1360.
61. Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. *J Am Stat Assoc*. 2011;106(496):1566–1577.
62. Rothman AJ. Positive definite estimators of large covariance matrices. *Biometrika*. 2012;99(3):733–740.
63. Liu H, Wang L, Zhao T. Sparse covariance matrix estimation with eigenvalue constraints. *J Comput Graph Stat.* 2014;23(2):439–459.
64. Wen F., Yang Y., Liu P., and Qiu R.C., Positive definite estimation of large covariance matrix using generalized nonconvex penalties. *arXiv preprint arXiv:1604.04348*, 2016.
65. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *Ann Stat.* 2008;199–227.
66. Bickel P, Lindner M. Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics. *Theor Probab Appl*. 2012;56(1):1–20.
67. Rothman AJ, Levina E, Zhu J, A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, 2010: p. asq022.
68. Fan J , Wang W, Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *arXiv preprint arXiv:1502.04733*, 2015.
69. Holmes E, Cloarec O, Nicholson J. Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: application to HgCl2 toxicity. *J Proteome Res*. 2006;5(6):1313–1320.
70. Blaise BJ, Giacomotto J, Triba MN, et al. Metabolic profiling strategy of *Caenorhabditis elegans* by whole-organism nuclear magnetic resonance. *J Proteome Res*. 2009;8(5):2542–2550.
71. Blaise BJ, Shintu L, Elena B, et al. Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. *Anal Chem*. 2009;81(15):6242–6251.
72. Cloarec O, Dumas M, Craig A, et al. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Anal Chem*. 2005;77(5):1282–1289.
73. Vujačić I, Abbruzzo A, Wit E. A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *J Stat Comput Sim*. 2015;85(18):3628–3640.
74. Lian H. Shrinkage tuning parameter selection in precision matrices estimation. *J Stat Plan Infer*. 2011;141(8):2839–2848.
75. Mazumder R, Hastie T. The graphical lasso: new insights and alternatives. *Electron J Stat*. 2012;6:2125.
76. Witten DM, Friedman JH, Simon N. New insights and faster computations for the graphical lasso. *J Comput Graph Stat*. 2011;20(4):892–900.
77. Tan KM, Witten D, Shojaie A. The cluster graphical lasso for improved estimation of Gaussian graphical models. *Comp Stat Data An*. 2015;85:23–36.
78. Hsieh C-J, Dhillon IS, Ravikumar PK, Sustik MA. Sparse inverse covariance matrix estimation using quadratic approximation. *Adv Neur In*. 2011.
79. Hsieh C.-J., Sustik M.A., Dhillon I.S., Ravikumar P.K., and Poldrack R.. BIG & QUIC: Sparse inverse covariance estimation for a million variables. *Adv. Neur. In*. 2013.
80. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat.* 2006;1436–1462.
81. Yuan M. High dimensional inverse covariance matrix estimation via linear programming. *J Mach Learn Res.* 2010;11(Aug):2261–2286.
82. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *J Am Stat Assoc*. 2012;
83. Rocha GV, Zhao P, Yu B, A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *arXiv preprint arXiv:0807.3734*, 2008.
84. Khare K, Oh SY, Rajaratnam B. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J Roy Stat Soc B Met.* 2015;77(4):803–825.
85. Cai T, Liu W, Luo X. A constrained $\ell$ 1 minimization approach to sparse precision matrix estimation. *J Am Stat Assoc*. 2011;106(494):594–607.
86. Liu W, Luo X. Fast and adaptive sparse precision matrix estimation in high dimensions. *J Multivariate Anal*. 2015;135:153–162.
87. Sun T, Zhang C-H. Sparse matrix inversion with scaled lasso. *J Mach Learn Res*. 2013;14(1):3385–3418.
88. Liu H, Wang L, Tiger: a tuning-insensitive approach for optimally estimating gaussian graphical models. *arXiv preprint arXiv:1209.2437*, 2012.
89. Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann Stat.* 2007;2313–2351.
90. Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*. 2012;80(6):2369–2429.
91. Zhao T, Liu H. Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE T Inform Theory.* 2014;60(12):7874.
92. Li, X., J. Haupt, R. Arora, et al., A first order free lunch for SQRT-lasso. *arXiv preprint arXiv:1605.07950*, 2016.
93. Wang H, Banerjee A, Hsieh C-J, Ravikumar PK, Dhillon IS. Large scale distributed sparse precision estimation. *Adv Neural In*. 2013;
94. Cai TT, Liu W, Zhou HH, Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *arXiv preprint arXiv:1212.2882*, 2012.
95. Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*. 2006;93(1):85–98.
96. Levina E, Rothman A, Zhu J. Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann Stat*. 2008;2(1):245–263.
97. Pourahmadi M. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters. *Biometrika*. 2007;94(4):1006–1013.
98. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*. 2011;5(1):1.
99. Boyd SVL. *Convex Optimization*. New York, NY, USA: Cambridge University Press; 2004.


**How to cite this article:** Engel J, Buydens L, Blanchet L. An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics. *Journal of Chemometrics*. 2017;31:e2880. https://doi.org/10.1002/cem.2880