sciendo

DOI: 10.2478/bile-2019-0010

Biometrical Letters
Vol. 56 (2019), No. 2, 165-181

Covariance regularization for metabolomic data on the drought resistance of barley

Adam Mieldzioc[1], Monika Mokrzycka[2], Aneta Sawikowska[1,2]

[1]Department of Mathematical and Statistical Methods,
Poznań University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland
e-mail: adam.mieldzioc@mail.up.poznan.pl, aneta.sawikowska@mail.up.poznan.pl
[2]Institute of Plant Genetics, Polish Academy of Sciences, Strzeszyńska 34,
60-479 Poznań, Poland, e-mail: mmok@igr.poznan.pl

Summary

Modern chromatography largely uses the technique of gas chromatography coupled with mass spectrometry (GC–MS). For a set of data concerning the drought resistance of barley, the problem of the characterization of a covariance structure is investigated with the use of two methods. The first is based on the Frobenius norm and the second on the entropy loss function. For the four considered covariance structures – compound symmetry, three-diagonal and penta-diagonal Toeplitz and autoregression of order one – the Frobenius norm indicates the compound symmetry matrix and autoregression of order one as the most relevant, whilst the entropy loss function gives a slight indication in favor of the compound symmetry structure.

Key words: covariance structure, compound symmetry matrix, autoregression matrix, Toeplitz matrix, estimation, regularization, entropy loss function, Frobenius norm

1. Introduction

Metabolomics is one of the most recent omics technologies. It has been applied in many fields of science, including drug discovery, food science, nutrition, and systems biology. The modern approach to systems biology requires large-scale experiments studied for a large number of genotypes, under various environmental conditions, with the number of biological replications sufficient for proper estimation of the natural variation. Chromatography is a widely used laboratory technique for the separation of chemical compounds (e.g. Piasecka et al. (2016), Piasecka et al. (2017), Swarcewicz et al. (2017), Ożarowski et al. (2017), Sawikowska et al. (2018)). Currently used

chromatographs take advantage of new high-throughput and sensitive instruments and protocols that generate a huge amount of data, in which the number of variables exceeds the sample size. This makes the analysis of the data set challenging, because such an experiment has too many parameters to estimate. One way to solve this problem is to select an appropriate covariance structure, which is the main purpose of this paper. The covariance structure is determined using prior knowledge of the researchers conducting the experiment. If such knowledge is not available, then the structure is selected from among a set of covariance structures. Such covariance structure selection is called regularization; cf. Cui et al. (2016) and Lin et al. (2014). Regularization consists of two steps. The first step is to determine a sample covariance matrix or maximum likelihood estimator of a true covariance matrix, which in the high-dimensional case is singular or ill-conditioned. The second step is to find a structured estimator of an unknown covariance matrix among the class of available structures, which depends on a smaller number of parameters, and which minimizes the discrepancy between unstructured and structured covariance matrix estimators. In Section 2.1, the following commonly used covariance structures, which we use in regularization, are presented: compound symmetry (CS), three-diagonal Toeplitz (T_1), penta-diagonal Toeplitz (T_2), and autoregression of order one (AR(1)). The discrepancy functions considered are the Frobenius norm and the entropy loss function presented in section 2.2. The data used in this paper were obtained in a study of metabolomic changes in barley (*Hordeum vulgare*) leaves under drought stress.

In this study, the metabolomic data come from an investigation of the effects of water shortage on the levels of primary metabolites in varieties of barley, measured repeatedly during the drought period (performed as a pilot study for a larger systems biology project; see Swarcewicz et al. (2017)). For 9 varieties of barley in drought treatment and control conditions, data were obtained at 4 plant growth stages, in 4 biological replications and 2 technical replications. The total number of samples was 422.

The following nine spring barley genotypes were used: the European varieties Georgia, Maresi, Lubuski, Sebastian and Stratus; Morex, bred in the USA; and Cam/B1/CI 08887//Cl 0576, Harmal, and Maris Dingo/Deir Alla 106, being lines bred in Syria. Barley plants were cultivated under partially controlled greenhouse conditions; see Chmielewska et al. (2016) for more details. The primary metabolites were recognized by gas chromatography coupled with mass spectrometry (GC–MS), which is the most widely applied

technique playing an important role in the identification and quantification of chemical compounds in analyzed samples. Such metabolite profiling was described by Chmielewska et al. (2016). It was performed using a 6890 N gas chromatograph (Agilent, USA) and a GCT Premier mass spectrometer (Waters, USA) (see https://github.com/metabolomicdata/Figure_GC-MS.git). The instrument is composed of two major parts: a gas chromatograph and a mass spectrometer. The gas chromatograph uses a capillary column. Each sample with barley leaves first turns into gas. Then the column provides separation of the compounds. The compounds are retained by the column and elute at different retention times. The mass spectrometer breaks each compound into ionized fragments and detects these fragments based on mass-to-charge ratio. After the last step the metabolites can be easily identified by chemical names.

The set of GC–MS data has a three-dimensional nature, with intensity measurements (peak height) as a function of retention time (elution time of the ion) [min] and mass-to-charge ratio [m/z]; cf. Khakimov et al. (2016). The raw data in this paper consist of 51 135 traits for 422 samples. Trait values are intensities of absorbance for the list of retention times and mass-to-charge ratios where nonzero values are detected. After averaging the data over technical replications, 211 biological samples are considered. In the GC–MS data the total ion current (TIC) chromatograms may be analyzed. Each TIC represents the summed intensity across the entire range of masses detected at every retention time point. After summation over mass-to-charge ratio 781 traits are obtained, which correspond to the retention time. After this step each biological sample is represented by one TIC chromatogram. Observations are transformed by logarithm with base 1.2 to ensure normality of the data.

2. Statistical background

In this section we define the covariance structures considered in the paper, and we present two methods of regularization based on minimization of the Frobenius norm and the entropy loss function. The considered covariance structures are commonly used in practice, especially in statistical analysis of time series (applications in nature, medicine, economics, signal and image processing, etc.). Interesting examples of the use of these structures include problems related to climate research, image restoration, stock price forecasting, precipitation forecasting, analysis of pre-earthquake ionospheric anomalies, etc.

2.1. Covariance structures

Let us assume that the variances of observations are homogeneous and all observations are equally correlated, which means that the correlation coefficient does not depend on the distance between characteristics (in our example, related to retention time). Then the covariance matrix may have a CS structure of the form

$$\mathbf{\Psi}_{CS} = \sigma^2 \begin{pmatrix} 1 & \varrho & \varrho & \dots & \varrho \\ \varrho & 1 & \varrho & \dots & \varrho \\ \varrho & \varrho & 1 & \dots & \varrho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varrho & \varrho & \varrho & \dots & 1 \end{pmatrix} = \sigma^2 \left((1-\varrho)\mathbf{I}_m + \varrho \mathbf{1}_m \mathbf{1}_m^\top\right),$$

where $\mathbf{I}_m$ is the identity matrix of order m, and $\mathbf{1}_m$ is an $m-$dimensional vector of ones. To ensure the positive definiteness of the matrix $\mathbf{\Psi}_{CS}$, we assume $\sigma^2 > 0$ and $\varrho \in \left(-\frac{1}{m-1}; 1\right)$, cf. Lin et al. (2014). This matrix is also called the equicorrelation matrix, and can be used to design split-plot experiments.

Let us assume that a covariance structure has homogeneous variances and heterogeneous correlations between elements. Additionally, the correlations between elements from the i-th and j-th diagonals are homogeneous and depend only on the lag between them ($i, j \in \{0, \dots, p\}$). Then the covariance matrix may have the banded Toeplitz structure ($\mathbf{\Psi}_{T_p}$, $p \leqslant m-1$) of the form

$$\mathbf{\Psi}_{T_p} = \sigma^2 \begin{pmatrix} 1 & \varrho_1 & \dots & \varrho_p & 0 & \dots & 0 \\ \varrho_1 & 1 & \varrho_1 & \dots & \varrho_p & \ddots & \vdots \\ \vdots & \varrho_1 & 1 & \varrho_1 & & \ddots & 0 \\ \varrho_p & & \ddots & \ddots & \ddots & & \varrho_p \\ 0 & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \varrho_1 \\ 0 & \dots & 0 & \varrho_p & \dots & \varrho_1 & 1 \end{pmatrix} = \sigma^2 (\mathbf{I}_m + \sum_{i=1}^{p} \varrho_i \mathbf{H}_i),$$

where $\mathbf{H}_i$ is a symmetric matrix with i-th superdiagonal and subdiagonal elements equal to 1 and all other elements equal to 0. In this paper, we

consider the Toeplitz covariance matrix with $p = 1$ and $p = 2$. The matrix $\mathbf{\Psi}_{T_1}$ is positive definite (p.d.) when

$$\sigma^2 > 0 \quad \text{and} \quad \varrho_1 \in \left(-\frac{1}{2\cos(\pi/(m+1))}; \frac{1}{2\cos(\pi/(m+1))} \right).$$

The conditions under which the matrix $\mathbf{\Psi}_{T_p}$ $(p > 1)$ is p.d. can be expressed using principal minors. It is usually numerically determined; cf. Filipiak et al. (2018d). It worth noting that the matrix $\mathbf{\Psi}_{T_1}$ is also known as the first-order moving average covariance structure.

Finally, let us assume that variances of observations are homogeneous, and the correlations decline exponentially with the lag, which means that the magnitude of the correlation between two observations depends on the distance (retention time in the GC-MS data) between them. Then the covariance matrix may have an AR(1) structure of the form

$$\mathbf{\Psi}_{AR} = \sigma^2 \begin{pmatrix} 1 & \varrho & \varrho^2 & \dots & \varrho^{m-1} \\ \varrho & 1 & \varrho & \dots & \varrho^{m-2} \\ \varrho^2 & \varrho & 1 & \dots & \varrho^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varrho^{m-1} & \varrho^{m-2} & \varrho^{m-3} & \dots & 1 \end{pmatrix} = \sigma^2 \sum_{i=0}^{m-1} \varrho^i \mathbf{H}_i$$

with $\mathbf{H}_0 = \mathbf{I}$. The matrix $\mathbf{\Psi}_{AR}$ is positive definite for $\sigma^2 > 0$ and ϱ belonging to the interval $(-1; 1)$, cf. Cui et al. (2016). The AR(1) structure is a special case of Toeplitz matrices with $p = m - 1$.

2.2. Covariance regularization methods

In the regularization problem we use two discrepancy functions, the Frobenius norm and entropy loss function, for the four structures under consideration. To obtain the closest positive definite matrices in the sense of these functions, we should find the argument for which the value of a given function is the smallest (minimum of function), e.g. by determining the appropriate derivatives. Since for $\Psi_{CS}, \mathbf{\Psi}_{T_1}, \mathbf{\Psi}_{T_2}$ finding the minimum for these structures is a convex problem, we can achieve a global minimum. In the case of Ψ_{AR} the problem is not convex; therefore only a local minimum can be attained (cf. Cui et al., 2016, p. 128 and Lin et al., 2014, p. 317). The solution of this problem is considered in Cui et al. (2016) for the Frobenius norm and in Lin et al. (2014) for the entropy loss function. In the next part

we introduce notation and present solutions for the case of each considered structure with the given discrepancy functions.

Let $\mathcal{A} = \{\mathbf{\Psi}_{CS}, \mathbf{\Psi}_{T_1}, \mathbf{\Psi}_{T_2}, \mathbf{\Psi}_{AR}\}$. Regularization consists in choosing the structure $\mathbf{\Psi} \in \mathcal{A}$ which minimizes the discrepancy between a given covariance matrix $\mathbf{\Omega}$ and the structure $\mathbf{\Psi}$; that is

$$\xi = \min_{\mathbf{\Psi} \in \mathcal{A}} f(\mathbf{\Omega}, \mathbf{\Psi})$$

with f being the discrepancy function.

Since the true $\mathbf{\Omega}$ is unknown, to find the best approximation of $\mathbf{\Omega}$ by $\mathbf{\Psi} \in \mathcal{A}$ we use the maximum likelihood estimator (MLE) of $\mathbf{\Omega}$; that is, $\mathbf{S}$ defined as

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{Q}_{1_n}\mathbf{X}^\top \tag{1}$$

with $\mathbf{X}$ being an observation matrix and $\mathbf{Q}_{1_n} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

Let $||\mathbf{A}||_F = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^\top)}$ denote the Frobenius norm of matrix $\mathbf{A}$. In this paper we use the Frobenius norm

$$f_F(\mathbf{S}, \mathbf{\Psi}) = ||\mathbf{S} - \mathbf{\Psi}||_F$$

and the entropy loss function

$$f_E(\mathbf{S}, \mathbf{\Psi}) = \text{tr}\left(\mathbf{S}^{-1}\mathbf{\Psi}\right) - \ln\left|\mathbf{S}^{-1}\mathbf{\Psi}\right| - m$$

(cf. e.g. James and Stein, 1961; Dey and Srinivasan, 1985; Lin et al., 2014; Filipiak et al., 2018b; Filipiak et al., 2018c) as the discrepancy functions. Cui et al. (2016) considered the square of the Frobenius norm, whilst in this paper the plain Frobenius norm is considered. Observe that since the Frobenius norm is a convex function, its square is also convex, and therefore the minimum is the same.

The minimum of the discrepancy function will be denoted respectively by

$$\xi_F = \min_{\mathbf{\Psi} \in \mathcal{A}} f_F(\mathbf{S}, \mathbf{\Psi}) = f_F(\mathbf{S}, \widehat{\mathbf{\Psi}}) \tag{2}$$

and

$$\xi_E = \min_{\mathbf{\Psi} \in \mathcal{A}} f_E(\mathbf{S}, \mathbf{\Psi}) = f_E(\mathbf{S}, \widetilde{\mathbf{\Psi}}). \tag{3}$$

Both the Frobenius norm and the entropy loss function are convex, and this property implies that there is one minimum. It is a global minimum in the case of CS and T_p, and a local minimum in the case of AR(1) (cf. Cui et al., 2016 and Lin et al., 2014).

The formulae for the estimator of a structured covariance matrix with the Frobenius norm as the discrepancy function are described by Cui et al. (2016) for CS, T_1 and AR(1) and by Filipiak et al. (2018d) for T_p. These formulae are as follows.

- CS structure

$$\begin{cases} \varrho &= \dfrac{\delta}{(m-1)\operatorname{tr}(\mathbf{S})} \\ \sigma^2 &= \dfrac{\operatorname{tr}(\mathbf{S}) + \varrho\delta}{m + m(m-1)\varrho^2} \end{cases}$$

  with $\delta = \operatorname{tr}[\mathbf{S}(\mathbf{1}_m\mathbf{1}_m^\top - \mathbf{I}_m)]$,

- Toeplitz structure

  - T_p for $p = 1$

$$\begin{cases} \sigma^2 &= \dfrac{\operatorname{tr}(\mathbf{S})}{m} \\ \varrho_1 &= \dfrac{m\operatorname{tr}(\mathbf{S}\mathbf{H}_1)}{2(m-1)\operatorname{tr}(\mathbf{S})} \end{cases} . \tag{4}$$

  Filipiak et al. (2018d) observed that the matrix obtained from (4) given by Cui et al. (2016) may be indefinite. Thus, they proposed an algorithm for determination of the minimum of the Frobenius norm (cf. Filipiak et al., 2018d, p. 77). It can be shown that for the σ^2 and ϱ_1 given in (4), the estimator of T_1 can be given as

$$t\sigma^2\Big(\mathbf{I}_m + \frac{1}{2\cos\frac{\pi}{m+1}}\mathbf{H}_1\Big) \quad \text{if} \quad \varrho_1 > 0$$

  and

$$t\sigma^2\Big(\mathbf{I}_m - \frac{1}{2\cos\frac{\pi}{m+1}}\mathbf{H}_1\Big) \quad \text{if} \quad \varrho_1 < 0$$

with

$$t = \Big(m + \frac{\varrho_1(m-1)}{\cos\frac{\pi}{m+1}}\Big) \Big/ \Big(m + \frac{m-1}{2(\cos\frac{\pi}{m+1})^2}\Big).$$

- T_p for $p > 1$

  In this case the formulae for the estimator of the T_p structure cannot be given in explicit form. To determine the estimator, the algorithm proposed by Filipiak et al. (2018d), p. 78, can be used.

- AR(1) structure

  To determine the estimator of the AR(1) structure the following system of equations should be solved:

$$\begin{cases} -\sum\limits_{i=1}^{m-1} i\varrho^{i-1}\operatorname{tr}(\mathbf{S}\mathbf{H}_i) + \dfrac{2\sum\limits_{i=0}^{m-1}\varrho^i\operatorname{tr}(\mathbf{S}\mathbf{H}_i)\sum\limits_{i=1}^{m-1}(m-i)i\varrho^{2i-1}}{m+2\sum\limits_{i=1}^{m-1}(m-i)\varrho^{2i}} = 0 \\ \sigma^2 = \dfrac{\sum\limits_{i=0}^{m-1}\varrho^i\operatorname{tr}(\mathbf{S}\mathbf{H}_i)}{m+2\sum\limits_{i=1}^{m-1}(m-i)\varrho^{2i}} \end{cases}$$

  with $\mathbf{H}_0 = \mathbf{I}$. The above system of equations provides the local minimum of the discrepancy function (cf. Cui et al., 2016).

For CS and AR(1), the formulae described by Filipiak et al. (2017) relating to the separable structure $\mathbf{\Psi} \otimes \mathbf{\Sigma}$ ($\mathbf{\Psi} : p \times p$ and $\mathbf{\Sigma} : q \times q$) with $q = 1$ can also be used.

The formulae for the estimator of structured covariance matrix with the entropy loss function as the discrepancy function for CS, T_1 and AR(1) are as follows (cf. Lin et al., 2014):

- CS structure

$$\begin{cases} \varrho &= -\delta/\big((m-1)\operatorname{tr}(\mathbf{S}^{-1}) + (m-2)\delta\big) \\ \sigma^2 &= m/\big(\operatorname{tr}(\mathbf{S}^{-1}) + \varrho\delta\big) \end{cases}$$

  with $\delta = \operatorname{tr}(\mathbf{S}^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \mathbf{I}_m))$,

- Toeplitz structure
  - T_p for $p = 1$

$$\begin{cases} \sigma^2 = \sum\limits_{j=1}^{m} \dfrac{2s_j}{1+2\varrho_1 s_j} / \operatorname{tr}(\mathbf{S}^{-1}\mathbf{H}_1) \\ \sum\limits_{j=1}^{m} \dfrac{2s_j}{1+2\varrho_1 s_j} - \dfrac{m \operatorname{tr}(\mathbf{S}^{-1}\mathbf{H}_1)}{\operatorname{tr}(\mathbf{S}^{-1}) + \varrho_1 \operatorname{tr}(\mathbf{S}^{-1}\mathbf{H}_1)} = 0 \end{cases}$$

  with $s_j = \cos(\pi j/(m+1))$,

  - T_p for $p > 1$

  Similarly as in the previous case, for $p > 1$ the estimator cannot be given in explicit form. To determine the estimators the algorithm proposed by Lin et al. (2014), p. 322 can be used.

- AR(1) structure

  To determine the estimator of the AR(1) structure the following system of equations should be solved:

$$\begin{cases} \dfrac{m \sum\limits_{i=1}^{m-1} i\varrho^{i-1} \operatorname{tr}(\mathbf{S}^{-1}\mathbf{H}_i)}{\sum\limits_{i=0}^{m-1} \varrho^i \operatorname{tr}(\mathbf{S}^{-1}\mathbf{H}_i)} + \dfrac{2(m-1)\varrho}{1-\varrho^2} = 0 \\ \sigma^2 = m / \sum\limits_{i=0}^{m-1} \varrho^i \operatorname{tr}(\mathbf{S}^{-1}\mathbf{H}_i) \end{cases},$$

  with $\mathbf{H}_0 = \mathbf{I}$. The above system of equations provides the local minimum of the discrepancy function (cf. Lin et al., 2014).

For CS and AR(1) the formulae described by Filipiak et al. (2018b) relating to the separable structure $\mathbf{\Psi} \otimes \mathbf{\Sigma}$ ($\mathbf{\Psi} : p \times p$ and $\mathbf{\Sigma} : q \times q$) with $q = 1$ can also be used.

3. Results

To overcome the problem of the high-dimensionality of the data, among 781 traits three subsets were selected: the first with traits from 1 to 200, the second with traits from 201 to 400 and the third with traits from 401 to

600. Respective measurements for each subset are collected in a 200×211 matrix $\mathbf{X}_i$ ($\mathbf{X}_i \sim N_{m,n}(\boldsymbol{\mu}_\mathbf{i} \otimes \mathbf{1}_n^\top, \boldsymbol{\Sigma}_i, \mathbf{I}_n)$), $i = 1, 2, 3$. Each data matrix consists of 211 columns ordered according to the sample number given by the biologists (such that one column concerns one variety under drought treatment or in control conditions at one plant growth stage and in one biological replication) and 200 rows concerning traits. Each trait represents the total ion current chromatogram calculated by the sum of intensity measurements over masses and log transformed. For every $\mathbf{X}_i$ the relevant nonsingular matrix $\mathbf{S}_i$ of order 200, as given in (1), is calculated.

Cui et al. (2016) and Lin et al. (2014) conducted simulations to verify the correctness of the procedure of finding a suitable structure using discrepancy functions. Since they did not analyze the T_2 structure, and since an error was identified in the formula for T_1 (cf. Filipiak et al., 2018d) our own simulations are performed. This makes it possible to verify whether the considered discrepancy function properly detects the true covariance structure from data simulated from a normal distribution with a given covariance structure. Therefore, we generated 100 data matrices $\mathbf{X} \sim N_{m,n}(\boldsymbol{\mu} \otimes \mathbf{1}_n^\top, \boldsymbol{\Sigma}, \mathbf{I}_n)$, where $n = 1000$ (sample size) and $m = 100$ (matrix order) and $\boldsymbol{\Sigma}$ has one of the considered structures CS, T_1 and AR(1) with $\sigma^2 = 2$, $\varrho = \varrho_1 = 0.25$ and T_2 with $\sigma^2 = 2$, $\varrho_1 = 0.25$, $\varrho_2 = 0.1$. We chose the same parameters as in Cui et al. (2016) and Lin et al. (2014) to obtain comparable results, and the number of simulations was taken as 100, since the algorithm proposed by Filipiak et al. (2018d) is time-consuming (it depends on the dimension of the matrix). Table 1 gives the results obtained from our simulations for all of the considered structures. The first column represents the true covariance matrix $\boldsymbol{\Sigma}$, whilst in the following columns the discrepancy indices ξ_F and ξ_E for four estimates of covariance structures are given. Optimal results for ξ_F and ξ_E for each structure are highlighted in bold.

It can be seen that both functions recognize the structure properly. Since ξ_F and ξ_E are not related to each other, it is difficult to compare them. Moreover, it is difficult to verify how good the approximations are. Therefore, the discrepancies are normalized to show how far they are from each other and to see if the goodness of approximation is at the same level.

The discrepancy obtained by the Frobenius norm is normalized by the Frobenius norm of the matrix $\mathbf{S}_i$, that is $\xi_F^N = \xi_F / ||\mathbf{S}_i||_F$, whilst the adjusted discrepancy based on the entropy loss function is obtained as $\xi_E^N = 1 - 1/[1 + \log(\xi_E + 1)]$. To make the results of both methods comparable, the aim of

this adjustment is to flatten the results into the interval $[0;1)$, where 0 means that the original matrix is close to the appropriate structure, while 1 means that it is distant from it. Normalized discrepancies used in a real data example are presented in Table 2.

Table 1. Simulation results for CS, T_1 and AR ($m = 100$, $\sigma^2 = 2$, $\varrho = \varrho_1 = 0.25$) and T_2 ($m = 100$, $\sigma^2 = 2$, $\varrho_1 = 0.25$, $\varrho_2 = 0.1$)

True structure $\boldsymbol{\Sigma}$	Set of structures			
	CS	T_1	T_2	AR
CS	$\boldsymbol{\xi_F = 1.78}$	$\xi_F = 49.25$	$\xi_F = 48.75$	$\xi_F = 14.93$
	$\boldsymbol{\xi_E = 0.56}$	$\xi_E = 3.11$	$\xi_E = 3.10$	$\xi_E = 3.11$
T_1	$\xi_F = 6.97$	$\boldsymbol{\xi_F = 0.12}$	$\xi_F = 0.15$	$\xi_F = 1.61$
	$\xi_E = 7.80$	$\boldsymbol{\xi_E = 0.56}$	$\xi_E = 5.63$	$\xi_E = 1.04$
T_2	$\xi_F = 7.45$	$\xi_F = 2.80$	$\boldsymbol{\xi_F = 0.15}$	$\xi_F = 1.08$
	$\xi_E = 6.37$	$\xi_E = 1.70$	$\boldsymbol{\xi_E = 0.56}$	$\xi_E = 0.85$
AR	$\xi_F = 15.74$	$\xi_F = 8.08$	$\xi_F = 4.02$	$\boldsymbol{\xi_F = 0.21}$
	$\xi_E = 21.96$	$\xi_E = 5.38$	$\xi_E = 1.72$	$\boldsymbol{\xi_E = 0.56}$

Table 2. The discrepancies and normalized discrepancies for the considered covariance structures and three data sets

structure	subset	ξ_F	ξ_E	ξ_F^N	ξ_E^N
CS	1	1399.1336	655.5110	0.3271	0.7380
	2	1387.0465	588.9344	0.2992	0.7348
	3	1512.8434	619.9874	0.2852	0.7364
T_1	1	4235.6872	659.3827	0.9902	0.7382
	2	4594.1416	589.3869	0.9910	0.7348
	3	5257.9262	623.2153	0.9912	0.7365
T_2	1	4219.1889	659.3437	0.9864	0.7382
	2	4572.1144	589.0878	0.9862	0.7348
	3	5232.9662	623.1698	0.9865	0.7365
AR	1	1404.7870	659.3951	0.3284	0.7382
	2	1379.1035	589.0559	0.2975	0.7348
	3	1477.2560	623.2280	0.2785	0.7365

From the results given in Table 2 we can conclude that when the Frobenius norm is used, the most relevant structure is compound symmetry or autoregression of order one. Such structures lead to a completely different image of covariance dependencies in the data sets under consideration, and this depends on the selected subsets. Of course, it is possible that the assessments of different estimators are very similar or even the same. Therefore, in order to make the best choice, additional studies should be carried out.

When using the entropy loss function the most relevant structure is compound symmetry; however, the differences between the discrepancies for the considered covariance structures are very small. Since all of the results obtained with the use of the entropy loss function are very similar (see the last column of Table 2) we do not recommend this discrepancy function for regularization.

In Table 3 we compare the estimators $\hat{\boldsymbol{\Psi}}$ and $\tilde{\tilde{\boldsymbol{\Psi}}}$, given in (2) and (3) respectively, for every structure under consideration. Moreover, to compare the obtained estimators with those commonly used in statistics, we determine the MLEs ($\check{\boldsymbol{\Psi}}$) of the parameters of the considered covariance structures. Formulae for the MLEs of unknown components of CS and AR(1) can be found in Filipiak et al. (2017). Since maximum likelihood estimation of a banded Toeplitz matrix is challenging and no explicit form of MLEs for T_1 and T_2 exist (cf. Christensen, 2007), they are not determined in this paper.

From Table 3 it can be seen that the estimates $\tilde{\tilde{\boldsymbol{\Psi}}}$ are much different than $\hat{\boldsymbol{\Psi}}$ in all cases. We obtain different estimates, since we use different methods in which functions lead to different solutions. However, it may be observed that $\hat{\boldsymbol{\Psi}}_{CS}$, $\hat{\boldsymbol{\Psi}}_{T_1}$ and $\hat{\boldsymbol{\Psi}}_{T_2}$ are comparable. Moreover, $\hat{\boldsymbol{\Psi}}_{CS} = \check{\boldsymbol{\Psi}}_{CS}$.

Indeed, using the spectral decomposition of $\boldsymbol{\Psi}_{CS}$, that is

$$\boldsymbol{\Psi}_{CS} = \sum_{i=1}^{2} \alpha_i \mathbf{P}_i$$

with $\mathbf{P}_1 = \mathbf{Q}_{1_m}$, $\mathbf{P}_2 = \mathbf{I}_m - \mathbf{Q}_{1_m}$ and $\alpha_1 = 1 - \varrho$, $\alpha_2 = 1 + (p-1)\varrho$, and differentiating $f_F(\mathbf{S}, \boldsymbol{\Psi})$ with respect to α_i we obtain

$$\frac{\partial f_F}{\partial \alpha_i} = \frac{-1}{2\sqrt{\operatorname{tr}(\mathbf{S} - \boldsymbol{\Psi}_{CS})^2}} \operatorname{vec}^\top \mathbf{I}_m \left[(\mathbf{S} - \boldsymbol{\Psi}_{CS}) \otimes \mathbf{I}_m + \mathbf{I}_m \otimes (\mathbf{S} - \boldsymbol{\Psi}_{CS}) \right] \operatorname{vec} \mathbf{P}_i.$$

Using Magnus and Neudecker (1986) and after equating the above to zero, we obtain

$$\operatorname{vec}^\top \mathbf{I}_m \cdot \{ \operatorname{vec} \left[\mathbf{P}_i \left(\mathbf{S} - \boldsymbol{\Psi}_{CS} \right) \right] + \operatorname{vec} \left[\left(\mathbf{S} - \boldsymbol{\Psi}_{CS} \right) \mathbf{P}_i \right] \} = 0$$

which is equivalent to

$$\operatorname{tr} \left[\mathbf{P}_i (\mathbf{S} - \boldsymbol{\Psi}_{CS}) \right] = 0. \tag{5}$$

For $\mathbf{X} \sim N_{m,n}(\boldsymbol{\mu} \otimes \mathbf{1}_n^\top, \boldsymbol{\Psi}_{CS}, \mathbf{I}_n)$ the log-likelihood function has the form

$$\ln L = -\frac{mn}{2} \ln(2\pi) - \frac{n}{2} \ln |\boldsymbol{\Psi}_{CS}| - \frac{1}{2} \operatorname{tr} \left[\left(\mathbf{X} - \boldsymbol{\mu} \otimes \mathbf{1}_n^\top \right)^\top \boldsymbol{\Psi}_{CS}^{-1} \left(\mathbf{X} - \boldsymbol{\mu} \otimes \mathbf{1}_n^\top \right) \right]$$

and, for $\widehat{\mu} = \frac{1}{n}\mathbf{X}\mathbf{1}_n$, it can be written as

$$\ln L = -\frac{mn}{2}\ln(2\pi) - \frac{n}{2}\ln|\mathbf{\Psi}_{CS}| - \frac{n}{2}\operatorname{tr}\left(\mathbf{S}\mathbf{\Psi}_{CS}^{-1}\right).$$

Table 3. Entropy loss estimators, Frobenius norm estimators and maximum likelihood estimators of parameters of the covariance structure for three data sets

estimator	subset	σ^2	ϱ	ϱ_1	ϱ_2
$\widehat{\mathbf{\Psi}}_{CS}$	1	30.7322	0.6555	-	-
	2	29.5329	0.7475	-	-
	3	33.2362	0.7635	-	-
$\widetilde{\mathbf{\Psi}}_{CS}$	1	0.0632	0.5145	-	-
	2	0.0599	0.4558	-	-
	3	0.0443	0.3737	-	-
$\breve{\mathbf{\Psi}}_{CS}$	1	30.7322	0.6555	-	-
	2	29.5329	0.7475	-	-
	3	33.2362	0.7635	-	-
$\widehat{\mathbf{\Psi}}_{\mathrm{T}_1}$	1	34.4749	-	0.5001	-
	2	35.8953	-	0.5001	-
	3	40.5434	-	0.5001	-
$\widetilde{\mathbf{\Psi}}_{\mathrm{T}_1}$	1	0.0310	-	0.0495	-
	2	0.0340	-	0.1329	-
	3	0.0281	-	0.0015	-
$\widehat{\mathbf{\Psi}}_{\mathrm{T}_2}$	1	36.0277	-	0.5351	0.4136
	2	39.1610	-	0.5405	0.4114
	3	44.3696	-	0.5398	0.4117
$\widetilde{\mathbf{\Psi}}_{\mathrm{T}_2}$	1	0.0310	-	0.0015	-0.0004
	2	0.0341	-	0.0049	0.0013
	3	0.0281	-	0.0015	-0.0004
$\widehat{\mathbf{\Psi}}_{AR}$	1	20.7750	0.9996	-	-
	2	23.4344	0.9991	-	-
	3	27.9402	0.9985	-	-
$\widetilde{\mathbf{\Psi}}_{AR}$	1	0.0310	0.0485	-	-
	2	0.0342	0.1494	-	-
	3	0.0281	0.0525	-	-
$\breve{\mathbf{\Psi}}_{AR}$	1	30.6336	0.6808	-	-
	2	29.4393	0.8230	-	-
	3	32.9569	0.8277	-	-

Differentiating the above with respect to α_i we obtain

$$\frac{\partial \ln L}{\partial \alpha_i} = -\frac{n}{2}\operatorname{vec}^\top \mathbf{\Psi}_{CS}^{-1}\operatorname{vec}\mathbf{P}_i + \frac{n}{2}\operatorname{vec}^\top \mathbf{S}\cdot\left(\mathbf{\Psi}_{CS}^{-1}\otimes\mathbf{\Psi}_{CS}^{-1}\right)\operatorname{vec}\mathbf{P}_i.$$

Since $\operatorname{vec}\mathbf{ABC} = (\mathbf{C}^\top \otimes \mathbf{A})\operatorname{vec}\mathbf{B}$ (cf. Magnus and Neudecker, 1986), after equating the above to zero, we obtain

$$\operatorname{vec}^\top \mathbf{S} \cdot \operatorname{vec}\left(\boldsymbol{\Psi}_{CS}^{-1}\mathbf{P}_i\boldsymbol{\Psi}_{CS}^{-1}\right) - \operatorname{tr}\left(\mathbf{P}_i\boldsymbol{\Psi}_{CS}^{-1}\right) = 0,$$

which is equivalent to

$$\operatorname{tr}\left[\mathbf{P}_i(\boldsymbol{\Psi}_{CS}^{-1}\mathbf{S}\boldsymbol{\Psi}_{CS}^{-1} - \boldsymbol{\Psi}_{CS}^{-1})\right] = 0,$$

and finally

$$\operatorname{tr}\left[\boldsymbol{\Psi}_{CS}^{-1}\mathbf{P}_i\boldsymbol{\Psi}_{CS}^{-1}(\mathbf{S} - \boldsymbol{\Psi}_{CS})\right] = 0.$$

From the fact that

$$\boldsymbol{\Psi}_{CS}^{-1}\mathbf{P}_i\boldsymbol{\Psi}_{CS}^{-1} = \frac{1}{\alpha_i^2}\mathbf{P}_i$$

we obtain (5).

Filipiak and Klein (2018a) compared the MLEs of $\boldsymbol{\Psi}_{CS} \otimes \boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}_{AR} \otimes \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is an unstructured matrix and $\boldsymbol{\Psi}_{CS}$, $\boldsymbol{\Psi}_{AR}$ are correlation matrices, with the estimators based on the Frobenius norm using simulation studies. In Filipiak et al. (2018b) and Filipiak et al. (2018c) the MLEs of the above structures and the estimators based on the entropy loss function were compared by simulations. In all of these papers it can be seen that the estimators are comparable for small sample sizes; however, the biasedness of the estimators based on the entropy loss function increases with the dimension of $\mathbf{S}$.

The main difference between the two methods considered in this paper is that in the case of the Frobenius norm we use the matrix $\mathbf{S}$, whilst in the case of the entropy loss function we use its inverse. Observe that if the sample size is close to the number of parameters, then $\mathbf{S}$ is close to singular or ill-conditioned, which makes the determination of $\widetilde{\boldsymbol{\Psi}}$ impossible. It is observed that the entropy loss function is widely used in statistics (see Cui et al., 2016); however,

$$\check{\boldsymbol{\Psi}}_{CS} = \widehat{\boldsymbol{\Psi}}_{CS} \neq \widetilde{\boldsymbol{\Psi}}_{CS}.$$

Concluding, for regularization as well as for the estimation we recommend the Frobenius norm as a more relevant discrepancy function.

4. Concluding remarks

The regularization method using the Frobenius norm indicates the compound symmetry matrix and autoregression of order one as the most relevant, whilst the entropy loss function gives a slight indication in favor of the compound symmetry structure. However, all normalized discrepancies in the case of the entropy loss function are close to each other. Therefore we cannot draw strong conclusions using this method. From simulations performed in Filipiak and Klein (2018a), Filipiak et al. (2018b) and Filipiak et al. (2018c) it is known that the Frobenius norm estimators are less biased than the entropy loss estimators. Therefore we would recommend the Frobenius norm as the most relevant discrepancy function, and the compound symmetry matrix or autoregression of order one as the most suitable structures. Nevertheless, a wider set of covariance structures should be considered for metabolomic data. Moreover, in this paper we assume the most general possible situation, when we do not distinguish whether the sample comes from one or another variety, treatment or plant stage. This is the initial stage, where we regularize the covariance matrices by the methods known from the literature. We realize that the estimates obtained as a result of regularization are not the best. The structures considered in the literature are still weak and require more study. Consideration of the block structure of the covariance matrix will be the subject of our further research.

It is also worth noting that one should be careful with the assumption concerning the covariance structure, because not all data sets are suitable for direct analysis using the methods presented in this paper.

Acknowledgments

The authors are very grateful to A. Kuczyńska, P. Ogrodowicz, K. Mikołajczak, K. Krystkowiak, M. Surma, and T. Adamski from the Institute of Plant Genetics, Polish Academy of Sciences, for plant material, and to B. Swarcewicz and M. Stobiecki from the Institute of Bioorganic Chemistry, Polish Academy of Sciences, for chemical analysis. This study was supported by the European Regional Development Fund through the Innovative Economy Program for Poland 2007–2013, project POLAPGEN-BD no. WND-POIG.01.03.01-00-101/08. Some of the computations were performed using a grant provided by Poznań Supercomputing and Networking Center. We are also very grateful to K. Filipiak from Poznan University of Technology

and A. Markiewicz from Poznan University of Life Sciences for helpful suggestions and remarks.

References

Chmielewska K., Rodziewicz P., Swarcewicz B., Sawikowska A., Krajewski P., Marczak Ł., Ciesiołka D., Kuczyńska A., Mikołajczak K., Ogrodowicz P., Krystkowiak K., Surma M., Adamski T., Bednarek P., Stobiecki M. (2016): Analysis of drought-induced proteomic and metabolomic changes in barley (Hordeum vulgare L.) leaves and roots unravels some aspects of biochemical mechanisms involved in drought tolerance. Frontiers in Plant Science 7: 1108.

Christensen L.P.B. (2007): An EM-algorithm for Band-Toeplitz Covariance Matrix Estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing III, Honolulu, 3: 1021–1024.

Cui X., Li X., Zhao J., Zeng L., Zhang D., Pan J. (2016): Covariance structure regularization via Frobenius norm discrepancy. Linear Algebra Appl. 510: 124–145.

Dey D.K., Srinivasan C. (1985): Estimation of a covariance matrix under Stein's loss. Ann. Statist. 13(4): 1581–1591.

Filipiak K., Klein D. (2018a): Approximation with Kronecker product structure with one component as compound symmetry or autoregression. Linear Algebra and Its Applications 559: 11–33.

Filipiak K., Klein D., Markiewicz A., Mokrzycka M. (2018b): Approximation with a Kronecker product structure via entropy loss function. Submitted.

Filipiak K., Klein D., Mokrzycka M. (2018c): Estimators comparison of separable covariance structure with one component as compound symmetry matrix. Electronic Journal of Linear Algebra 33: 83–98.

Filipiak K., Klein D., Roy A. (2017): A comparison of likelihood ratio tests and Rao's score test for three separable covariance matrix structures. Biometrical J. 59: 192–215.

Filipiak K., Markiewicz A., Mieldzioc A., Sawikowska A. (2018d): On projection of a positive definite matrix on a cone of nonnegative definite Toeplitz matrices. Electronic Journal of Linear Algebra 33: 74–82.

James W., Stein C. (1961): Estimation with quadratic loss. In: Neyman, J. (ed.) Proceedings of the Fourth Berkeley Symposium. In: Mathematical Statistics and Probability, 1: 361–379. The Statistical Laboratory, University of California Press.

Khakimov B., Gürdeniz G., Engelsen S.B. (2016): Trends in the application of chemometrics to foodomics studies. Acta Alimentaria 44: 4–31.

Lin L., Higham N. J., Pan J. (2014): Covariance structure regularization via entropy loss function. Computational Statistics and Data Analysis 72: 315–327.

Magnus J., Neudecker H. (1986): Symmetry, 0-1 matrices and Jacobians, a review. Econom. Theory 2: 157–190.

Ożarowski M., Piasecka A., Gryszczyńska A., Sawikowska A., Pietrowiak A., Opala B., Mikołajczak P.Ł., Kujawski R., Kachlicki P., Buchwaldg W., Seremak-Mrozikiewicz A. (2017): Determination of phenolic compounds and diterpenes in roots of Salvia miltiorrhiza and Salvia przewalskii by two LC–MS tools: Multi-stage and high resolution tandem mass spectrometry with assessment of antioxidant capacity. Phytochemistry Letters 20: 331–338.

Piasecka A., Sawikowska A., Krajewski P., Kachlicki P. (2016): Combined mass spectrometric and chromatographic methods for in-depth analysis of phenolic secondary metabolites in barley leaves. Journal of Mass Spectrometry 50: 513 – 532.

Piasecka A., Sawikowska A., Kuczyńska A., Ogrodowicz P., Mikołajczak K., Krystowiak K., Gudyś K., Guzy-Wróbelska J., Krajewski P., Kachlicki P. (2017): Drought related secondary metabolites of barley (*Hordeum vulgare L.*) leaves and their association with mQTLs. Plant Journal 89: 898–913.

Sawikowska A., Piasecka A., Kachlicki P., Krajewski P. (2018): Separation of co-eluted compounds by clustering and by functional data analysis. Submitted.

Swarcewicz B., Sawikowska A., Marczak Ł., Łuczak M., Ciesiołka D., Krystkowiak K., Kuczyńska A., Piślewska-Bednarek M., Krajewski P., Stobiecki M. (2017): Effect of drought stress on metabolite contents in barley recombinant inbred line population revealed by untargeted GC–MS profiling. Acta Physiol Plant 39: 158.