Two heuristic approaches to describe periodicities in genomic microarrays

Jörg Aßmus[1,2], Hans Arnfinn Karlsen[2] and Dag Tjøstheim[2]

[1]UNIFOB Health, Christies gate 13, N-5015 Bergen, Norway
[2]Department of Mathematics, Johannes Bruns gate 12, N-5008 Bergen, Norway

Corresponding author: Jörg Aßmus, joerg.assmus@unifob.uib.no, phone: +47 55583239

Abstract

In the first part we discuss the filtering of panels of time series based on singular value decomposition. The discussion is based on an approach where this filtering is used to normalize microarray data. We point out effects on the periodicity and phases for time series panels.

In the second part we investigate time dependent periodic panels with different phases. We align the time series in the panel and discuss the periodogram of the aligned time series with the purpose of describing the periodic structure of the panel. The method is quite powerful assuming known phases in the model, but it deteriorates rapidly for noisy data.

Keywords: microarray, time series, periodic panels, singular value decomposition, filtering, aligned data, Fourier transformation, periodogram

1 Introduction

Time dependent processes have a wide range of applications in most disciplines of life science such as courses of diseases, concentrations of toxic substances or genetic activity in cells. The investigation of such processes is often done in parallel for several patients, tissues or genes, so that several time series are produced. Considering the same experiment for different patients or genes one can expect common structures for all or at least groups of the time series contained in the panel. Those structures can for example be trends or periodicities.

With the introduction of the microarray technique for monitoring the genetic activity in cells, large data panels are produced both in biology (Spellman et al., 1998; Alter et al., 2000) and medical research (Cooper, 2001). There are both time dependent and time independent studiess. For time dependent experiments one gets panels of time series that typically contain a much larger number of time series than time points. If there are common structures in the panel the isolated investigation of single time series can lead to loss of information because the knowledge of the common structures is not used. In this paper we will discuss two heuristic approaches to benefit from common structures in data panels. Motivated by the investigation of cell cycles, we study periodicities in a genetic microarray sample.

One of the most used benchmark data sets in the analysis of microarray data is the monitoring of the cell cycles for the yeast *Saccharomyces cerevisiae* by Spellman et al. (1998). Even if this data set is quite old and not studied with a medical problem in mind it is well suited for illustrating and comparing techniques in the analysis of microarray data. Alter et al. (2000) use these data to develop a normalization method of the data set by filtering with respect to a singular value decomposition (SVD). One can get the impression that this is an attempt to extract periodically expressed genes, even though they are not periodic. In the first part we investigate, using simulated data, how the normalization influences the data. We show that the interpretation of SVD-normalized data is not as simple as it may look. So we hope that our work can contribute to reduce the number of pitfalls in the interpretation of microarray data.

On the other hand the discussion whether there are periodicities raises the question how such periodicities should be detected and estimated in a data set. Little has been done in this field and in the description of time dependent microarrays, generally. There are tests to detect periodicities in each single time series like that of Fisher (1929) with applications to panel data sets like microarrays, see e.g. Wichert et al. (2004). These tests are based on the periodogram, i.e. the Fourier analysis which generally deals with periodicities in time series (Brillinger, 1981). They are mostly concerned with the description of one function. There are many cases where it is reasonable to assume that there are more than one process going on in the same basic cycle or with the same base frequency, like the cell cycle in gene expressions or many phenomena connected to daily or annual cycles in biology or geophysics. It is therefore desirable to develop methods to take advantage of the similarities. Such a process often contains several frequencies and in such a case the periodogram calculated from the data gives important information. Diggle and al Wasel (1997) discuss the approach of average periodograms in a panel of biological time series. The problem is that time series in microarray analysis are often very short such that the number of frequencies which can be investigated is very limited. We will discuss an approach to avoid this problem for panels

of periodic time series where the frequency structure is the same for all of the time series, but the phase may vary in the panel. By aligning the time series so that the aligned series have (approximately) the same phase, we produce a data vector with very dense observations so that we also can resolve higher frequencies.

2 The singular value decomposition approach

Alter et al. (2000) discuss the structure of microarray data sets using the SVD approach. They detect a cyclic behavior in SVD-normalized gene expressions of the yeast *Saccharomyces cerevisiae* monitored by Spellman et al. (1998). After a brief introduction to the mathematical procedure they use, we will discuss the interpretation of the results using several simulated data sets.

2.1 Normalization of gene expression microarrays using SVD

The SVD is a multivariate technique to transform a matrix into a representation in an orthogonal system. Given a matrix $\mathbf{Y} = \{Y_{ij}\}_{i=1,..,N,\, j=1,..,M}$ containing the expression of N genes at M time points and defining $L = \min\{N, M\}$ the SVD is a linear decomposition of $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{UDV}^T \tag{2.1}$$

into an $N \times L$ *eigengene*-matrix $\mathbf{U}$, a diagonal $L \times L$ *eigenexpression*-matrix $\mathbf{D}$ and an $M \times L$ *eigenarray*-matrix $\mathbf{V}$. The columns of U and V are eigengenes and eigenarrays, respectively. This decomposition is also known as the Karhunen-Loéve decomposition or the Principal Component Analysis (PCA) if the matrix $\mathbf{Y}$ is quadratic and positive semi-definite, i.e. a covariance matrix.

The diagonal values of the eigenexpression-matrix determine how much of the behavior of $\mathbf{Y}$ can be explained by the corresponding eigencomponents in the orthogonal system. Manipulating the eigenexpressions, for example, by removing components, we can filter the matrix $\mathbf{Y}$ with respect to the properties described by the filtered eigencomponents. This idea is used in Alter et al. (2000) to normalize microarray data sets by removing undesirable components. We will now briefly introduce the procedure with its main steps to enable us to discuss the results in that paper and extend it to simulated examples. We will not describe the procedure comprehensively and ignore some aspects (pattern inference, degenerated subspace rotation). There is a very detailed description of the entire procedure in Alter et al. (2000).

Let us consider the panel

$$\mathbf{Y} = \mathbf{E}(\mathbf{Y}) + \boldsymbol{\varepsilon} \quad , \tag{2.2}$$

represented in an experiment with realizations $\mathbf{y} = \{y_{ij}\}_{i=1,..,N,\, j=1,..,M}$. $\mathbf{E}(\mathbf{Y})$ is the expectation of $\mathbf{Y}$ and $\boldsymbol{\varepsilon} = \{\varepsilon_{ij}\}_{i=1,..,N,\, j=1,..,M}$ is a zero mean noise term.

1. **Normalization of the trend.** Let $\mathbf{y}$ be a realization of model (2.2). We compute the SVD (2.1)
$$\mathbf{y} = \mathbf{udv}^T \quad .$$
Motivated by the shape of most microarray data sets we assume $M < N$ such that we have $L = M$. We will denote the columns $\{u_{ij}\}_{i=1,...,N}$ and $\{v_{ij}\}_{i=1,...,M}$ by $\mathbf{u}_j$ and $\mathbf{v}_j$, respectively, and the jth diagonal element of $\mathbf{d}$ by d_j. To normalize the data we remove the first eigencomponent from the data set,
$$\mathbf{y}_C = \mathbf{y} - d_1 \mathbf{u}_1 \mathbf{v_1}^T \quad . \tag{2.3}$$
This first eigenexpression represents the eigencomponent explaining the largest proportion of the variance in the data set. In the investigations of Alter et al. (2000) it is interpreted as a constant trend in the data set.

2. **Normalization of the variation.** Using the trend-normalized data panel $\mathbf{y}_C$, we introduce
$$y_{LV,ij} = \ln(y_{C,ij}^2) \quad .$$
We normalize this scaling parameter as above by computing the SVD of $\mathbf{y}_{LV}$
$$\mathbf{y}_{LV} = \mathbf{u}_{LV}\mathbf{d}_{LV}\mathbf{v}_{LV}^T \quad ,$$
removing its first eigencomponent which represents the largest variation in the data set,
$$\mathbf{y}_{CLV} = \mathbf{y}_{LV} - d_{LV,1}\mathbf{u}_{LV,1}\mathbf{v}_{LV,1}^T \quad , \tag{2.4}$$
and transforming it back,
$$y_{N,ij} = \mathrm{sgn}(y_{C,ij})\sqrt{\exp\{y_{CLV,ij}\}} \quad ,$$
where $\mathrm{sgn}(y_{C,ij})$ denotes the sign of $y_{C,ij}$. The panel $\mathbf{y}_N = \{y_{N,ij}\}$ is now normalized both with respect to the trend (2.3) and the variation (2.4).

3. **Data Sorting.** After the normalization we sort the genes with respect to the relative correlation ψ_i between their normalized expression $\mathbf{y}_{N,i}$ and the first two eigenarrays $\mathbf{v}_{N,1}$ and $\mathbf{v}_{N,2}$ of the normalized panel, i.e. we compute the SVD of the normalized data
$$\mathbf{y}_N = \mathbf{u}_N\mathbf{d}_N\mathbf{v}_N^T$$
and then the relative correlation ψ_i given by
$$\tan\psi_i = \frac{\mathbf{Corr}(\mathbf{y}_{N,i}, \mathbf{v}_{N,1})}{\mathbf{Corr}(\mathbf{y}_{N,i}, \mathbf{v}_{N,2})} \tag{2.5}$$
Looking at the two correlations in (2.5) as Cartesian coordinates the relative correlation ψ_i is the corresponding angle in polar coordinates. Note that the pair
$$[\mathbf{Corr}(\mathbf{y}_{N,i}, \mathbf{v}_{N,1}), \mathbf{Corr}(\mathbf{y}_{N,i}, \mathbf{v}_{N,2})]$$
always will be inside the unit circle, as shown in Alter et al. (2000).

We next explore the properties of the technique used in Alter et al. (2000) by applying it to simulated data sets.

2.2 Application of the SVD approach to different data sets

Alter et al. (2000) apply the SVD-normalization to the gene expression data of the yeast *Saccharomyces cerevisiae* monitored by Spellman et al. (1998). This data set is often used as a benchmark in microarray analysis. Different synchronization techniques are used. We will concentrate here on one of them, the elutriation synchronization.

The data set contains the gene expression for $N = 5981$ genes measured at 30min intervals in approximately one cell cycle with the period length $T \approx 390$min, i.e. we have $M = 14$ time points. Spellman et al. (1998) classify 784 genes as cell cycle regulated. The behavior of these genes with respect to the normalization and sorting procedure is comprehensively discussed by Alter et al. (2000). Very interesting are the different periodicities appearing in the results. Even if not all genes are classified as cell cycle regulated the normalized and sorted data set is periodic both in time and over genes as we can see in the upper left plot of figure 1. In addition there is a discussion of how well the elutriation synchronization is able to extract the periodic properties of gene expression for example in Shedden and Cooper (2002). The arguments are supported by empirical studies in Wichert et al. (2004) and Aßmus (2006). This leads to the question of how to interpret the different periodicities in the normalized and sorted data.

Let us consider a panel $\mathbf{y} = \{y_{ij}\}_{i=1,..,N,\, j=1,..,M}$ using the model

$$y_{ij} \;=\; f_i(t_j + \phi_i) + \alpha + \beta_i + \gamma_j + \varepsilon_{ij} \qquad (2.6)$$

for $M = 15$ time points $t_j = 0, \frac{1}{M-1}, \frac{2}{M-1}, ..., 1$ and $N = 5000$ genes, where
- α is a general constant effect,
- $\boldsymbol{\beta} = \{\beta_i\}_{i=1,...,N}$ a gene specific effect (vertical),
- $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1,...,M}$ an array or time specific effect (horizontal),
- $\boldsymbol{\phi} = \{\phi_i\}_{i=1,...,N}$ gene specific phases and
- $\boldsymbol{\varepsilon} = \{\varepsilon_{ij}\}_{i=1,...,N, j=1,...,M}$ a zero mean noise term.

We investigate this model by examining several functions f_i and noise terms $\boldsymbol{\varepsilon}$:

- `cos0`: The cosine function

$$f_i(t_j, \phi_i) \;=\; \cos(2\pi(t_j + \phi_i)) \qquad (2.7)$$

  with independent standard normal noise $\boldsymbol{\varepsilon}$ added

- `cos1`: The cosine function (2.7) with correlated noise generated by transforming the columns $\boldsymbol{\varepsilon}'_j$ of an independent standard normal matrix $\boldsymbol{\varepsilon}'$

$$\boldsymbol{\varepsilon}_j = \mathbf{A}\boldsymbol{\varepsilon}'_j, \quad j = 1, ..., M \quad ,$$

  and using the vectors $\boldsymbol{\varepsilon}_j$ as columns of the noise matrix $\boldsymbol{\varepsilon}$. The elements of $\mathbf{A}$ are random, uniformly distributed in the interval $[-1, 1]$.

- `cos2`: A sum of cosine functions with different frequencies

$$f_i(t_j, \phi_i) \;= \cos(2\pi(t_j + \phi_i)) + 0.5\cos(6\pi(t_j + \phi_i))$$

  with independent standard normal noise $\boldsymbol{\varepsilon}$ added

- `cos500`: A partial cosine function

$$f_i(t_j, \phi_i) \;=\; \begin{cases} \cos(2\pi(t_j + \phi_i)) & i \le 500 \\ 0 & i > 500 \end{cases} \quad ,$$

  with independent standard normal noise $\boldsymbol{\varepsilon}$ added

- `randn`: Independent standard normal noise $\boldsymbol{\varepsilon}$, i.e. $f_i = 0$.

We investigated the cosine functions for different choices of the parameters α, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ in model (2.6). While the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are more or less motivated by controlling possible artefacts, we will pay more attention to the choice of α and $\boldsymbol{\phi}$.

Table 1: Parameter choices in model (2.6)

code	α	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\phi}$
`a`	0	0	0	**U**[-0.5,0.5]
`b`	0.01	0	0	**U**[-0.5,0.5]
`c`	5	0	0	**U**[-0.5,0.5]
`d`	5	0	**N**(0,1)	**U**[-0.5,0.5]
`e`	5	**N**(0,1)	0	**U**[-0.5,0.5]
`f`	5	0	0	$0.0625\chi^2_1$
`g`	0	0	0	$0.0625\chi^2_1$
`h`	0	0	0	0
`i`	5	0	0	0

We simulated each of the functions `cos0`, `cos1`, `cos2` and `cos500` with each of the parameter sets given in table 1. For the noise model `randn` we will only use `a`, `c`, `d` and `e` because all other choices are either already included or do not make sense.

Note that we treat neither $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ or $\boldsymbol{\phi}$ as random variables. These are arbitrary fixed values which are randomly chosen. Only for $\boldsymbol{\phi}$ we will consider the distribution of the chosen values.

The singular value decomposition is a linear method that rotates a matrix into a system of orthogonal bases, which are sorted with respect to how much of the values in the data set is explained by the several base vectors. Generally, it is a linear method and not designed to take care of time dependencies, i.e. ordered index sets, or nonlinear functions. Nevertheless, assuming a time dependent nonlinear function in the rows we will find a decomposition into a function base in the columns of $\mathbf{v}$ and the columns of $\mathbf{u}$. Removing the first component both with respect to trend and variation in (2.3) and (2.4), we remove the component explaining the major part of the trend and the variation in the data set.

The constant α. Alter et al. (2000) interpret the removed first component as a constant trend. Since

the normalization removes the component explaining the largest values, we can expect (ignoring the effect of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) that α is removed when it is larger than the values of $f_i(t_j + \phi_i)$. And indeed we obtain very different patterns in the normalized and sorted data assuming $\alpha = 0$ and $\alpha = 0.01$ (fig. 2:`cos0b`) on one hand and $\alpha = 5$ (fig. 2:`cos0c`) on the other. While we get uniformly distributed phases of $\mathbf{y}_N$ for $\alpha = 5$ we roughly observe only two different values of the phases of $\mathbf{y}_N$ for the small values of α. In the latter case the largest part of the variation of $\mathbf{y}$ is explained by the cosine oscillation and since the cosine function is decomposed into

$$\begin{aligned} & A\cos(\omega(t_j + \phi_i)) \\ &= A\cos(\omega\phi_i)\cos(\omega t_j) + A\sin(\omega\phi_i)\sin(\omega t_j) \\ &= B_i\cos(\omega t_j) + C_i\sin(\omega t_j) \end{aligned} \tag{2.8}$$

$B_i\cos(\omega t_j)$ is filtered out by the normalization. The first two eigenarrays $\mathbf{v}_{N,1}$ and $\mathbf{v}_{N,2}$ confirm this conclusion (fig. 3). In figure 3:`cos0c` we find for $\alpha = 5$, that this constant is taken out by the normalization, leaving a clear cosine for $\mathbf{v}_{N,1}$ (red) and a clear sine for $\mathbf{v}_{N,2}$ (blue), while $\alpha = 0.01$ in figure 3:`cos0b` produces a sine for the first eigenarray $\mathbf{v}_{N,1}$ (which was the second eigenarray in the other case) and a second eigenarray $\mathbf{v}_{N,2}$, whose shape is not interpretable as a periodic function. The cosine part is not expressed anymore. In the remaining data the phases ϕ_i are only contained in the coefficients C_i and filtered out by the scale filtering (2.4), except their sign. That is why we find two phases with a difference of a half period length for small values of $\alpha = 0$ and $\alpha = 0.01$.

The gen specific effect $\boldsymbol{\beta}$ and the array specific effect $\boldsymbol{\gamma}$. A vertical effect $\boldsymbol{\beta}$ did not affect the results of our experiments. We do not present the corresponding plots, since they are essentially equal to figure 2:`cos0c` for $\mathbf{y}_N$ and 3:`cos0c` for the eigenarrays $\mathbf{v}_{N,1}$ and $\mathbf{v}_{N,2}$. Obviously, the absolute values of vertical effects are filtered out by the scale normalization (2.4) since the sign is kept as we saw above for C_i in (2.8).

In contrast to $\boldsymbol{\beta}$ we find $\boldsymbol{\gamma}$ clearly expressed as a horizontal structure in $\mathbf{y}_N$ (fig. 2:`cos0d`) and as the deviation from a sine function in figure 3:`cos0d`.

The phases $\boldsymbol{\phi}$. Considering the example `cos0c` assuming uniformly distributed phases $\boldsymbol{\phi}$ one can obtain a relation between the phases ϕ_i, the relative correlation ψ_i and the phases in the normalized data $\mathbf{y}_N$ since all are uniformly distributed: We fitted a Beta(r,s) distribution to ψ_i and estimated both parameters very close to 1, i.e. a uniform distribution. For comparison we did the same for ϕ_i, which we know are uniformly distributed (see table 2). Furthermore we find a clear linear relation between ϕ_i and ψ_i as it is shown in the upper right plot of figure 4. Translating the cluster $\Omega = \{\psi_i > 0,\ \phi_i < 0.45\}$ on the right hand side by one period length, we consider the transformed phase ϕ'_i,

$$\phi'_i = \begin{cases} \phi_i + 1 & \phi_i \in \Omega \\ \phi_i & \phi_i \notin \Omega \end{cases} ;$$

we find the linear regression equation:

$$\phi' = 0.16\,\psi + 0.78 \quad ,$$

with $R^2 = 0.9529$.

Table 2: Estimated parameters of the Beta(r,s) distributions

	$\hat{r}$	$\hat{s}$
`cos0c` ϕ_i	0.9739	0.9805
`cos0c` ψ_i	1.0059	1.0037
`cos0f` ψ_i	1.0100	1.0303

Nevertheless, we can not conclude from the knowledge of ψ_i to the properties of ϕ_i. Assuming for example a very skewed distribution of ϕ_i like we did using a downscaled χ^2_1 distribution (see the histogram of ϕ_i in figure 4:`cos0f` on the left hand side) the distribution of ψ_i is uniform, as the estimated parameters of a Beta(r,s) distribution indicate (see table 2), and the clearly visible relation we obtained in the other case disappears as seen in lower right plot of figure 4:`cos0f`. The phases of the normalized data $\mathbf{y}_N$ are uniformly distributed as well (fig. 2:`cos0f`). The first two eigenarrays in figure 3:`cos0f`) do not differ from the corresponding eigenarrays in figure 3:`cos0c`.

Assuming now a constant phase, all periodic structure is filtered out in the normalized data $\mathbf{y}_N$ as seen in figure 2:`cos0i`. Here, the coefficient C_i in (2.8) vanishes such that the second component disappears and the first is filtered out. The relative correlations ψ_i are still uniformly distributed.

Obviously, the normalization changes the phase structure of a cosine function. We did not investigate other functions here. Possibly this is an effect of the addition theorem allowing the decomposition (2.8) and so a property of the trigonometric functions.

The different models. Comparing the two cosine models `cos0` and `cos500` the behavior of the normalized and sorted data $\mathbf{y}_N$ is very similar if we have a strong constant term α (fig. 2:`cos0c` and 1:`cos500c`), although the pattern looks somewhat clearer for `cos0`, where the eigenarrays are much closer to cosine/sine functions than those for `cos500` (fig. 3:`cos0c` and `cos500c`). If not all genes follow a cosine, the first component of (2.8) can not explain the behavior of all genes, and we can therefore expect that a common trend will be filtered out first. For `cos500` there are 90% non trigonometrically explainable genes and indeed none of the two trigonometric components are removed (fig. 1:`cos500b`) as it happens for `cos0` (recall fig. 2:`cos0b`). We note for all experiments using `cos500` that even if only a small part of the genes are cyclic (10%) the normalized expressions behave cyclic for most of the genes. The periodicity "leaks out" to the entire data set. Considering only one arbitrary non-cyclic gene it may seem like the normalization creates a periodicity.

The differences between cos0 and cos500 get clearer when we consider the correlations between the normalized data and the first two eigenarrays $\mathbf{v}_{N,1}$ and $\mathbf{v}_{N,2}$. Considering `cos0` both the first two eigenarrays and the normalized genes are cyclic, such that we find a large number of high correlations (fig. 6:`cos0c`) and very few correlations around zero. For `cos500` the eigenarrays $\mathbf{v}_{N,1}$, $\mathbf{v}_{N,2}$ and the normalized data are cyclic as well but more disturbed, so that the correlations are much lower (fig. 6:`cos500c`). For model `cos0` with a small constant effect α such that the first cyclic component is filtered out, all rows of $\mathbf{y}_N$ are still cyclic but the second eigenarray is not, such that we get high correlations with respect to $\mathbf{v}_{N,1}$ but not with respect to $\mathbf{v}_{N,2}$ producing the non radial figure 6:`cos0b`.

Assuming now the mixed model of two cosine functions (`cos2`) we get four cyclic eigenarrays since generalizing to K cosine functions we have

$$\begin{aligned}
&\sum_{k=1}^{K} A_k \cos(k\omega(t_j+\phi_i)) \\
&= \sum_{k=1}^{K} A_k \cos(k\omega\phi_i)\cos(k\omega t_j) + A_k \sin(k\omega\phi_i)\sin(k\omega t_j) \\
&= \sum_{k=1}^{K} B_{ki}\cos(k\omega t_j) + C_{ki}\sin(k\omega t_j) \quad .
\end{aligned}$$

Assuming now a small constant effect α and filtering out the first cyclic component, the second eigenarray becomes the first component with a higher frequency as shown in figure 3:`cos2b`, leading to a clearly visible high frequency structure in the normalized data (fig. 1:`cos2b`). Assuming a large α the pattern of the sorted normalized expressions $\mathbf{y}_N$ is similar to the other two discussed models but the high and the low value regions are narrower (fig. 1:`cos2c`). The high frequency is not visible because it is less expressed in the data. So the correlations shown in figure 6:`cos2c` are only slightly lower than for `cos0` even though the first two eigenarrays show the same clear sine/cosine as in figures 3:`cos0c` and 3:`cos0f`. In this case it is very difficult to distinguish between the models `cos0` and `cos2` only considering the plots of one data set for each model.

The cosine models with independent noise lead to clear patterns. In contrast to that we can not find a similar behavior by adding correlated noise. Except an oscillation at the first two time points of the normalized data $\mathbf{y}_N$ (fig. 2:`cos1c`), which is difficult to interpret, we can not see any structures, neither in the normalized data nor in the eigenarrays (fig. 3:`cos1c`).

The normalized data or the eigenarrays of the pure Gaussian noise model did not show any clear structures so we did not present them here.

We observed only one effect common for all considered data sets: The first two components of the ordered eigengenes of the normalized data set show an oscillation with a phase difference of $N/4$ even for the independent noise data set (fig. 5) as well as the observed data of Alter et al. (2000). Probably, this is not a property of the data set. On the other hand, all data sets we considered were very regular and they all contained periodic structures if we consider the constant function $f_i(t_j) = 0$ as a degenerated periodic function. Maybe we have to use data sets with more irregularities like strongly correlated clusters or nonperiodic functions to change this.

The elutriation data. For the yeast *Saccharomyces cerevisiae* we consider the entire data set containing the elutriation synchronized genes which was investigated by Alter et al. (2000) (denoted in the plots by `elutriation`) and the sub-sample of cell cycle regulated genes given in Spellman et al. (1998) (`elutriation CCR`). Assuming that the CCR genes follow the cell cycle, i.e. are periodically expressed and the entire data set is a mixture of periodically and nonperiodically expressed genes we expect that the normalized elutriation and cell cycle regulated elutriation data sets behave similar to `cos500` and `cos0` respectively. The plots of $\mathbf{y}_N$ in the figures 1:`elutriation` and 1:`elutriation CCR` indicate that. In both cases we obtain the typical structures of a time dependent effect which is described and interpreted by Alter et al. (2000) and observed in our experiments (fig. 2:`cos0d`). On the other hand the behavior of the correlations differ a lot. Figure 6:`elutriation` shows that the correlations of the normalized data with respect to the first two eigenarrays fill the unit circle. They reach values closer to 1 than we observed in any simulation, including the pure cosine models `cos0` and `cos2`. At the same time we do not have the empty space around zero as in figure 6:`cos0c`. Obviously, there must be genes, whose normalized expression follow a linear combination of the first two eigenarrays very tightly, and genes which do not behave like that. The assumption that the CCR-genes could be the first ones is not supported by their correlation plot in figure 6:`elutriation CCR`. We can not observe special properties like for example a higher frequency of high correlations. It seems to be a subsample of the entire data set.

Considering the phases, we note that we should not be led to conclude from the plots of $\mathbf{y}_N$, where the phases are uniformly distributed, that the phase distribution in the raw data set is uniform. In Aßmus (2006) a Beta distribution is fitted to the data with the result that the phases of the periodically expressed genes are not uniformly distributed. A uniform distribution would also have been difficult to interpret from a biological point of view.

Finally, to summarize we conclude that the SVD-normalization as it is investigated in Alter et al. (2000) is a powerful method to remove unwanted effects from a data set if we are able to locate them in the eigenarrays. It should, however, be used cautiously to avoid the removing of effects, that are important for the description of the data and even more because the interpretation of the normalized data is not as obvious as it may appear.

3 A Discrete Fourier Approach

In the previous section we discussed the singular value decomposition approach to describe microarray panels. As we saw it is difficult to interpret the result since the time dependent structure is changed by the method. Now we will introduce a heuristic method to investigate periodicities in microarray panels.

3.1 Basics

Let us first consider only one gene which is periodically expressed and assume that it follows different harmonic cycles with the corresponding frequencies ω_k and amplitudes A_k. In addition we assume a non frequency dependent phase ϕ such that we get at a time point t the expression Y,

$$\begin{aligned} Y(t) &= \sum_{k=1}^{K} A_k \cos(\omega_k(t+\phi)) + \varepsilon_t \\ &= \sum_{k=1}^{K} B_k \cos(\omega_k t) + C_k \sin(\omega_k t) + \varepsilon_t \quad , \end{aligned} \tag{3.1}$$

where ε_t denotes a zero mean noise term which is independent for different time points.

Considering measurements at M equidistant time points $t_j,\ j = 1, ..., M$ and the specific frequencies $\omega_k = k$, a very common and useful approach is the spectral analysis using the discrete Fourier transformation (DFT) where we transform the data from time domain into frequency domain,

$$\begin{aligned} \tilde{Y}(\omega_l) &= \sum_{j=0}^{M-1} Y_j \mathrm{e}^{-2\pi i j l} \quad , \\ \omega_l &= \frac{2\pi l}{M\Delta t}, \quad l = 0, ...M-1 \quad , \end{aligned}$$

where $\Delta t = t_j - t_{j-1}$ denotes the length of a time interval and i the complex imaginary unit. The values of the Fourier transformed expression $\tilde{Y}(\omega_l)$ are complex where the real part corresponds to B_k, and the imaginary part to C_k in model (3.1). To find how much of the variation in the function is explained by the frequencies ω_l we use the periodogram $I(\omega_l)$

$$I(\omega_l) = \tilde{Y}(\omega_l)\tilde{Y}^*(\omega_l), \quad l = 0, ..., L \quad , \tag{3.2}$$

where $\tilde{Y}^*(\omega_l)$ is the conjugated complex Fourier transformed expression at ω_l and L the largest integer smaller than $M/2$.

In model (3.1) the phase ϕ is contained in B_k and C_k. This complex structure is eliminated by the periodogram, i.e. we decompose the time series into the frequency domain and need no knowledge about the phase ϕ.

The weakness of this approach is that the number of frequencies which can be used is strongly limited especially if the time series is very short as is the case in microarray analysis. But in most microarray experiments we have many short time series of similar shape, each of them containing information about the frequency structure. Furthermore, if we have, for example, uniformly distributed phases as in the normalized elutriation data in the previous section, the entire panel contains more information than each single gene expression. We will use this information by aligning the gene expressions to achieve more dense observations.

3.2 The Aligned DFT Approach

Let us now consider a panel $\mathbf{Y} = \{Y_{ij}\},\ i = 1, ..., N,\ j = 1, ..., M$ of time dependent gene expressions for N genes at M time points. Furthermore we assume for all genes common amplitudes A_k, a common frequency ω_0 and gene specific phases ϕ_i. Using (3.1) with $\omega_k = k\omega_0$ for each gene we get at each of the time points $\mathbf{t} = [t_1, ..., t_M]$ the model

$$Y_{ij} = \sum_{k=1}^{K} A_k \cos(k\omega_0(t_j + \phi_i)) + \varepsilon_{ij} \tag{3.3}$$

with the realizations $\mathbf{y} = \{y_{ij}\},\ i = 1, ..., N,\ j = 1, ..., M$.

The expression functions of the genes differ only in the phase ϕ_i, i.e. assuming known ϕ_i we have in fact NM realizations of the same function. Diggle and al Wasel (1997) use that the periodogram is phase independent, compute the periodogram for each row corresponding to the genes in microarrays, and investigate the average of the periodograms. This method would be able to improve on the periodogram. On the other hand the number of frequencies we can estimate is strongly restricted in the microarray case since we usually have quite few time points. Assuming $M = 14$ time points as in the elutriation data we can not consider more than $M/2 = 7$ frequencies as seen in (3.2).

In Aßmus (2006) the maximum likelihood estimation and a mixed model to estimate parameters in the model are investigated. But since we have to solve complicated equations numerically we are quite restricted in the number of parameters. The maximum likelihood approach gives a possibility to estimate the phases ϕ_i. Once knowing the phases we can align the genes by relocating the time points such that all genes have different time points but the same phase. Now we can put them in one vector such that we have NM data points in an interval which is maximally twice as large as the original time window if we restrict the phases ϕ to one period $\phi \in [-\frac{\pi}{\omega_0}, \frac{\pi}{\omega_0}]$.

Since the time points of the aligned vector are not necessarily equidistant, we have interpolated them to an equidistant lattice $\boldsymbol{\tau}$ with M_τ points. The interpolated vector can be analyzed by Fourier analysis.

Assuming now a panel $\mathbf{y} = \{y_{ij}\},\ i = 1, ..., N,\ j = 1, ..., M$ as realizations of the model (3.3), as it is shown for one frequency and $N = 2$ in figure 7a, we suggest the algorithm

1. Estimate the phases ϕ_i for all genes for example using the maximum likelihood estimation denoting the estimates by $\hat{\phi}_i$.

2. Translate the time points for each gene i by $\hat{\phi}_i$,

$$\tilde{t}_{i,j} := t_j - \hat{\phi}_i \quad . \tag{3.4}$$

3. Create the aligned data vector $[\tilde{\mathbf{t}}, \tilde{\mathbf{y}}]$ using the alignment time points $\tilde{t}_j$,

$$[\tilde{\mathbf{t}}, \tilde{\mathbf{y}}] := [[\tilde{t}_{11}, y_{11}], ..., [\tilde{t}_{1M}, y_{1M}], [\tilde{t}_{21}, y_{21}], ... \\ ..., [\tilde{t}_{NM}, y_{NM}]] \quad , \tag{3.5}$$

as it is illustrated in figure 7(b). Considering now $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{y}}$ respectively as one vector we denote the jth element of the vector defined in (3.5) as $\tilde{t}_j$ and $\tilde{y}_j$.

4. Choose an equidistant lattice

$$\boldsymbol{\tau} := [\tau_1, ..., \tau_{M_\tau}] \tag{3.6}$$

and interpolate the function on τ estimating the conditional expectation

$$y_{\tau,j} := \mathbf{E}[\tilde{Y} |\ \tilde{t} = \tau_j] \quad , \tag{3.7}$$

treating $\tilde{Y}(\tilde{t})$ as a random variable which the aligned data $\tilde{\mathbf{y}}(\tilde{\mathbf{t}})$ are realizations of. An example is given in figure 7(c), where the estimator in (3.12) is used to estimate the conditional expectation..

5. Compute the periodogram of the interpolated time series $y_{\tau,j}$ illustrated in figure 7(d).

Aliasing and maximal resolution. Assuming $Y(t)$ in model (3.1) known for all $t \in \mathbb{R}$ the frequencies ω_k are uniquely determined, because for two cosines with fixed amplitude A, a fixed phase ϕ_0 and frequencies ω and ω' it follows from $A\cos(\omega(t+\phi_0)) = A\cos(\omega'(t+\phi_0))$ for all $t \in \mathbb{R}$ that $\omega = \omega'$.

On the other hand if we consider a data set where we do not know the function for its entire range but only for the measured discrete time points $\mathbf{t} = t_1, ..., t_M$, then the frequencies are not uniquely determined. We have to restrict the frequency range. So the DFT is only computed for a number of discrete frequencies. Assuming one function measured at equidistant time points t_j the data do not contain information about how often the function oscillates between two time points so there is an upper bound of the resolution we can achieve. From the Nyquist-Shannon Sampling Theorem which is found in most literature about signal analysis or DFT, e.g. Brillinger (1981), p.179, we know that the highest frequency we can resolve is the so called Nyquist-frequency for the model (3.1),

$$\omega_{ny} = \frac{\pi}{\Delta t} \quad ,$$

where $\Delta t = t_j - t_{j-1}$. Considering non equidistant time scales, for example assuming missing data, the usual approach is to interpolate the function to an equidistant time scale. We use the Nadaraya-Watson-estimation (3.12) in our experiments to estimate (3.7). There is a wide range of alternatives such as polynomials, splines, Newton or Lagrange methods (Schwarz, 1988). The maximal resolution is now the Nyquist-frequency of the interpolated time scale.

In our approach the interval length of the time scale for one gene is $\Delta t = (t_M - t_1)/(M-1)$ such that the Nyquist-frequency is

$$\omega_{ny} = \frac{\pi(M-1)}{t_M - t_1} \quad .$$

After aligning and interpolating the Nyquist-frequency is analogously

$$\omega_{ny} = \frac{\pi(M_\tau - 1)}{\tau_{M_\tau} - \tau_1} \quad .$$

Since the time intervals $[t_1, t_M]$ and $[\tau_1, \tau_{M_\tau}]$ are of similar size but M_τ is much larger than M the resolution of the aligned data set is much higher than for each gene.

We note that the maximal resolution is only dependent on the time scale. This means that assuming an appropriate distribution of phases we can reach a resolution as high as we want for every fixed M even in the extreme case also for $M = 1$. The only thing we have to do is to increase the number of genes N. We will not discuss here, what properties an appropriate phase distribution must have. The more uniform the distribution of time points in the aligned time vector (3.5) the more appropriate will the distribution be. A single or two point distribution, i.e. one or two fixed phases, will for example not be appropriate.

The Nyquist-frequency does not contain information of which frequencies we actually can find in the aligned data because this depends not only on the time scale but on the distribution of the phases ϕ_i, the estimated phases $\hat{\phi}$, the interpolation method and the noise level $\mathbf{Var}\varepsilon_{ij}$

Phase estimation. Since the estimated phases are subtracted from the time points in the aligned data $[\tilde{\mathbf{t}}, \tilde{\mathbf{y}}]$, the quality of the phase estimation is a strongly limiting factor of the entire procedure. To take the estimation error into account we have to consider each time point in $\tilde{\mathbf{t}}$,

$$\tilde{t}_j = \tilde{\tilde{t}}_j + \eta_j$$

as a sum of a non random $\tilde{\tilde{t}}_j$ and a noise term η_j leading to the model

$$\tilde{Y}_j = \sum_{K=1}^{K} A_k \cos(k\omega_0(\tilde{\tilde{t}}_j + \eta_j + \phi_0)) + \varepsilon_j, \qquad j = 1, ..., NM \quad . \tag{3.8}$$

It is to be expected that even small variations of $\hat{\phi}_i$ can destroy a lot of the high frequency structure of the data set. This is a serious problem since the estimation of ϕ_i can be difficult.

In Aßmus (2006) the maximum likelihood estimation of the phases and the amplitudes simultaneously

is investigated, assuming Beta(r,s) distributed phases. We fit a Beta distribution to the estimated phases and observe a bias of the estimated parameters $\hat{r}$ and $\hat{s}$ for small values of M. Especially for larger true values of r and s there is a strong bias while the distribution of the estimated phases is for $r = s = 1$ quite close to a uniform distribution.

Definition of the lattice $\boldsymbol{\tau}$. Since the elements of the aligned time scale are random the definition of the lattice $\boldsymbol{\tau}$ deserves some attention.

Naively one could assume that it is a natural choice to split the entire interval covered by the aligned time scale into $NM-1$ intervals such that we get the lattice $\boldsymbol{\tau}$:

$$\tau_j \;=\; \min(\tilde{\mathbf{t}}) + \frac{j-1}{NM-1}\left(\max(\tilde{\mathbf{t}}) - \min(\tilde{\mathbf{t}})\right) \quad . \quad (3.9)$$

If not all phases are equal, the aligned time scale $\tilde{\mathbf{t}}$ covers a larger range than the original time scale $\mathbf{t}$. Furthermore, assuming uniformly distributed phases ϕ_i there are less observations at the ends of the interval $[\min(\tilde{\mathbf{t}}), \max(\tilde{\mathbf{t}})]$ (fig. 8a). If we nevertheless use the naive lattice (3.9) we have at the ends of the covered interval much more time points than observations. Since it makes no sense to interpolate too finely, it is convenient to reduce the range of the lattice to ensure that we have enough observations. One possibility is to define the lattice within the interval $[t_1, t_M]$.

Using a reduced range for the lattice makes it necessary to reduce M_τ as well. In the example seen in figure 8a we can see, that the density of the aligned time points is not necessarily constant on the interval $[t_1, t_M] = [0, 1]$. Again, it makes no sense to interpolate too finely. We suggest using the area with least number of observations to adjust M_τ. One approach is to compute the histogram of the aligned time points in the lattice range, find the interval Δ_l (one bar of the histogram) with the lowest histogram value and choose M_τ

$$M_\tau \leq \frac{\#\{\text{observations in } \Delta_l\}}{\text{length of } \Delta_l} \cdot \#\{\text{intervals}\} \quad . \quad (3.10)$$

We suggest only an upper bound for M_τ because a meaningful choice of M_τ depends also on the distribution of the phases. To illustrate this we will only mention two extreme examples, where we use true (not estimated) phases to align the time points.

If the phases are chosen from $\{k\Delta t,\ k = 0, ..., M\}$ the aligned time points $\{t_1, ..., t_i\}$ are found on the lattice $t_1 + j\Delta t,\ j = 1, ..., 2M$, such that they are equidistant with many observations at each time point (fig. 9a). In this case it makes no sense to interpolate at all. We can only use the mean of the aligned expressions for each time point such that we only get more time intervals of the same length, i.e. no increased Nyquist-frequency. The result will be similar to the average periodogram. This case is not very realistic but it illustrates the problem of the dependence of the lattice on the distribution of the phases.

In the other extreme case, namely phases uniformly distributed at one time interval Δt, there is no overlap of the $t_j - \phi_i$ for the different values of j such that we have a uniform distribution as shown in figure 8b for a larger data set. Nevertheless the aligned time scale is not equidistant (fig. 9b) such that we have to interpolate. On the other hand we can use the naive lattice (3.9) or a lattice very close to that because of the uniformly distributed values of $\tilde{\mathbf{t}}$.

Assuming the phases uniformly distributed in $[-\frac{\pi}{\omega}, \frac{\pi}{\omega}]$ (equivalent to fig. 9c) as we do in our experiments motivated by the elutriation data set in the previous section, we can locate the interval Δ_l in (3.10) at the ends of the interval $[t_1, t_M]$. Because of the special triangle shape of the density (fig. 8a) this is a value close to $\frac{1}{2}NM$ such that we use this as an upper bound for M_τ, leading analogously with (3.9) to the lattice

$$\tau_j \;=\; t_1 + \frac{j-1}{\frac{NM}{2}-1}(t_M - t_1), \quad j = 1, ..., \frac{NM}{2} \quad (3.11)$$

for $M_\tau = \frac{1}{2}NM$.

Interpolation. To interpolate on a chosen lattice we have to specify the estimator for the conditional expectation in (3.7). There is a wide range of non- or semi-parametric smoothers. We chose the Nadaraya-Watson-estimator:

$$y_{\tau,j} \;:=\; \frac{\sum_{l=1}^{M_\tau} \tilde{y}_l \mathcal{K}(\frac{\tau_j - \tilde{t}_l}{b_\tau})}{\sum_{l=1}^{M_\tau} \mathcal{K}(\frac{\tau_j - \tilde{t}_l}{b_\tau})} \quad , \quad (3.12)$$

where $\mathcal{K}$ denotes a kernel function and b_τ the bandwidth whose choice depends on τ. We will not investigate here the influence of these parameters on the estimate. The Nadaraya-Watson-estimation is a common method which is widely discussed in the literature, e.g. Györfi et al. (1989).

3.3 Application to data sets

The focus. In this section we will use simulated data to investigate the effects discussed in the previous section. As we saw we have many possibilities to vary the procedure. There are roughly three main points:
- choice of the estimation procedure in step 1,
- choice of the lattice $\boldsymbol{\tau}$ (3.6),
- choice of the interpolation method and its parameters.

We will focus on the influence of the estimation of the phases ϕ_i and not discuss the influence of the interpolation method and the choice of the lattice τ.

The data set. We generated data sets containing N genes at M time points using the model

$$Y_{ij} = \sum_{k=1}^{4} A_k \cos(\omega_k(t_j + \phi_i)) + \varepsilon_{ij}$$

for $N = 500,\ 1000,\ 5000$, $M = 4,\ 15$, the time points $\mathbf{t} = [0, 1/(M-1), ..., 1]$ and the frequencies

$\boldsymbol{\omega} = [2\pi, 20\pi, 40\pi, 100\pi]$. All amplitudes are 1. The noise in all experiments consists of independent standard normal variables and the phase is uniform on $[0, 1]$. In all experiments we use the lattice (3.11),

$$\tau_j = \frac{j-1}{NM/2-1}, \quad j = 1, ..., \frac{NM}{2}$$

All Nyquist frequencies corresponding to the different lattices defined by N and M are much larger than the highest frequency in the data (table 3).

Table 3: Nyquist frequencies for the different generated data sets

M	single gene	aligned data		
		$N = 500$	$N = 1000$	$N = 5000$
4	3π	999π	1999π	9999π
15	14π	3749π	7499π	37499π

As interpolation method we use the Nadaraya-Watson-estimation (3.12) with a Gaussian kernel and the bandwidth $b_\tau = 5\Delta\tau$.

The experiments. For these data sets we will do two experiments, one where we estimate the phases and one where we use the known phases but simulate an estimation error (see (3.8)) such that we can control the variance of these pseudo-estimates of phases, i.e. we start in the second step of the algorithm and use in (3.4) the pertubated true phases,

$$\hat{\phi}_i = \phi(\sigma_\eta, i) = \phi_i + \eta_i, \quad \eta_i \sim \mathbf{N}[0, \sigma_\eta^2] \quad . \quad (3.13)$$

instead of the estimated phases.

Assuming that we know the phases exactly, i.e. $\sigma_\eta = 0$, all frequencies in the data are clearly resolved in the periodograms of the aligned data (upper plots of fig. 11 and 12. Even if we only have 4 time points we resolved the highest frequency $\omega_4 = 100\pi$.

Increasing the standard deviation of the σ_η, the higher frequencies disappear quite fast. Figure 10 shows, how the behavior of the periodogram value corresponding to the frequencies in the data set depends on σ_η. The noise levels at which the frequencies are not observable anymore (periodogram value reach zero level) are lowest for the highest frequencies. It matches our anticipation that the highest frequencies disappear first. Already for the low value of $\sigma_\eta = 0.007$ the highest frequency $\omega_4 = 100\pi$ is not detectable anymore, while $\omega_3 = 40\pi$ vanishes at $\sigma_\eta = 0.02$ and $\omega_3 = 20\pi$ at $\sigma_\eta = 0.035$. In the third and fourth row of the figures 11 and 12 we see that ω_2, ω_3 and ω_4 really are almost not detectable in the periodogram anymore.

Even if the periodogram value for $\sigma_\eta = 0$ is much lower for small than for the large sample experiments, the curves reach the zero level approximately for the same σ_η independent of the sample size, i.e. increasing the sample size in the investigated ranges gives no essential improvement of the histograms. The sample size determines only the level of the curves and the smoothness such that we have very similar curves for $N = 1000,\ M = 15$ ($M_\tau = 7500$) and $N = 5000,\ M = 4$ ($M_\tau = 10000$). Obviously it is the variance of η_i which determines if a frequency is resolved or not. Intuitively it is clear that the high frequency effects which determine fine scaled structures are not robust with respect to noise in the arguments. A frequency like $\omega_4 = 100\pi$ has a period length of $T = 0.02$ and assuming $\sigma_\eta = 0.08$ the probability that the error of an aligned time point is more than a half period is $Prob(|\phi(\sigma_\eta, i) - \phi_i| > 0.01) = Prob(|\tilde{t}_j - \tilde{\tilde{t}}_j| > 0.01) \approx 0.2$.

Considering now the experiments using the estimated phases in (3.4) we need some additional remarks before starting. The idea of the alignment procedure is to detect frequencies in the data set, but for a maximum likelihood model we must know or at least estimate them to get an estimation of the phases. Ignoring that and only using a one frequency model with a known leading frequency of one gene leads to a dilemma. A strong expression of the high frequencies in the data will lead to bad estimates because we can expect a large model error. If the higher frequencies are weakly expressed in the data they are weakly expressed in the periodogram as well and will be easier dominated by noise effects.

We tried to solve this problem by smoothing the data set to remove the higher frequencies before estimating the phases. The estimates of the phases have standard deviations between 0.22 ($N = 5000, M = 15$) and 0.38 ($N = 500, M = 4$), i.e. much larger than the standard deviations of η_i where the peaks in the periodogram vanished. And indeed we do not obtain the higher frequencies in the periodograms using estimated phases as shown in the lowest row of the figures 11 and 12. The phase estimates are too bad to be used in a procedure like we discussed here. To solve this problem one has to know more about the phases or improve the estimator dramatically. Since we did not know anything about the phases of the data sets we have available we abstained from an investigation of real data using this method.

These results inspired us to leave this heuristic method and investigate the maximum likelihood estimation of parameters in a cosine model in Aßmus (2006).

Acknowledgements

Finally, we would like to thank Mette Langås, Kostas Fokianos and Vidar Hjellvik for the helping and inspiring comments to improve this work.

References

Alter, O., Brown, P., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97:10101 – 10106.

Aßmus, J. (2006). *Panels of time series with application to microarrays and modular tool systems*. PhD thesis, University of Bergen.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. International Series in Decision Processes. Holt, Rinehart, and Winston, New York, NY, USA.

Cooper, C. (2001). Application of microarray technology in breast cancer research. *Breast Cancer Research*, 3:159–175.

Diggle, P. J. and al Wasel, I. (1997). Spectral analysis of replicated biomedical time series. *Applied Statistics*, 46:31–71.

Fisher, R. (1929). Tests of significance in harmonic analysis. *Proc. Roy. Soc. A*, 125:54?59.

Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Springer, New York.

Schwarz, H. R. (1988). *Numerische Mathematik*. B.G. Teubner Stuttgart.

Shedden, K. and Cooper, S. (2002). Analysis of cell-cycle gene expression in s. cerevisiae using microarrays and multiple synchronization methods. *Nucleic Acids Research*, 30:2920–2929.

Spellman, P. T., Sherlock, G., and et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

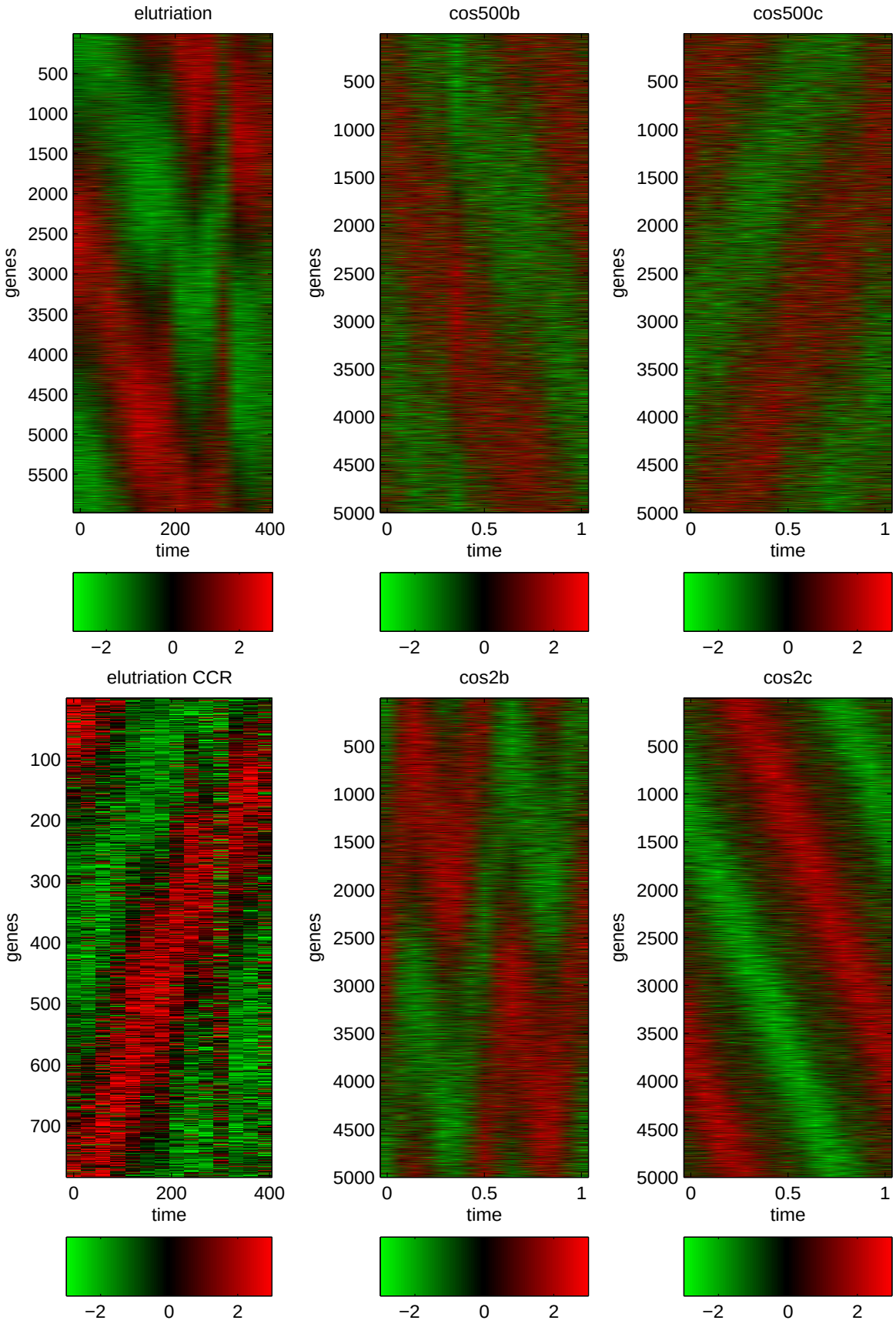

Figure 1: Plots of the normalized and sorted gene expressions $\mathbf{y}_N$ of the elutriation data and different simulated data sets (title codings are the different functions introduced in Section 2.2 and the parameter sets in table 1.

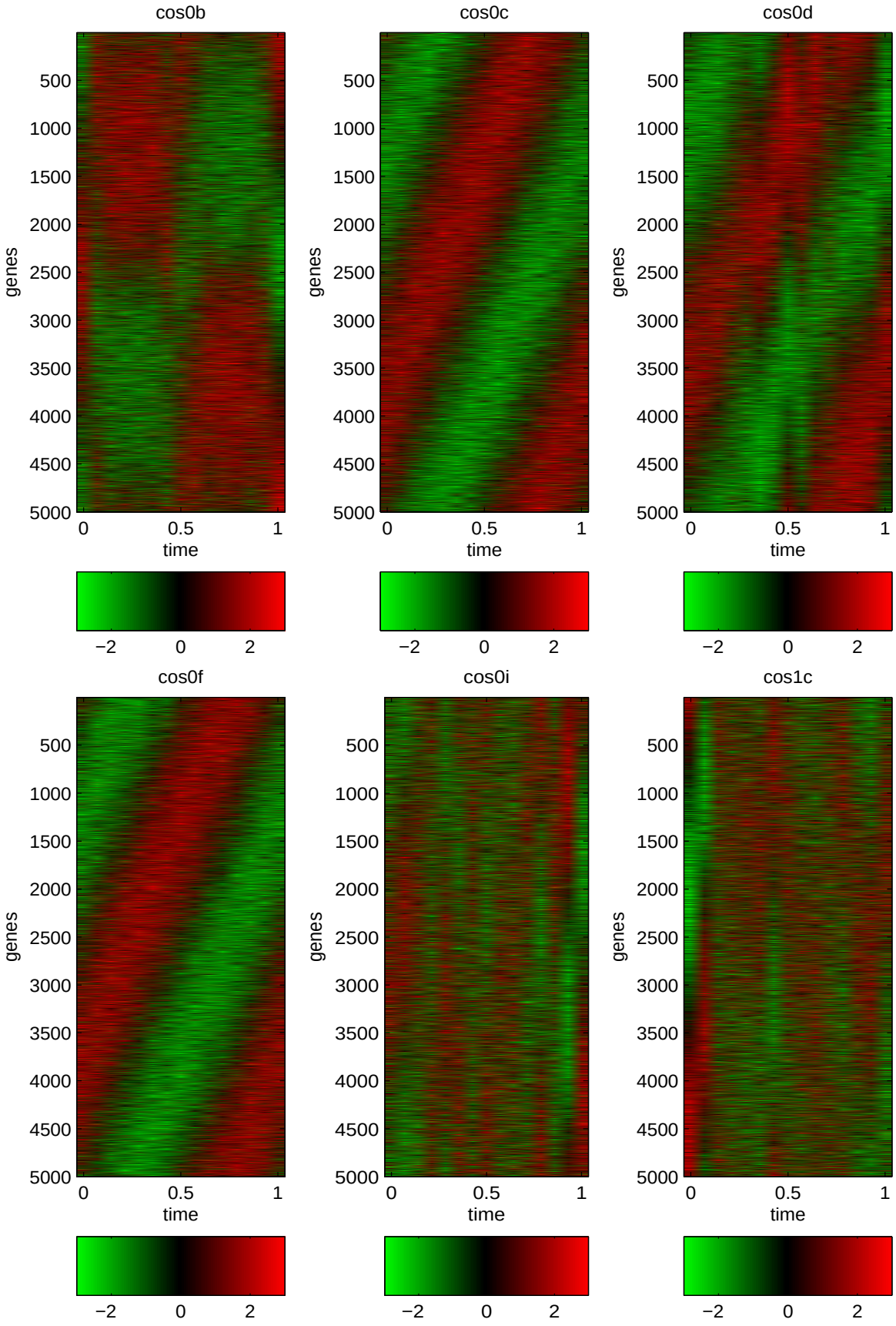

Figure 2: Plots of the normalized and sorted gene expressions $\mathbf{y}_N$ of different simulated data sets (title codings are the different functions introduced in Section 2.2 and the parameter sets in table 1.

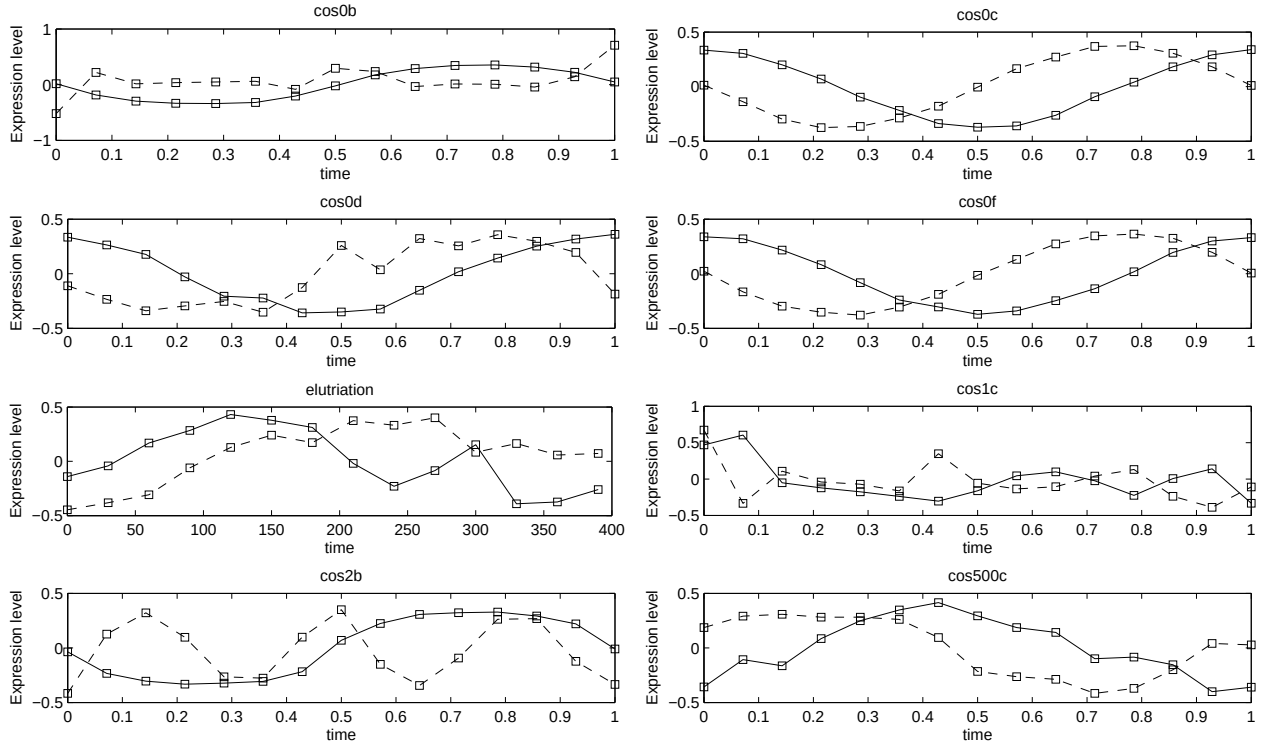

Figure 3: The first two eigenarrays $\mathbf{v}_{N,1}$ (solid) and $\mathbf{v}_{N,2}$ (dash) of the normalized and sorted gene expressions of the elutriation data and different simulated data sets (title codings are the different functions introduced in Section 2.2 and the used parameter sets in table 1.

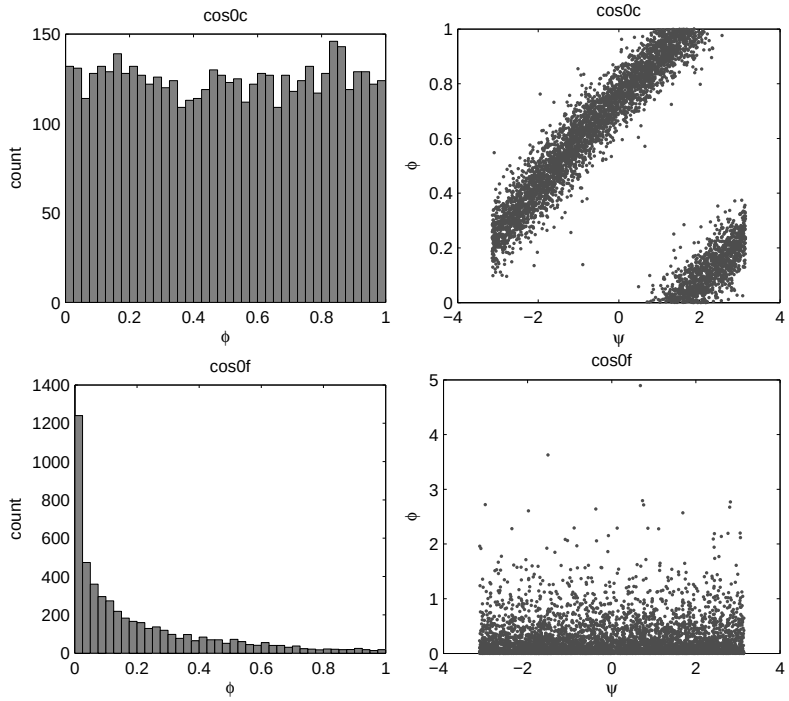

Figure 4: Distribution of the phases ϕ_i and the relative correlation ψ_i in the models `cos0c` and `cos0f`

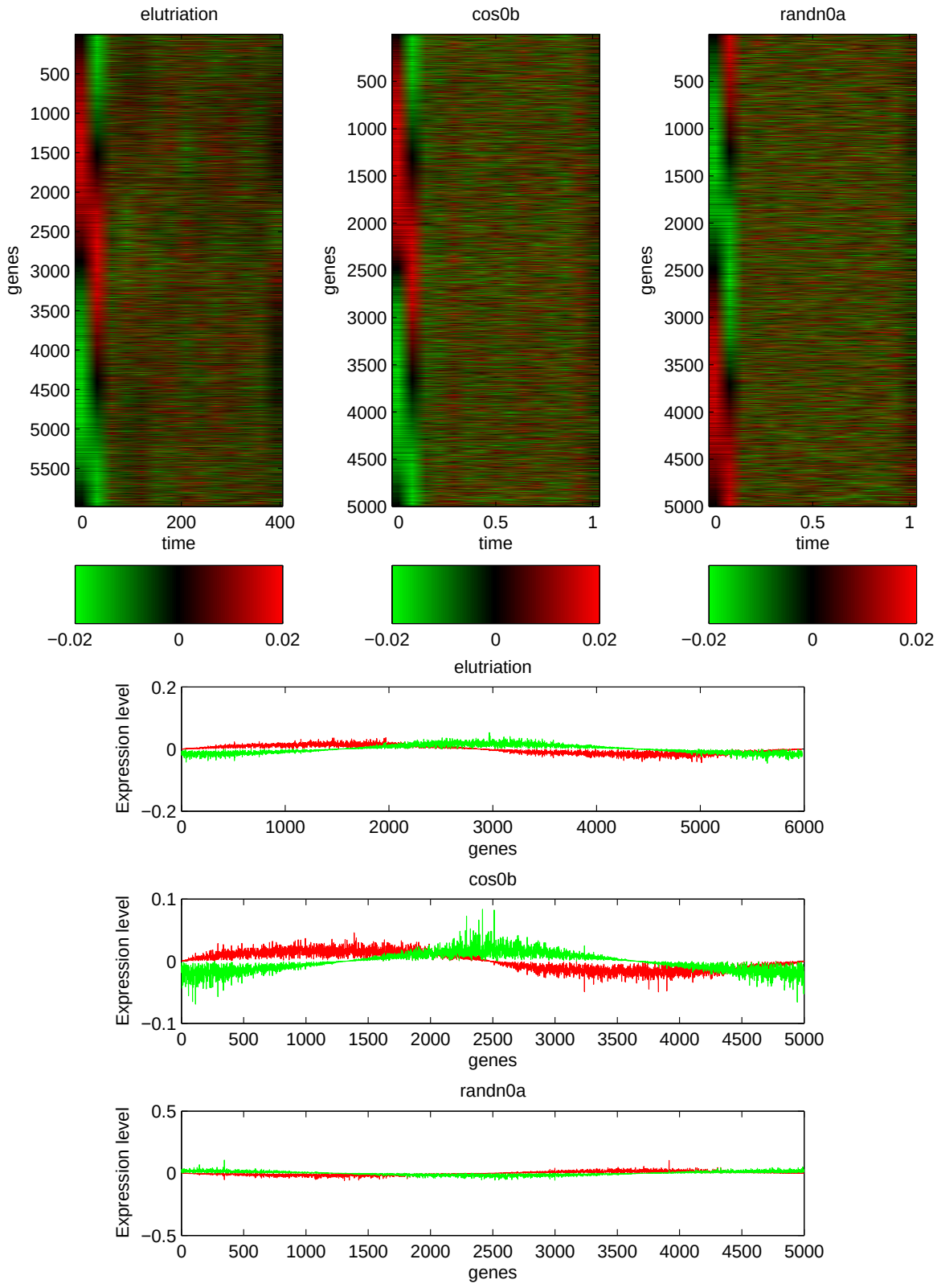

Figure 5: Eigengene plots (upper plots) and the first two eigengenes $\mathbf{u}_{N,1}$ and $\mathbf{u}_{N,2}$ (lower plots) of the normalized and sorted gene expressions of the elutriation data and different simulated data sets (title codings are the different functions introduced in Section 2.2 and the parameter sets in table 1.

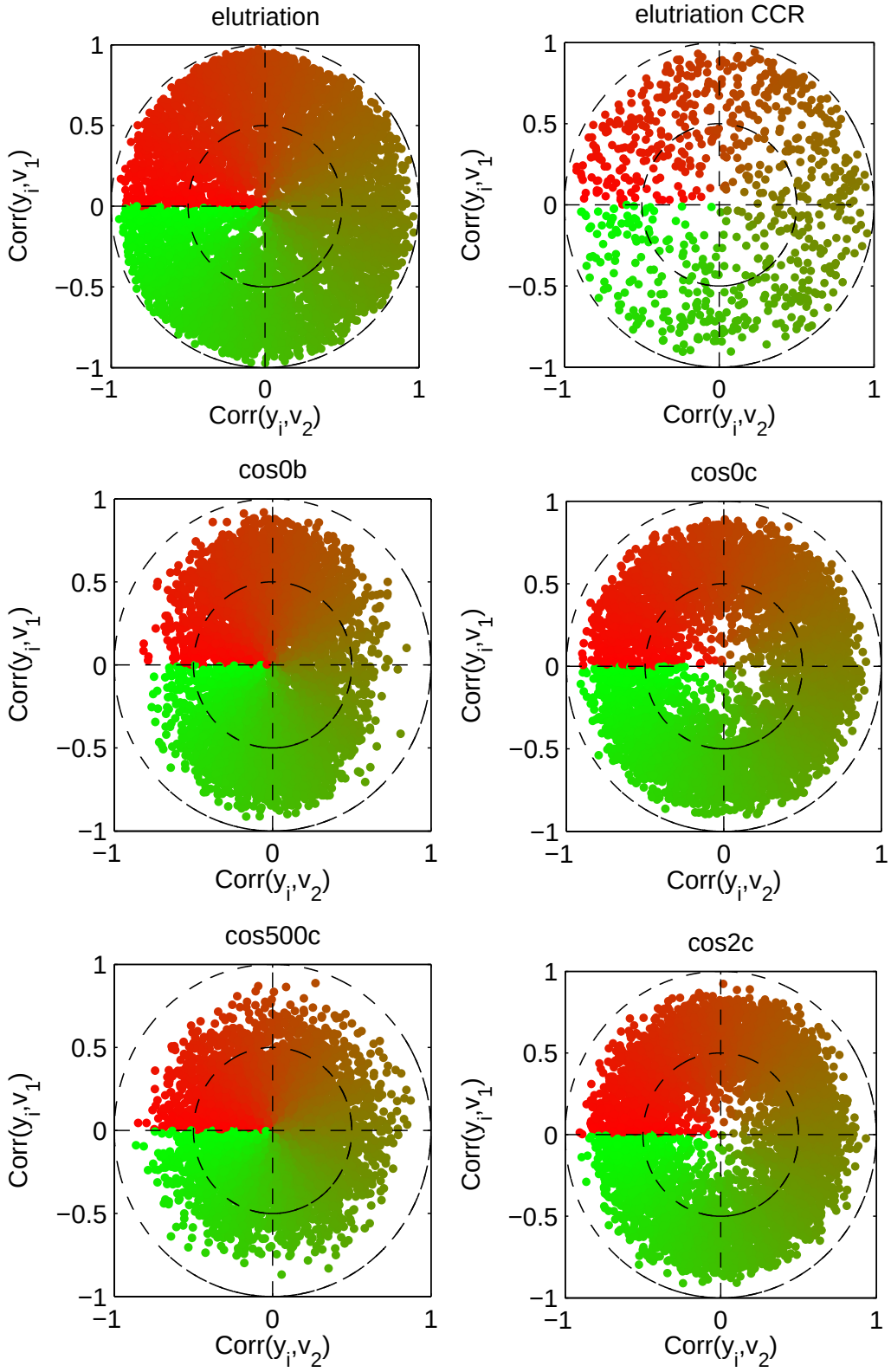

Figure 6: Correlations of the normalized gene expression with respect to their first two eigenarrays $\mathbf{v}_{N,1}$ and $\mathbf{v}_{N,2}$ for the elutriation data and different simulated data sets (title codings are the different functions introduced in Section 2.2 and the parameter sets in table 1.

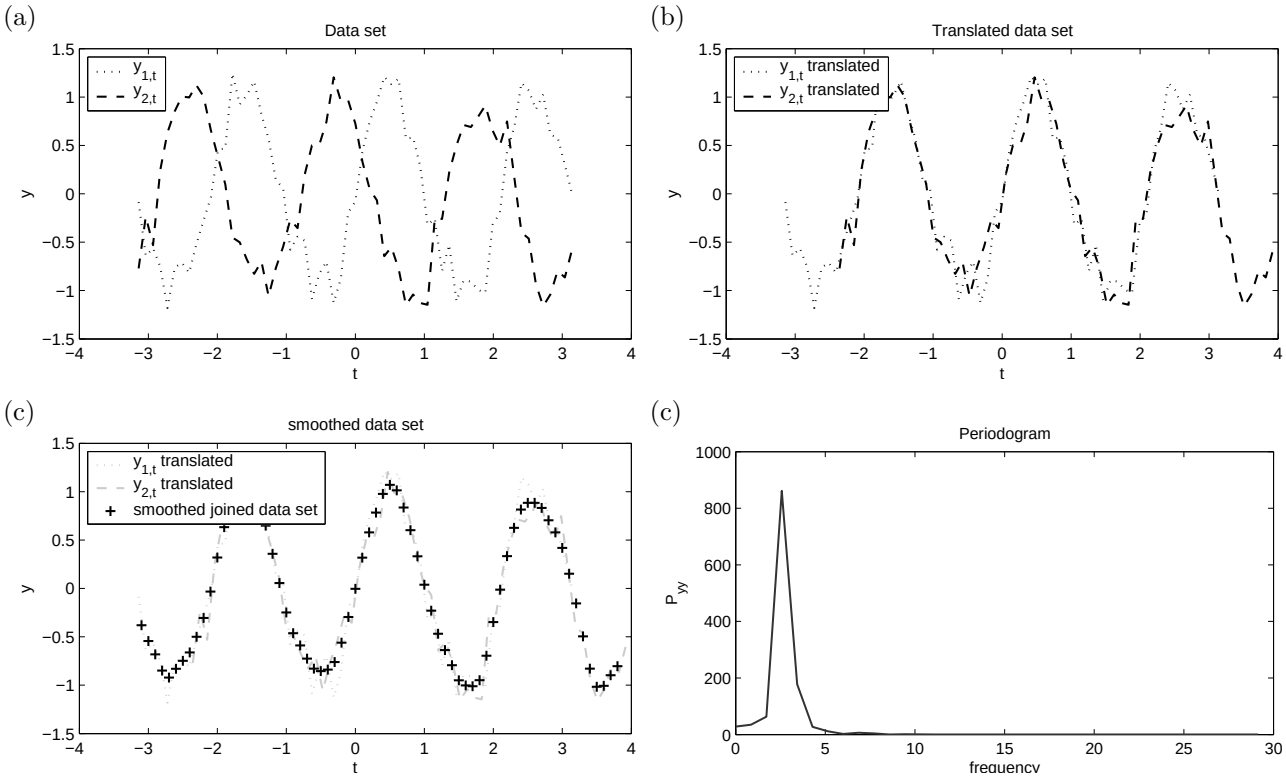

Figure 7: Algorithm for the Fourier analysis for two gene expression function (a) illustrating the alignment (b), the interpolation (c) and the DFT of the interpolated vector (d)

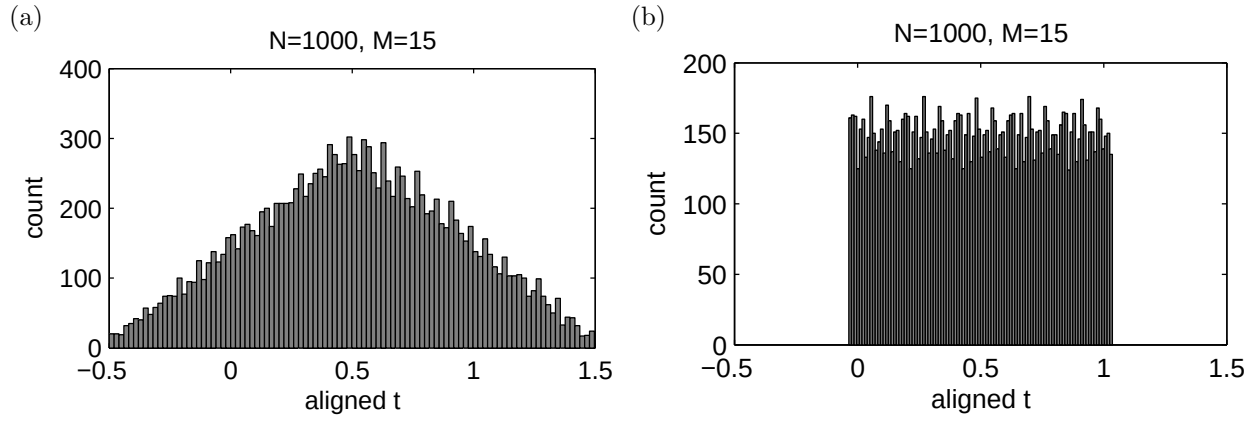

Figure 8: Histograms of the aligned time scales $\tilde{\mathbf{t}}$ for $N = 1000$ genes and $M = 15$ time points using uniformly distributed phases on $[-0.5, 0.5]$ and $[-\frac{1}{2}\Delta\mathbf{t}, \frac{1}{2}\Delta\mathbf{t}]$

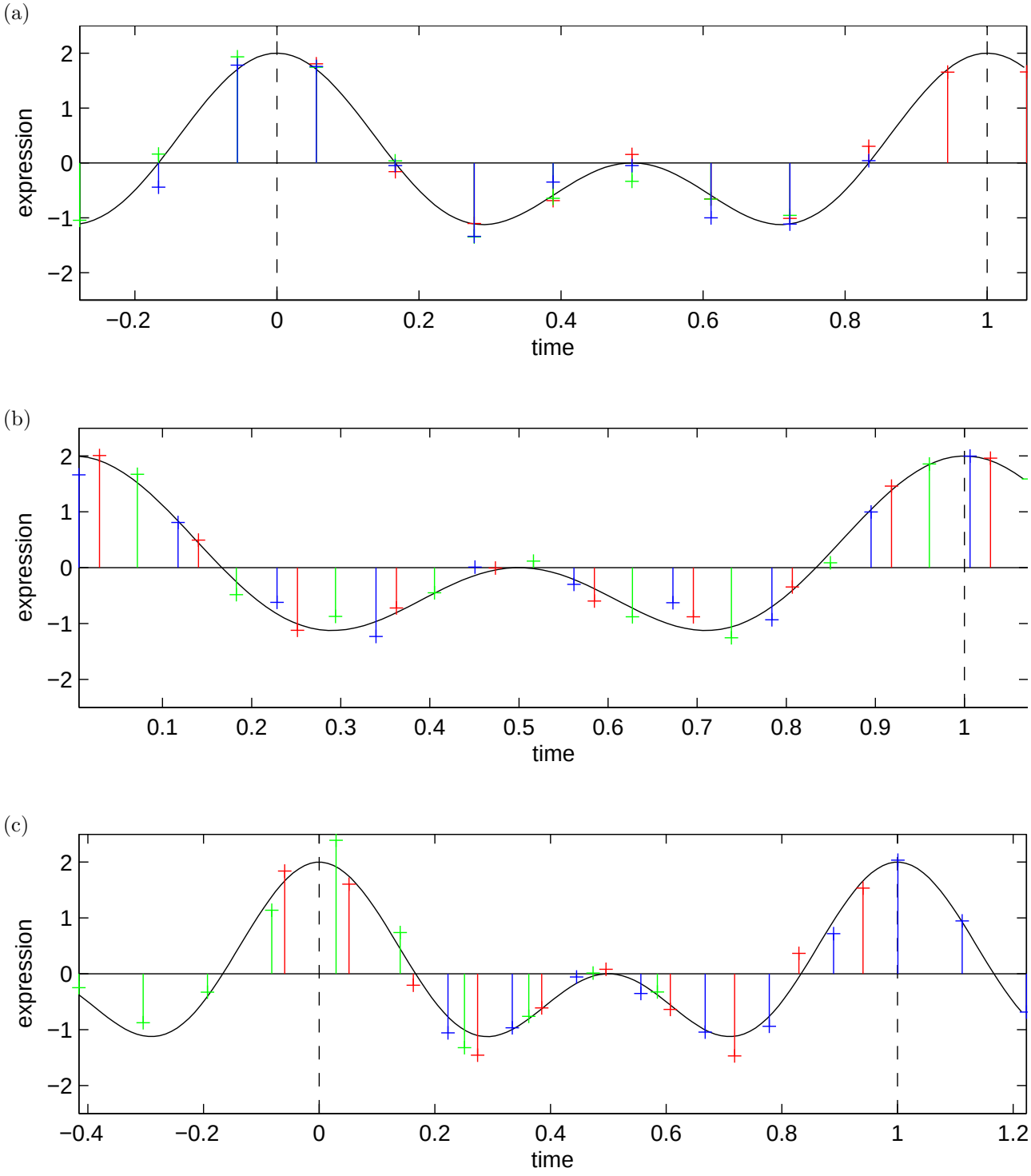

Figure 9: Aligned data set for $N = 3$ genes (represented by the different colors) at $M = 10$ time points and the true function $y_{ij} = \cos(2\pi(t_j + \phi_i) + 0.5\cos(4\pi(t_j + \phi_i)$ using $\mathbf{t} = 0, \frac{1}{M-1}, ..., 1$ for three different distributions of the phases:
(a) $\phi = [0.5 - 8\Delta t, 0.5 - 7\Delta t, 0.5 - 2\Delta t]$
(b) $\phi_i \sim \mathbf{U}[0, \frac{1}{M-1}]$
(c) $\phi_i \sim \mathbf{U}[0, 1]$

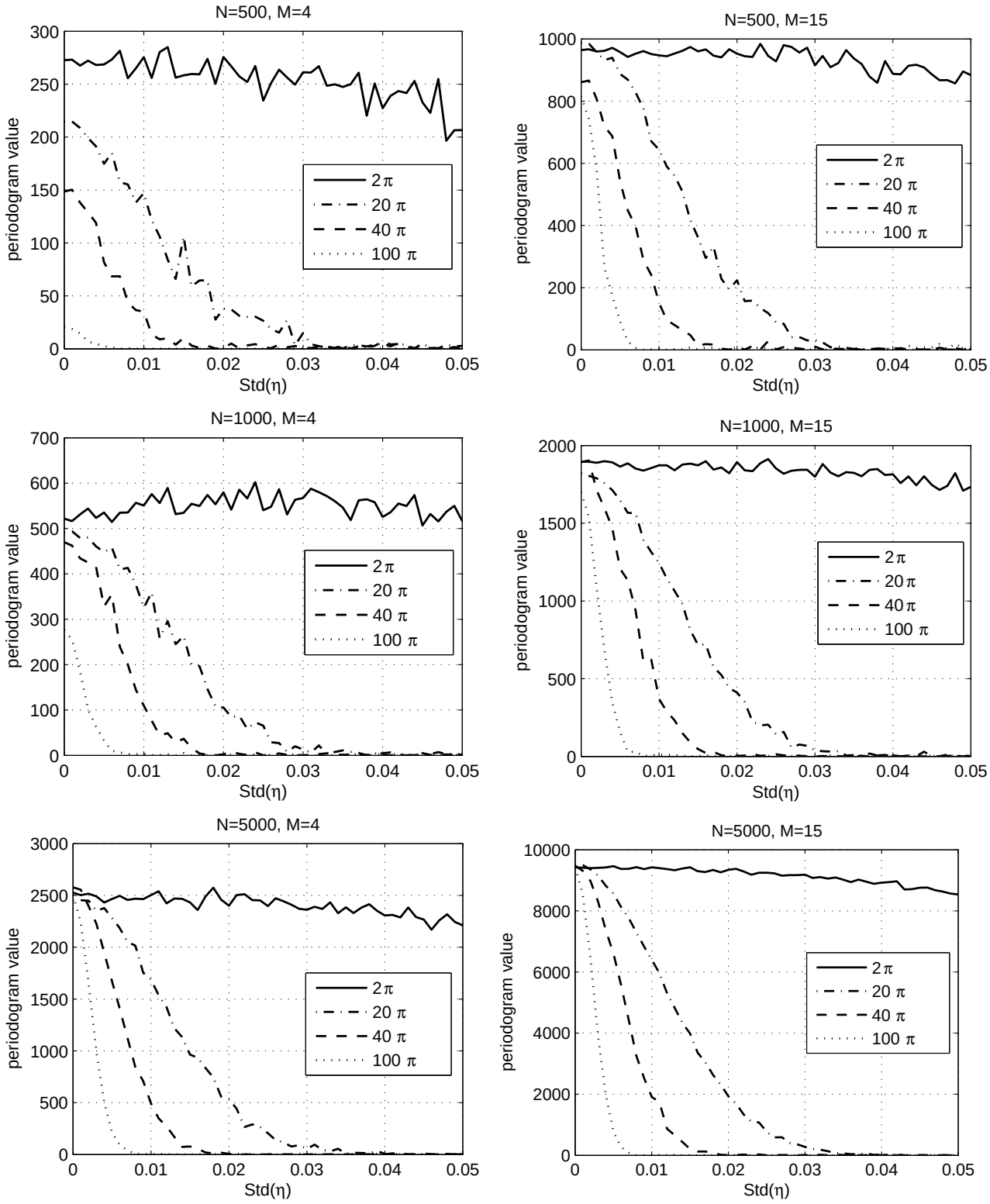

Figure 10: Periodograms of aligned time series: periodogram values for the true frequencies dependent on the noise standard deviation σ_η (3.13) for different numbers of genes N and time points M.

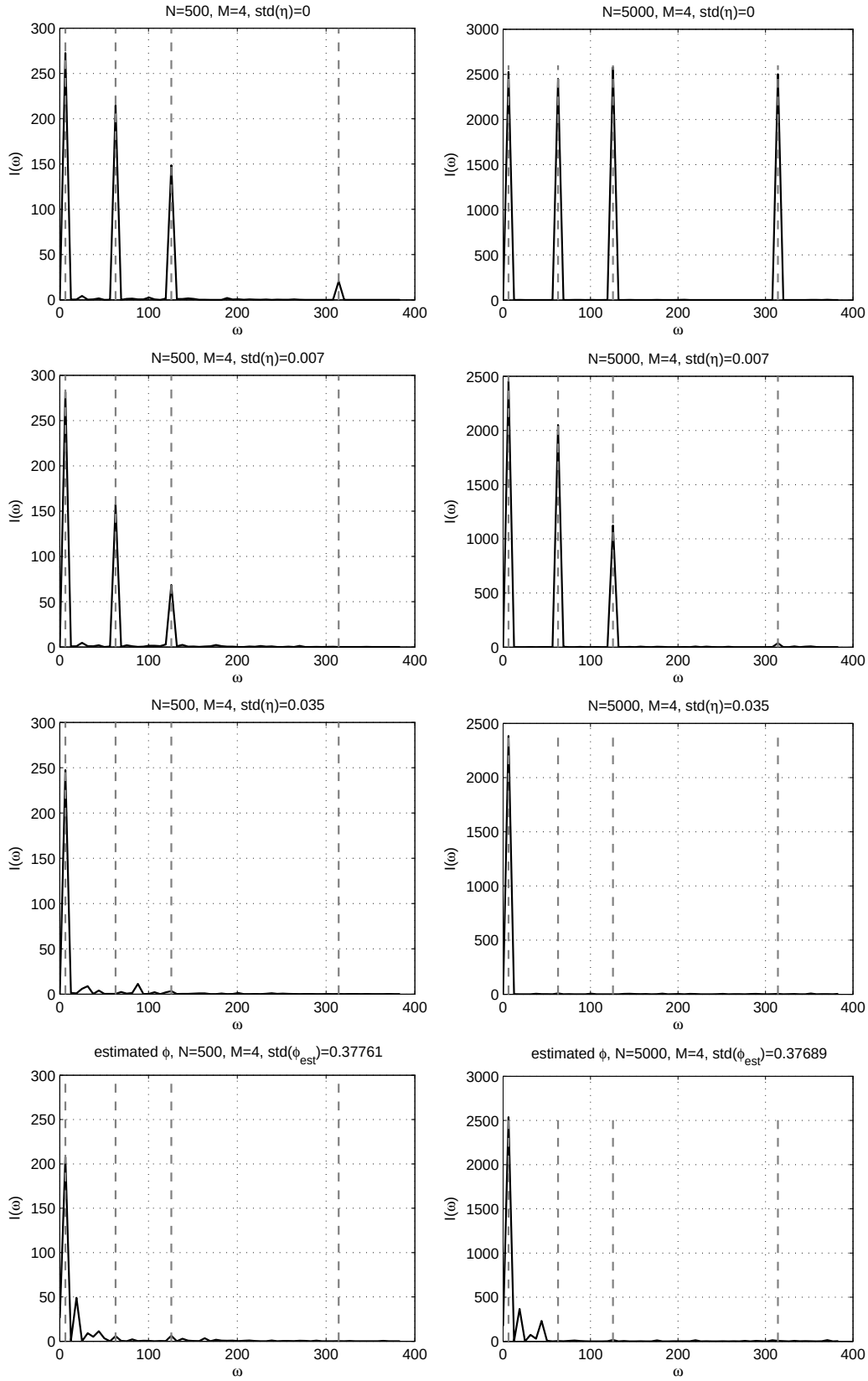

Figure 11: Periodograms of aligned time series: Periodograms for different noise standard deviations σ_η in (3.13) (upper three) and estimated phases (lowest).

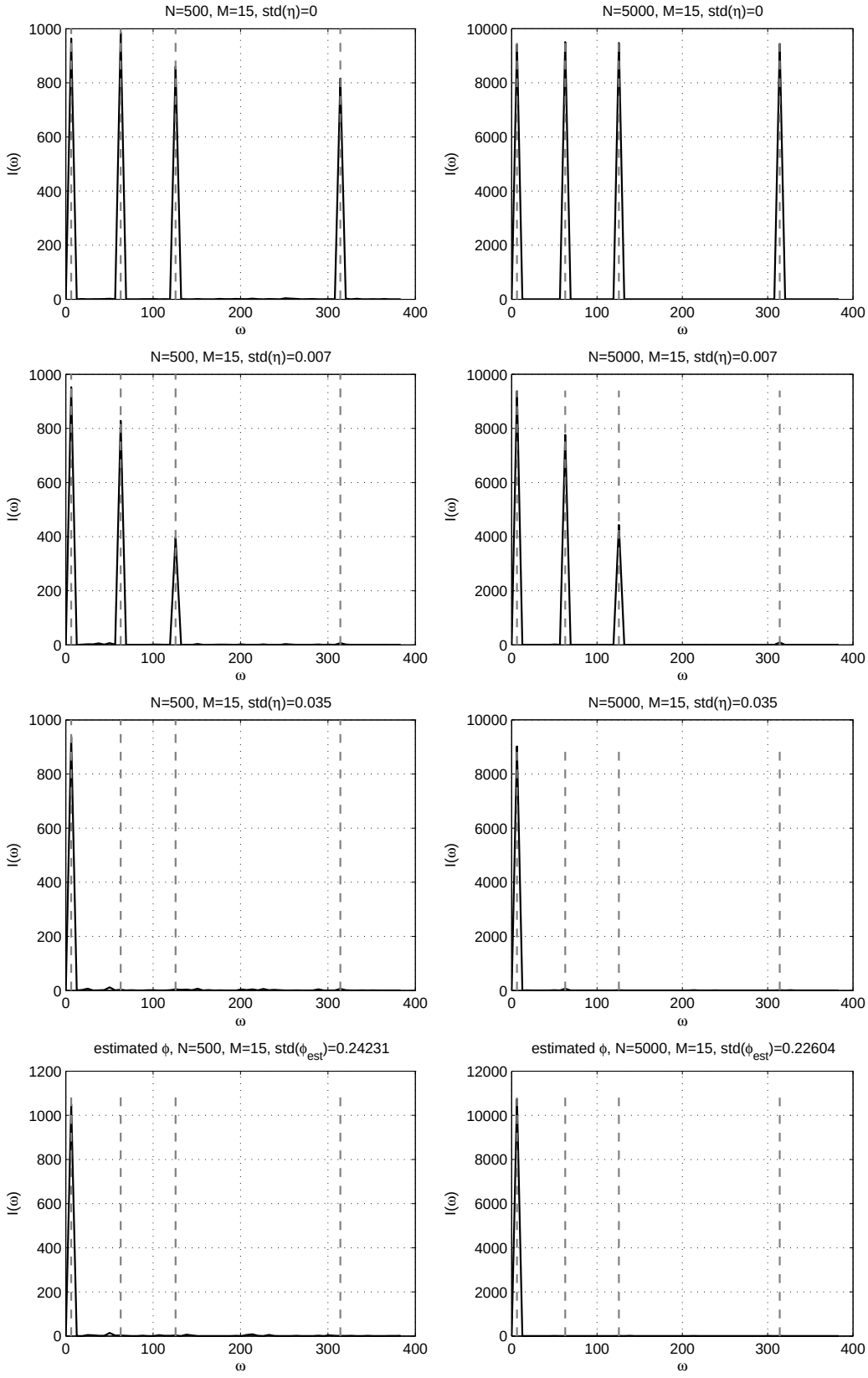

Figure 12: Periodograms of aligned time series: Periodograms for different noise standard deviations σ_η (3.13) (upper three rows) and estimated phases (lowest).