# Dynamic Modelling of Microarray Time Course Data

Javier Nuñez Garcia[†] and Olaf Wolkenhauer[‡*]
Control Systems Centre[†‡] and Dept. Biomolecular Sciences[‡]
UMIST
Manchester M60 1QD, UK
e-mail: javier@csc.umist.ac.uk, o.wolkenhauer@umist.ac.uk
http://www.umist.ac.uk/csc/people/wolkenhauer.htm

February 4, 2001

**Abstract**

The analysis of gene expression profiles, obtained from DNA microarray experiments, is used to discover relationships between genes and to discern groups of genes involved common processes. The principal aim of this paper is to introduce dynamic modelling of microarray time course data. A novel approach to identify similar gene expression profiles is presented. Using parametric modelling, we define a distance between expression profiles that identifies genes with similar dynamic responses opposed to distances among vectors. This approach provides an intuitive interpretation of similarity in the time domain and allows fast clustering using nearest neighbors.

## 1 DNA Microarray Data

For the data considered in this paper we assume that gene expression profiles were obtained from time course DNA microarray experiments. To illustrate the idea and without loss of generality, we use the yeast data set described by Eisen et al. [ESBB98]. The data set is chosen to illustrate the proposed concept and no reference is made to the biology or the experiments which generated the data. The data matrix $\mathbf{X}$ has $i = 1, \ldots, n$ rows representing genes and $j = 1, \ldots, r_1, r_1{+}1, \ldots, r_1{+}r_2, \ldots, \sum_{i=1}^{m-1} r_i{+}1, \ldots, \sum_{i=1}^{m} r_i = r$ columns of samples grouped into $m$ experiments consisting of $r_1$, $r_2$,..., $r_m$ measurements respectively. The data set discussed in [ESBB98] describes $n = 2647$ genes for $m = 9$ experiments. The number of measurements per experiment varies between 4 and 18.

A common approach is to combine the gene expression profiles of a number of experiments into a single row vector. Referring to the distances between row vectors, this makes sense and can be useful for clustering directly on the matrix $\mathbf{X}$. For example, by using the Euclidean distance, the same importance is given to any data point in a row, independently of the experiment to which it belongs. Other weighted distances could in principle be used to give different importance to each experiment. Although creating a larger sample, this approach makes the assumption that a similar response of genes in unrelated experiments provides stronger evidence for common function. Considering time course data, grouping times series of unrelated experiments, does not make any sense since a model for the combined experiments will explain individual characteristics only poorly. For no reason other than to illustrate the basic ideas, we have worked with the time series corresponding to the first experiment, where $n = 2647$ and $m = r_1 = 18$.

Each row vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{ij}, \ldots, x_{i18}]$ represents a particular gene expression profile. Throughout the paper we refer to a gene by its (row) number in $\mathbf{X}$. $x_{ij}$ is the gene expression level at time $j$ of gene $i$. $x_{ij} \in \mathbf{x}_i$ is the normalized $\log_2$ ratio $E_{ij}/R_{ij}$ where $E_{ij}$ is the expression

*Author to whom correspondence should be addressed. The research was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant GR/L95151 ('Diagnostic Signal Analysis') and GR/N21871 ('Genetic Systems').

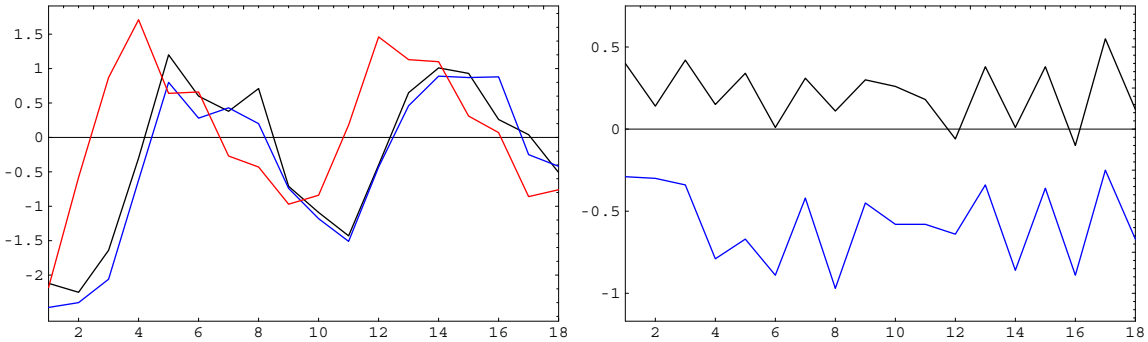


Fig. 1: On the left, three gene expression profiles (1274, 1281 and 1745) with the same dynamic response pattern but shifted in time. On the right, two signals with the same pattern but with difference levels of response (1613, 1384).

level or state at time $j$ of the gene $i$, and $R_{ij}$ is the reference state of the gene, which is a constant value throughout the experiment [BGL+00]:

$$x_{ij} = \frac{\log_2\left(\frac{E_{ij}}{R_{ij}}\right)}{\sqrt{\sum_{k=1}^{18}\left(\log_2\left(\frac{E_{ij}}{R_{ij}}\right)\right)^2}} \quad i = 1, \ldots, 2467 \quad j = 1, \ldots, 18 \tag{1}$$

With the normalization of (1), $x_{ij}$ is positive when $E_{ij} \geq R_{ij}$ and we say the gene is induced or "up-regulated". When $E_{ij} \leq R_{ij}$, $x_{ij}$ is negative and it is said that the gene is repressed or "down-regulated". Here we assume that there are no missing values, that is, measurements have been obtained for all points in the time course or missing values have been replaced by a suitable method.

## 2 Signal Selection and Clustering

Virtually all clustering studies published to this date, using the time course data presented in [ESBB98], have ignored the dynamic component of the microarray experiments. It is the principal aim of this paper to suggest a dynamic systems approach to microarray time course data. Using a distance or the correlation between groups of signals it is possible to identify genes with similar shapes of expression profiles. This approach will however discount signals which are similar although one is delayed with respect to the other or one has a stronger or weaker response, but the same dynamic behavior. Figure 1 illustrates this point.

The method is summarized by the following steps. Each signal is modelled as an individual time series by using a parametric technique such as ARIMA models. Signals with the same model structure such as AR(2) or MA(1) can then be grouped. Inside each group, similar signals have similar parameters and a clustering algorithm can be applied to identify patterns among the parameters in each group. In this paper we do not go as far as clustering by model structures but illustrate the idea by fitting a simple autoregressive model to the data and a then group genes as nearest neighbors in the parameter space of the models.

In microarray experiments studying considering a large number of genes, as in whole genome arrays, we often find that many genes do not show any response leading to noisy signals with no deterministic component. In [NW01], we suggested a number of statistical tests which may be used to select 'informative' signals and thereby to reduce the computational costs of the analysis. This selection, however, does remain subjective and with relatively small sample sizes such tests are inevitably unreliable.

To achieve the objective of the present paper, it is not necessary to select signals but to illustrate and visualize the idea, we wish to 'clear' the rather dense parameter space. For visualization purposes, we therefore discard signals which have no trend nor any dependency among elements of the sample. For this we use the 'Runs Test' [NW01] which assumes that such a signal oscillates above and below the median with equal probability. The second criterion that we apply, is to

define a relatively small threshold or interval around zero for the maximum of the absolute values of the time series. Gene expression profiles with a weak signal, never exceeding the boundaries, are discarded. The threshold used here is $[-0.5, 0.5]$, leading to 1802 signals out of 2467 selected.

# 3 Dynamic Modelling of Microarray Experiments

In this section, each signal selected using the criteria described in the previous section, is described by a parametric time series model. Because of small sample sizes only a simple autoregressive (AR) model structure with a limited number of parameter is considered. An autoregressive model of order $p$, denoted AR($p$), defined by [BJ76]

$$x_t = \theta_1 x_{t-1} + \theta_2 x_{t-2} + ... + \theta_p x_{t-p} + e_t \tag{2}$$

where $e$ is the random variable describing the error or difference between the forecast and the real value. In Box-Jenkins time series analysis the absence of a trend is an important condition to obtain accurate models in forecasting especially for larger samples. Since we do not intend to make forecasts and do not consider asymptotic properties which form the basis of Box-Jenkins models, we do not remove the trend of signals.

One criterion used to examine whether a model is adequate or not, is the Portmanteau test. The Pormanteau statistic $Q_h$ is calculated from the residual $e_t$ produced by the model. It is defined as

$$Q_h = n(n+2) \sum_{i=1}^{h} \hat{\rho}^2(i)/(n-i) \tag{3}$$

where $\hat{\rho}(i)$is the sample correlation function of the residuals. $Q_h$ is approximately chi-squared distributed with $h - p$ degrees of freedom. $h$ is a value smaller than $n$. The Portmanteau test checks whether the hypotheses $H_o$ "The model is adequate for the time series", is verified or not. We accept $H_o$ when $\chi^2_{1-\alpha}(h-p) > Q_h$. The level of confidence of the test $\alpha$ used in this paper is 0.05. We found that from the 1802 signals selected previously, 1617 signals are well modelled by an $AR(1)$ model, 1723 with an $AR(2)$ model, 1780 with an $AR(3)$ model and 1777 with an $AR(4)$ model. For the estimation of model parameters the Yule-Walker equations were used. We refer to [BJ76] for further details regarding AR($p$) time series modelling.

## 3.1 Studying the Parameters of the Models

Using the same model structure, the dynamic pattern of a gene response is represented by its model parameters. The closer these parameters in the parameter space, the similar is the dynamic pattern of the corresponding genes in an experiment. Here we use the Euclidean distance between parameter vectors. In Figure 2 the distributions of parameters for the 1617 AR(1) models are shown. We note that all parameters are inside the $[-1, 1]$ interval which defines stationary first order models. Without signal selection, as discussed previously, the number of models with parameters around zero is significantly increasing. Parameter around zero suggest that the signals in question are uncorrelated, providing evidence that mostly noisy signals were discarded. The lower histogram in Fig. 2 illustrates the large number of noisy signals. While the tails of both distributions are nearly identical in the vicinity of zero the selection process has had its effect.

For example, in Figure 3, the AR(1) model for gene 1613 has a parameter 0.37. Its nearest neighbors in the parameters space show other expression profile 2208, 2225 and 867 with similar frequency and amplitude. Although the signals appear unrelated (and may indeed biologically unrelated or noise), they nevertheless share a similar dynamic response pattern. Following this simple example, we now consider second order AR(2) models for a larger number of 1723 signals. A second order model will allow for an intuitive visualization of signal dynamics. Figure 4, shows the 1723 two-dimensional parameter vectors. The triangular region of the parameter space denotes stationary models. The parabolic curve divides the parameter space into models with real roots (upper region) and complex conjugate roots of the characteristic polynomial given by

$$\varphi(z) = z^2 - \theta_1 z - \theta_2 \ .$$

In Figure 5 on the left, we can see the $8, 12$ and 16 nearest neighbors of gene 1274 (identified by the Euclidean distance in the parametric space). In the same figure, the column on the

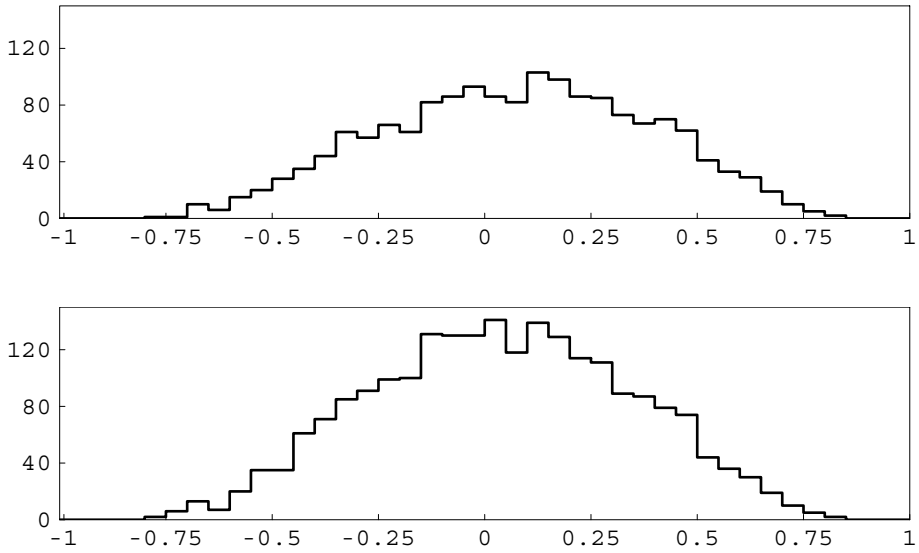


Fig. 2: Histograms of AR(1) model parameters for which residuals passed the Portmanteau test. Top: Signals selected using Runs Test and threshold. Bottom: Without signal selection.

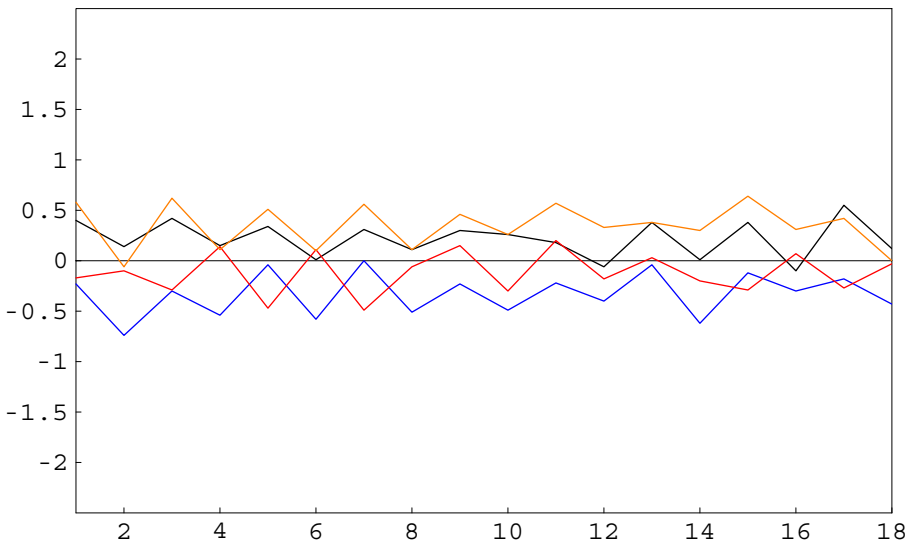


Fig. 3: Four gene expression with similar dynamic responses.

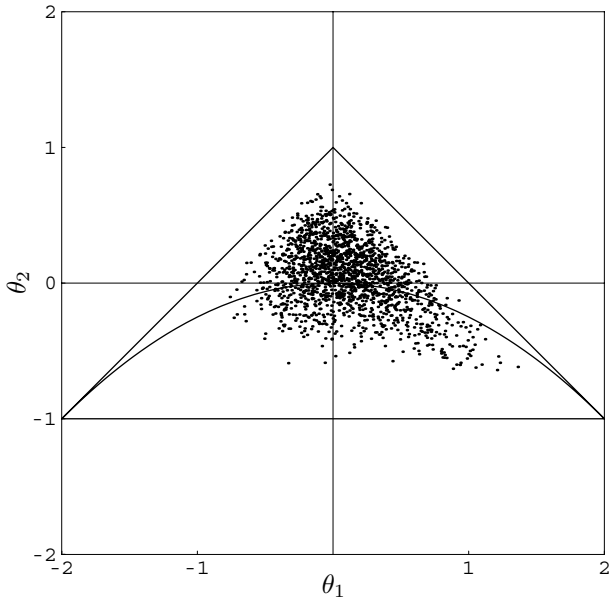


Fig. 4: Two-dimensional parametric space for the $AR(2)$ models

right shows the nearest neighbors of the same gene identified using a correlation coefficient as in [ESBB98]. The closer the correlation to $-1$ or 1, the more similar are two signals.

For eight nearest neighbors both methods identify very similar groups of genes. However, as we increase the number of nearest neighbors, the signals that are added to the clusters using the correlation coefficient are increasingly dissimilar to the rest. In contrast, genes selected in the parameter space maintain a similar dynamic response. Studying gene networks, two genes or groups of genes with similar dynamic pattern but shifted in time may suggest temporal or causal interactions. Such genes or gene clusters could more easily be identified in the parameter space than using for example a correlation coefficient or Euclidean distance in the $r$ dimensional space. In Figure 6 the AR(2) parameter space and the neighborhoods of gene 1274 ($\theta_1 = 0.99$, $\theta_2 = -0.52$) using the two methods are shown. Although both clusters overlap for a number of genes, clustering using the correlation coefficient obviously ignores a number of candidates which have a much more similar dynamic response to the gene in question.

With the proposed approach, the identification of co-expressed and interacting genes can proceed in two ways. Given a gene with known function or involvement in a particular process, other genes with a similar dynamic response are readily identified in the parameter space. If, on the other hand, we are to look for a specific temporal pattern we can search for genes in the corresponding region of the parameter space. In other words, each region of the parameter space corresponds to a known dynamic response, allowing an intuitive interpretation of the similarity of gene expression profiles in the time domain. For example, for the used data set described measurements were obtained from time courses during the cell division cycle after synchronization by alpha factor arrest. For this process a cyclic pattern is expected and related genes will be found in the lower right corner of the parameter space, as shown in Fig. 6. The basic structure of the parameter space, associated dynamic patterns and related alternative representations are discussed in the next section.

## 3.2 Discussion and Results

From previous sections it has become obvious that there exists no unique description of what is meant by the similarity of gene expression profiles. Ideally, clustering algorithms and the distance measure employed should reflect the biologists definition of co-expression. This will depend on the type of experiment and the biological context. With a number of similarity measures available, 'hard' clustering of genes into non-overlapping groups, for example using hierarchical or hard-$c$-means clustering, may lead to misleading results. A 'fuzzy' clustering, either in the data space of the matrix $\mathbf{X}$ or in the parameter space of an autoregressive model, would allow

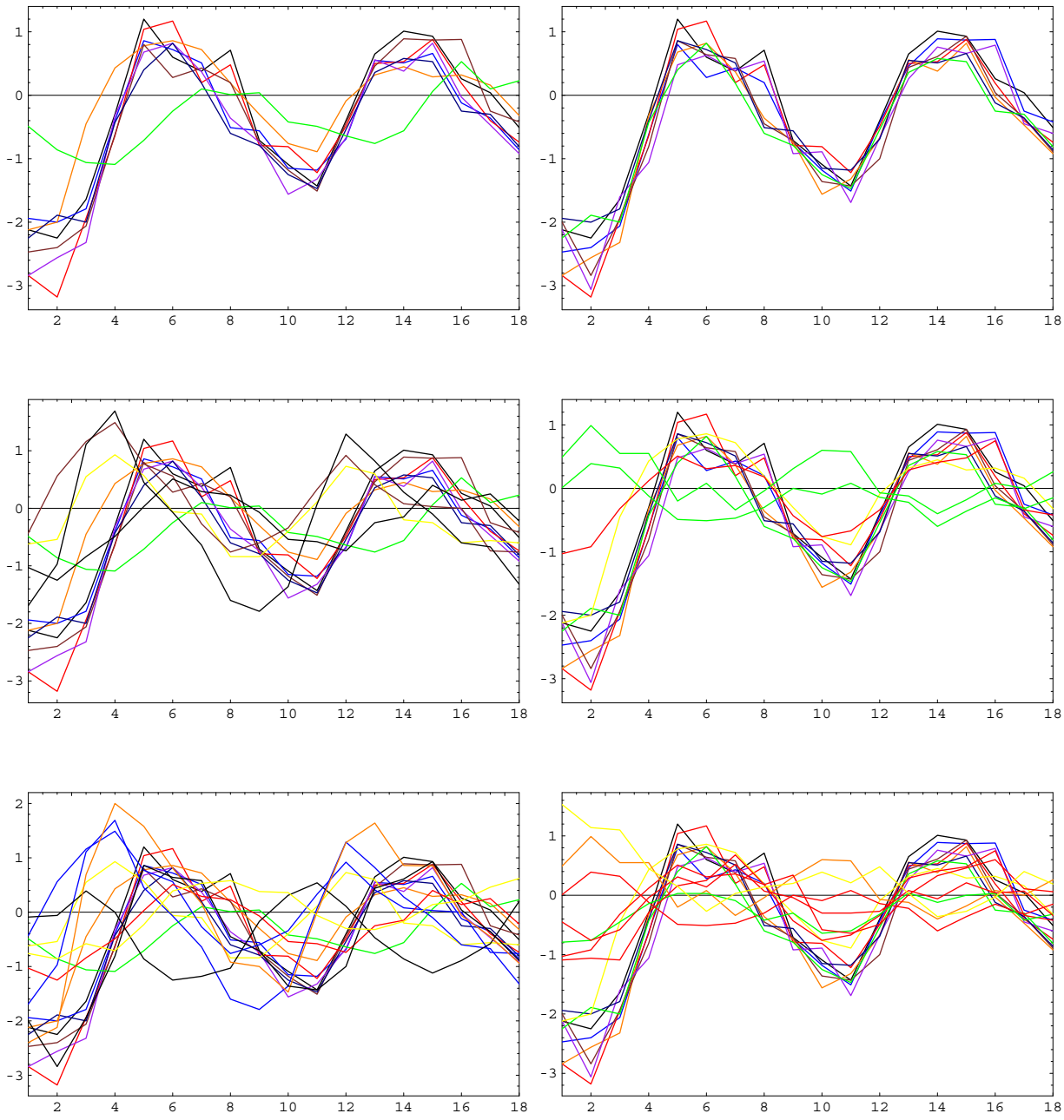

Fig. 5: Left: 8, 12 and 16 nearest neighbors of gene 1274 identified using the Euclidean distance on the parameters of the AR(2) models: 1274, 1278, 1277, 471, 1281, 1276, 1275, 891, 452, 349, 1233, 1749, 180, 1747, 1280 and 2095. For the signals on the right, a correlation coefficient is used to define the similarity. The : 1274, 1281, 1277, 1276, 1280, 1279, 1278, 1275, 471, 111, 1982, 2379, 299, 1347, 1009 and 212.

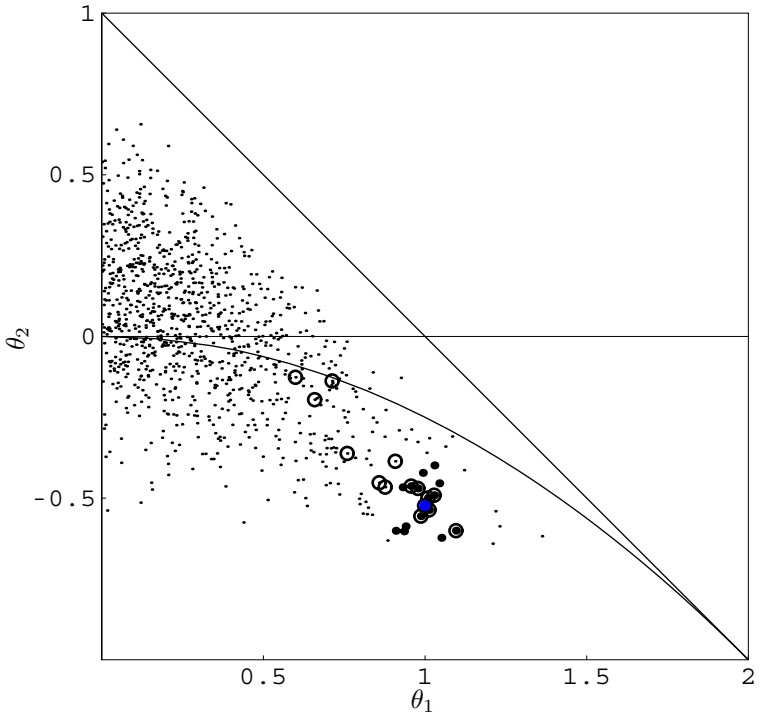


Fig. 6: The AR(2) parameter space, gene 1274 ($\theta_1 = 0.99$, $\theta_2 = -0.52$) and its nearest neighbors. Solid points denote genes clustered in the parameter space while circles denote genes clustered using a correlation coefficient.

for genes belonging to more than one group. Clustering in the parameter space would allow us to define conditions on dynamic patterns in the time domain. On the other hand, using the nearest neighbor approach we obtain a sequence of genes ordered according to the distance used and with respect to a chosen gene. The nearest neighbor approach is particularly suited for an interactive selection of related genes. As the number of nearest neighbors is increased the change of expression profiles can be observed. For example, in Figure 5 on the right, different numbers of nearest neighbors, based on the correlation of gene profiles, are shown. The analyst may then cut off the number of nearest neighbors when (s)he decides that they do not sufficiently follow a cyclical pattern anymore. Note how this method avoid a large quantity of computation time since it is only necessary to calculate a single vector of distances between the original profiles while hierarchical clustering would require the calculation of the complete distance matrix. Modelling microarray time course data in the parameter space adds extra information in that the location of the gene in the parameter space identifies it with a characteristic and known temporal pattern.

In many cases experiments are designed to identify genes related to particular processes. If any gene in such functional class is known, the approach would instantly select genes with similar or related responses rather than clustering the whole genome space which may be time consuming and at present requires a fast computer with a large memory to store the distance matrix as used for example in hierarchical clustering. Even if a reasonably fast computer would be available, repeated experiments with a number of distances would be discouraging. The multiplicity of choices for algorithms and distance measures suggests that one should be able to try a number of alternative techniques in reasonable time. The proposed method would certainly allow such experimentation with a minimal amount of computations.

If for any reason no gene is known, one may also 'design' an expression profile and proceed as described above. If such signal has a cyclical element, we know that it would appear in the lower right area of the triangle of Figure 4. Figure 7 illustrates this region of the parameter space using a gene with parameters $(1, -0.5)$ and its 15 nearest neighbors. Figures 8 and 9 demonstrate how the dynamic pattern changes as we move away from the specified region of Figure 7. Note how the cyclical pattern vanishes as we consider other regions. (Since we have used data describing

the yeast cell cycle, some regions do not describe informative expression profiles.)

Figures 7, 8 and 9 include plots of the roots for the characteristic polynomial in the top-right corner, the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the nearest neighbors. Nearest neighbors in the parameter space follow a specific pattern in the complex plane. This suggests an alternative space for clustering genes. Yet another method is to cluster genes using elements of the ACF or PACF. The dimension of the space in which to cluster could for example be determined by values that reach a specified confidence level (plotted as a dashed line). The autocorrelations chosen can then be compared using a metric. In Figure 10 we calculated the 15 nearest neighbors in the two-dimensional autocorrelation space, i.e., taking the autocorrelations $\rho_1$ and $\rho_2$ of the PACF for gene profile 1274. This approach leads to very similar results in the time domain. This method may be preferred to the parametric one since the estimation of the parameters is not necessarily avoiding the uncertainty that the estimation process includes in the analysis. In the same way we could use the $ACF$ or a combination of both.

# 4 Summary

The principal aim of this paper is to introduce dynamic modelling of microarray time course data. The data set used was chosen to illustrate the underlying idea without referring to the biological context of the data. A novel approach to identify similar gene expression profiles is presented. Using parametric modelling, we define a distance between expression profiles that identifies genes with similar dynamic responses opposed to distances among vectors. This approach provides an intuitive interpretation of similarity in the time domain and allows fast clustering using nearest neighbors. Although small sample sizes remain a major challenge in the analysis of microarray time course experiments, the theory of time series analysis is well established with a vast collection of reference literature available. The algorithms can easily be implemented but are also readily available in most software packages. The proposed technique is suited for an interactive study of genes with similar dynamic responses as for example required in the study of interactions in gene networks.

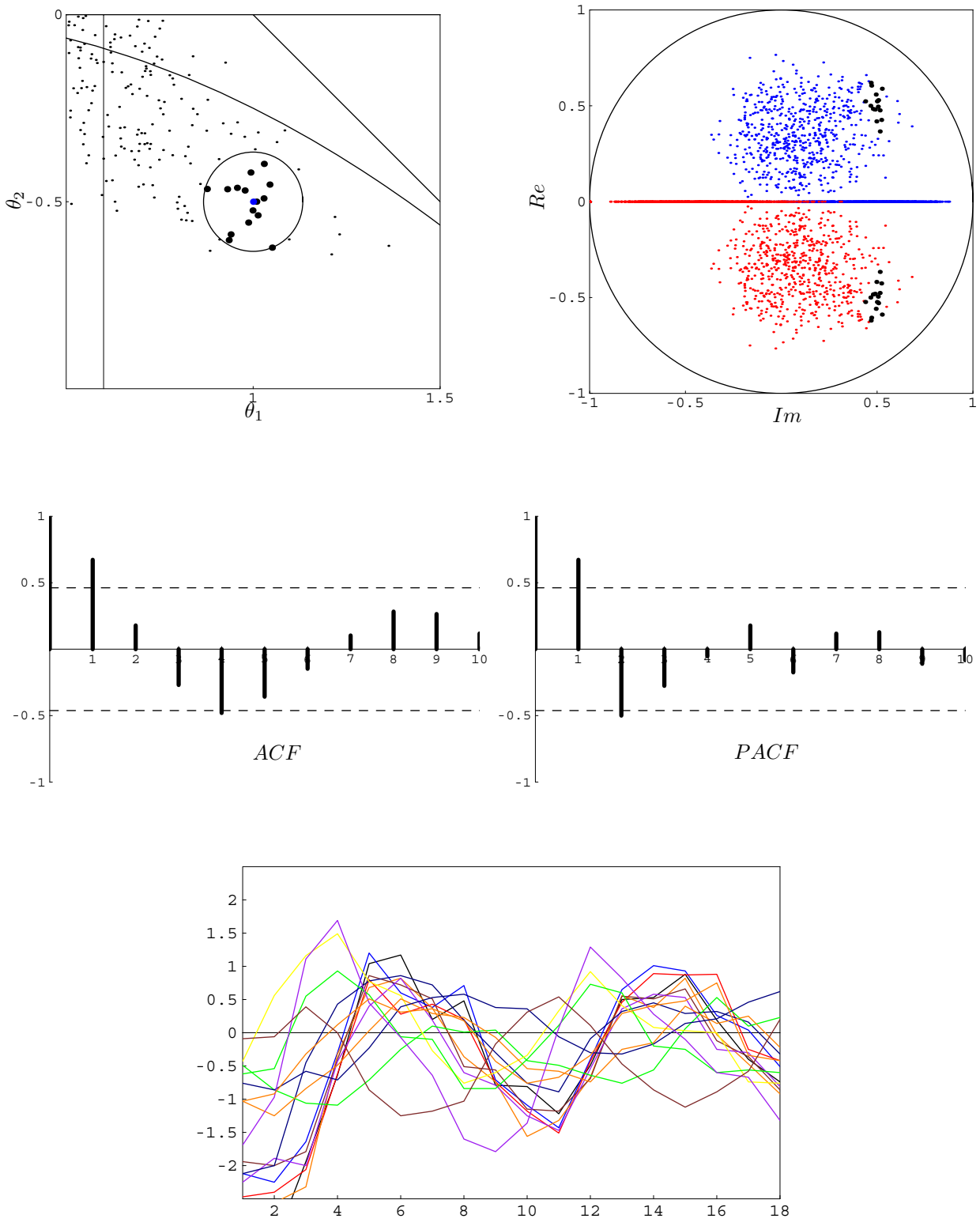


Fig. 7: The point $(1, -0.5)$ in the parametric space with its 15 nearest neighbors delimited by a circle; The roots of the characteristic polynomial of the models; the autocorrelation (ACF) and partial autocorrelation functions (PACF); the 15 time series corresponding to the nearest neighbors.

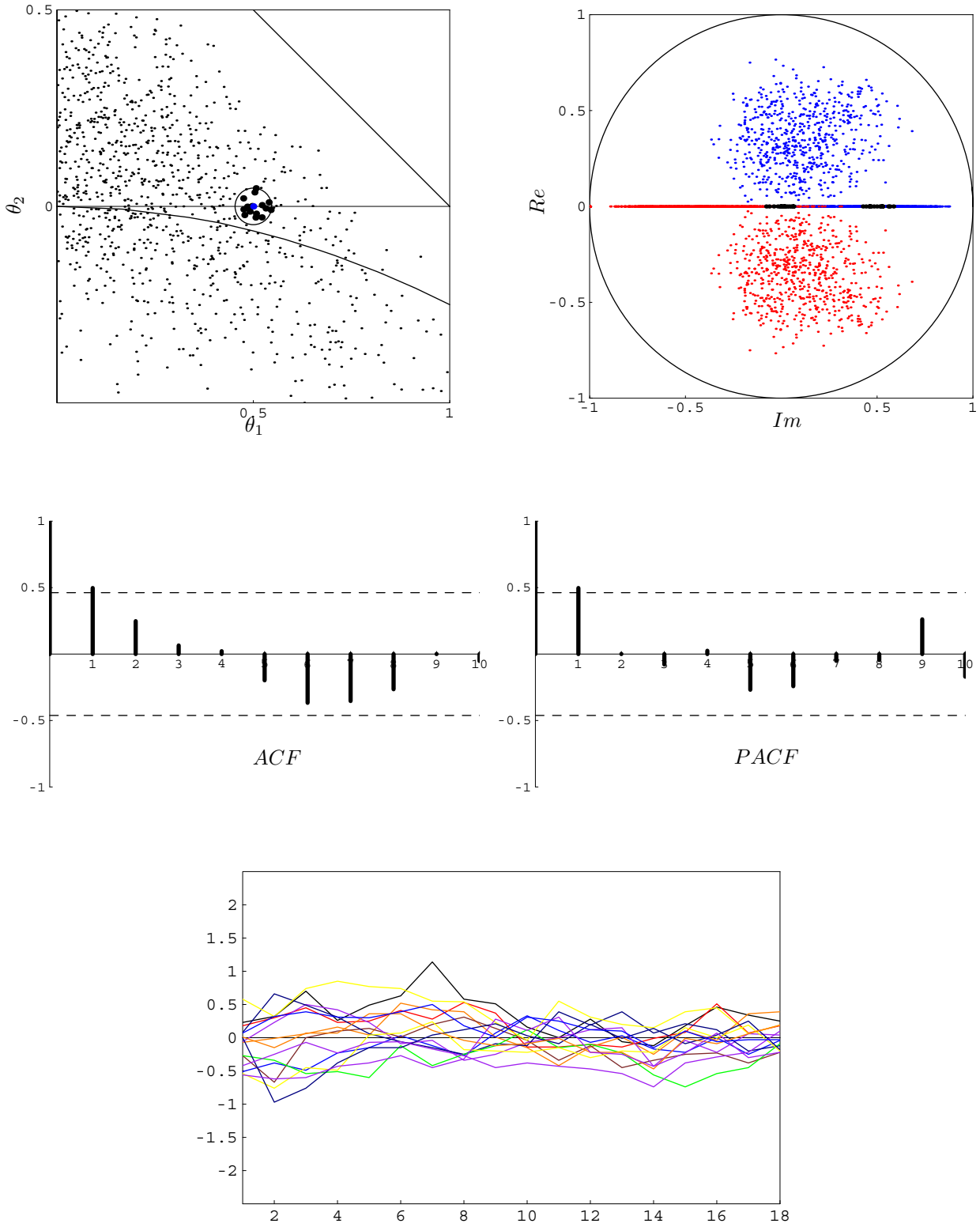


Fig. 8: The point $(0.5, 0)$ in the parametric space with its 15 nearest neighbors delimited by a circle; The roots of the characteristic polynomial of the models; the autocorrelation (ACF) and partial autocorrelation functions (PACF); the 15 time series corresponding to the nearest neighbors.

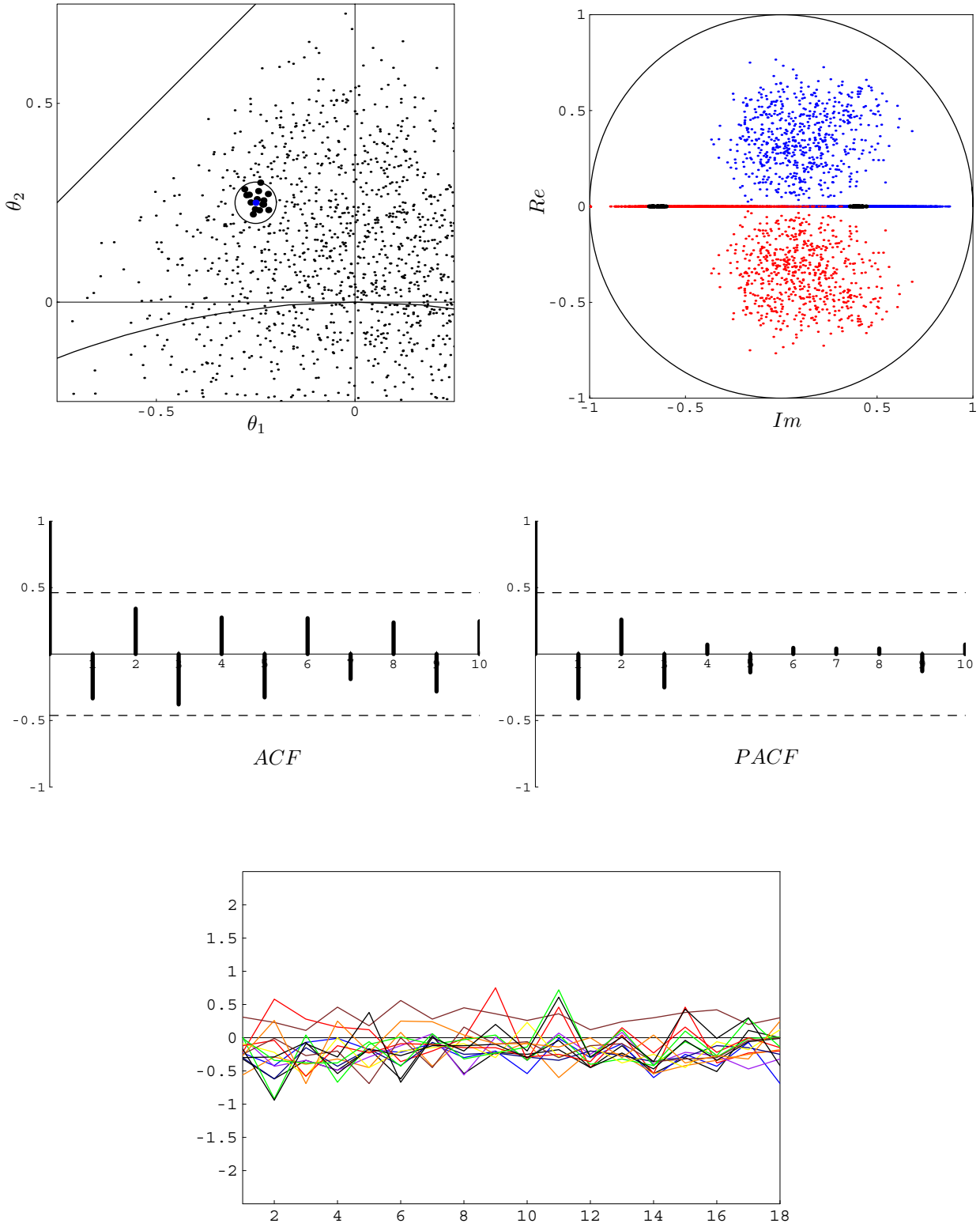


Fig. 9: The point $(-0.25, 0.25)$ in the parametric space with its 15 nearest neighbors delimited by a circle; The roots of the characteristic polynomial of the models; the autocorrelation (ACF) and partial autocorrelation functions (PACF); the 15 time series corresponding to the nearest neighbors.

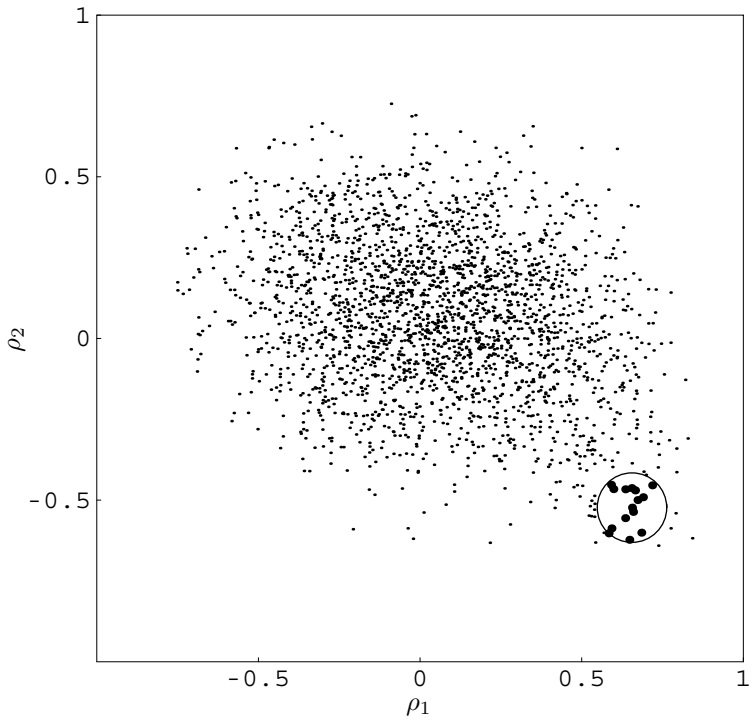


Fig. 10: Nearest neighbors in the two-dimensional autocorrelation space for gene profile 1274: 1274, 1278, 1277, 471, 1281, 1276, 1275, 349, 111, 1280, 452, 891, 1982, 180 and 1749.

# References

[BGL+00] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S.Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA (PNAS)*, pages 262–267, 2000.

[BJ76] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden Day, 1976.

[ESBB98] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA (PNAS)*, pages 14863–14868, 1998.

[NW01] J. Nunez and O. Wolkenhauer. Signal selection in microarray data analysis. Submitted for publication, January 2001.