Dynamic metabolomic data analysis: a tutorial review

Smilde, A.K.; Westerhuis, J.A.; Hoefsloot, H.C.J.; Bijlsma, S.; Rubingh, C.M.; Vis, D.J.; Jellema, R.H.; Pijl, H.; Roelfsema, F.; van der Greef, J.

DOI
10.1007/s11306-009-0191-1

Publication date
2010

Document Version
Final published version

Published in
Metabolomics

Metabolomics (2010) 6:3–17
DOI 10.1007/s11306-009-0191-1

REVIEW ARTICLE

Dynamic metabolomic data analysis: a tutorial review

A. K. Smilde · J. A. Westerhuis · H. C. J. Hoefsloot · S. Bijlsma · C. M. Rubingh · D. J. Vis · R. H. Jellema · H. Pijl · F. Roelfsema · J. van der Greef

Received: 28 January 2009 / Accepted: 9 November 2009 / Published online: 4 December 2009

Abstract In metabolomics, time-resolved, dynamic or temporal data is more and more collected. The number of methods to analyze such data, however, is very limited and in most cases the dynamic nature of the data is not even taken into account. This paper reviews current methods in use for analyzing dynamic metabolomic data. Moreover, some methods from other fields of science that may be of use to analyze such dynamic metabolomics data are described in some detail. The methods are put in a general framework after providing a formal definition on what constitutes a 'dynamic' method. Some of the methods are illustrated with real-life metabolomics examples.

Keyword Time-resolved · Dynamic systems · Multivariate modeling · Dimension reduction · Time series analysis · Basis functions

1 Introduction

In the last decade exciting science and innovation in Life Sciences are driven from a systems view. Systems biology has found a scientific focus due to advances in the high throughput enabling technologies, measuring quickly at different biological levels such as transcripts, proteins and metabolites (Hood 2003; van der Greef et al. 2007; Wolkenhauer 2002). The progress of the systems based approach is in a large part depending on developments in biostatistics and bioinformatics to integrate high-dimensional data to obtain a systems view. Many challenges remain in this area some of which will be discussed.

A systems view recognizes that at different levels of a complex system new properties are emerging and as a consequence we need to study a system as a whole and not by focusing on elements only. In addition to this, the multi-level, interconnected, non-linear and dynamic properties become the focus from which the self-organization of a system can be understood. The dynamic characteristics both from a measurement and biostatistics point of view are becoming mandatory to reveal new system information, e.g., to understand homeostasis and resilience after perturbation. Understanding health and disease based on concepts as resilience can be understood from a biological view, but the ability to measure and to analyze the complex longitudinal high-dimensional data is mandatory to make progress in research.

The concept of dynamic diseases was coined by Glass and Mackey (1988) and the importance of biorhythms in relation to health and disease as well as intervention are surfacing as topics in multi-factorial disease etiology. From a measurement point of view, metabolomics is an attractive tool as it reveals information close to the phenotypic level and it allows for large scale measurements in a robust way.

To set the stage, we first describe which type of dynamic metabolomic data will be the topic of this paper. We will not discuss approaches in metabolic flux analysis, since that topic is covered elsewhere (Kholodenko 2004;

A. K. Smilde (✉) · J. A. Westerhuis · H. C. J. Hoefsloot · D. J. Vis
Biosystems Data Analysis, Swammerdam Institute for LifeSciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
e-mail: a.k.smilde@uva.nl

S. Bijlsma · C. M. Rubingh · R. H. Jellema · J. van der Greef
TNO Quality of Life, Utrechtseweg 48, 3704 HE Zeist, TheNetherlands

H. Pijl · F. Roelfsema
Department of Endocrinology and Metabolic Diseases, Leiden University Medical Center, Leiden, The Netherlands

Stephanopoulos et al. 1998) and occurs at a different time scale involving other mechanisms than the data we want to discuss. Also metabolic network inference methods through dynamic data will not be discussed because this is a topic in itself (Samoilov et al. (2001); van Berlo et al. (2003)). The dynamic data we are going to discuss can originate from different sources, depending on the relevant biological question and the study design. In human metabolomics, dynamic data from a challenge test may be available: a person receives a challenge (e.g., a test meal) and blood is sampled afterwards (Bijlsma et al. 2006) (time scale of minutes/hours). This points to the topic of personalized food and medicine where each person is subjected to a challenge test that serves as a blueprint of the 'metabolic status' of that person (van der Greef et al. 2006). In animal studies, also serial tissue sampling (besides bodily fluids) might be available (Kleemann et al. 2007) (time scale of hours/days). Another example in animal studies is toxicology, where a toxic compound is administered at different dosage levels and samples of urine, blood and liver are collected in time (Heijne et al. 2005; Keun et al. 2004) (time scale of hours/days). A completely different type of dynamics is in microbial metabolomics, where time-resolved measurements are done on the intracellular metabolome of an organism in a fermentation process (Rubingh et al. 2009) (time scale of hours).

In this paper, we will focus on metabolomics data (mostly) obtained through instruments such as NMR, LC-MS and GC-MS. We will discuss methods to understand underlying dynamic behavior of biological systems based on analyzing metabolomics data. We will give an overview on approaches taken in other fields such as chemical engineering, systems theory and psychometrics. Since these fields are very large, only those approaches are discussed which are of potential use in metabolomics. The existing approaches in transcriptomics will also be discusssed in this framework. We will discuss the methods in the context of multivariate and high-dimensional data. Multivariate means that for a single sample, multiple metabolites are measured and in high-dimensional data, many more metabolites are measured than samples. For metabolomics, hardly any methods incorporating dynamics exist sofar, to our best knowledge. We will specify this statement later.

Three real-life data sets will be used as examples throughout this paper to illustrate the working of some of the methods. A brief introduction to the specifics of the examples; building blocks of dynamic metabolomic data analysis and a definition of 'dynamics methods' are also provided to set the stage. Many of the methods reviewed and proposed are not yet used in metabolomics, hence, the area of dynamic metabolomics data analysis is still very open for research.

2 Short description of the examples

Hormones are signaling and regulatory molecules. In humans many hormones exhibit a circadian rhythm. There are indications that the dynamic behavior of hormones are related to disease states and also change upon treatment (Kok et al. 2004, 2006). Hormones are secreted in pulses, delivered to the bloodstream and subsequently degraded. In the example, women were hospitalized for a study and during a 24 h period, blood samples were taken every 10 min ($n = 145$ per individual). These blood samples were analyzed for certain hormones, among them cortisol and luteinizing hormone (LH). These measured hormone levels are shown in Fig. 1 for one female. The data show clearly pulsatile patterns.

The second example concerns NMR spectra of urine of rhesus monkeys (*Macaca mulatta*) measured in time. Samples are taken of ten monkeys (five male and five female) at $n = 29$ days unevenly spread over a time course of 57 days. This is a normality study: the monkeys were kept in a non-stressed environment to study their natural biorhythms. Details of the study were published elsewhere (Jansen et al. 2004).

The third example is on nutrikinetics: the kinetic fate of nutritional compounds (van Velzen et al. 2009). In a randomized, placebo controlled double blind cross-over study, 20 healthy volunteers were subjected to a tea treatment. NMR measurements were performed of their urine which was collected over time (n varies between 9 and 14). This allowed for estimating kinetic parameters for metabolites in their urine.

3 Short description of methods

The potential viable methods are categorized in six groups of methods where each group shares a similar underlying idea. There is a loose ordering of the categories by the amount of *a priori* knowledge needed to perform the data analysis or, stated differently, by the strictness of the imposed assumptions.

The first group consists of methods that are based on *fundamental models*. This means that *a priori* knowledge should be available about the functional form of the dynamics for the metabolites. The second group consists of methods based on predefined *basis functions*; such as wavelets. Hence, some form of the dynamics must be reasonable given the underlying biology. The third group discusses *dimension reduction* methods, such as principal component analysis. These methods work if there is an underlying low dimensionality in the metabolomics data. Group four discusses *multivariate time series* models, which can be used if certain stationarity assumptions hold.

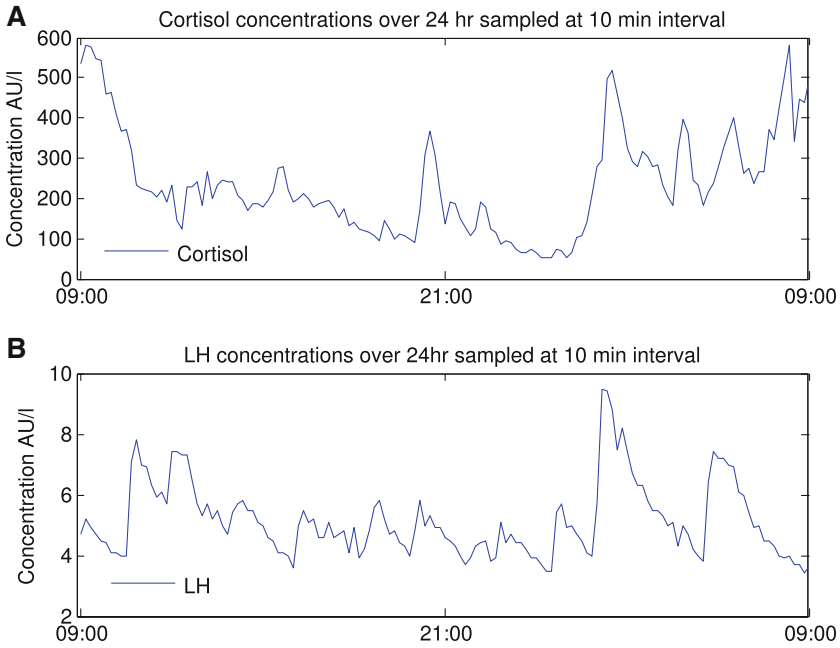

Fig. 1 Measured cortisol (**a**) and luteinizing hormone (**b**) levels in a woman during 1 day showing pulsatility and biorhythms

Group five deals with analysis-of-variance (*ANOVA*) type models and finally, the sixth group discusses methods imposing *smoothness*, using the intrinsic consecutiveness of time-resolved measurements.

When selecting a specific method for modeling dynamic metabolomics data, it is useful to think in terms of the 'data generating process'. A possible data generating mechanism is when the biological system under study is perturbed thereby inducing changes in unobservable biological processes. These in turn then affect the manifest variables, which are the measured metabolites. Variations on this theme are possible; this particular example is the idea behind data generating processes for which dimension reduction methods are suitable. Postulating a specific data generating process presupposes knowledge about the biological system under study and the way the experimental design has perturbed that system. Ideally, the knowledge about the form of the dynamic system behavior as made explicit in the data generation process is matched to the requirements for the data analysis method. We will therefore make assumptions of the methods as explicit as possible.

4 Building blocks

4.1 Fundamental models

Fundamental models of biological processes are usually put in the framework of differential or difference equations. These will therefore be discusses in some detail. Good introductory textbooks exists for both linear (Fortmann and Hitz 1977) and nonlinear dynamics (Strogatz 1994). The general form of a first-order differential equation for $x(t)$ is

$$\dot{x} = f(x(t), t; \alpha), \tag{1}$$

where $f(x(t), t;\alpha)$ is a (possible nonlinear) function of $x(t)$ and t and only the first order derivative of $x(t)$ is present; the function f contains (possibly unknown) parameters α. An example of a simple DE is

$$\dot{x} = \alpha x, \tag{2}$$

which is a first order (only first order derivatives), linear (only linear terms in x), autonomous (time appears only implicitly through $x(t)$) and homogenous (no forcing functions or inputs) differential equation and the solution of this equation is $x = ae^{\alpha t}$ for the initial condition $x(0) = a$. Depending on the value of α, Eq. 1 has a stable solution (a decreasing e-power for $\alpha < 0$) or an unstable solution (an increasing e-power for $\alpha > 0$); for $\alpha = 0$, $\dot{x} = 0$, and $x(t) = x(0)$ for all t. If $x(0) = 0$, the derivative $\dot{x} = 0$ and the solution is then $x(t) = 0$ for all t and this solution is indicated with x^* (a fixed-point of Eq. 1). These solutions show the typical behavior of linear first order differentail equations: constant, blow-up or decaying towards zero. Hence, no oscillations can take place.

A second order linear differential equation may look like

$$\ddot{x} = \alpha_1 x + \alpha_2 \dot{x}, \tag{3}$$

which can be rewritten as

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= \alpha_1 x + \alpha_2 y, \end{aligned} \tag{4}$$

or, using obvious matrix notation

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \alpha_1 & \alpha_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \end{pmatrix}. \quad (5)$$

Hence, higher-order linear differential equations can always be transformed to first-order systems. The solutions of a second-order differential equation are richer in behavior, e.g, oscillations are possible (Strogatz 1994). These solutions are characterized by the eigenvectors and eigenvalues of the system matrix **A**: there are oscillations if the imaginary parts of the eigenvalues differ from zero.

Another way of expressing dynamics is in the form of difference equations using discrete time points. Such an equation can look like

$$x_{t+1} = g(x_t, t; \alpha), \quad (6)$$

or, in a simple example

$$x_{t+1} = \alpha x_t. \quad (7)$$

There are relationships between Eqs. 1 and 6 (Fortmann and Hitz 1977) but that is beyond the scope of this paper. In the sequel, continuous time functions are denoted as $x(t)$ and discrete time functions as x_t.

Examples of using fundamental models and differential equations will be given in Sect. 7.1 using both the hormones and nutrikinetics data.

4.2 Time series models

A time series of a single metabolite can be approximated with time series models such as an autoregressive process of order 1 (AR(1))

$$x_{t+1} = \theta x_t + \epsilon_{t+1}, \quad (8)$$

where θ is the parameter to estimate and ε_t is a so called random shock. The parameter θ has to obey a regularity condition ($|\theta| < 1$ for the AR(1) process) to be meaningful. Alternatively, moving average (MA) processes can be used

$$x_{t+1} = \epsilon_{t+1} - \phi\epsilon_t, \quad (9)$$

or combinations of both (ARMA processes),

$$x_{t+1} = \theta x_t + \epsilon_{t+1} - \phi\epsilon_t, \quad (10)$$

which is an ARMA (1,1) model. These are also available for higher orders and in nonlinear ways. It is important to realize that Eqs. 8 and 9 make assumptions about the time series x_t. They both assume stationarity: the mean and standard deviation do not depend on t and the autocovariance only depends on τ (the lag time is defined as a time interval $\tau = t_2 - t_1$, where t_1 and t_2 are two time points of the process). The autocovariance of an AR(1) model decays exponentially as a function of the lag time and for an MA(1) model the autocovariance is zero for lag $\tau > 1$. Hence, such models are not suitable for modeling periodicity and oscillations since these would require

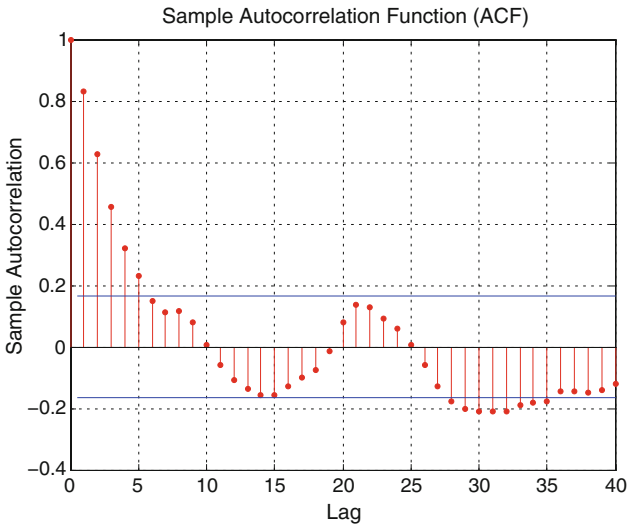

Fig. 2 An example of an autocorrelation function of the LH data showing periodicity

autocovariances with periodic lags. Although autocovariances or autocorrelations strictly speaking can only be used for stationary time series, they can also convey information from general time series. Figure 2 shows the autocorrelation function of the LH-hormone data of Fig. 1. This autocorrelation function shows clearly periodicity which relates to the periodicity in the original signals.

Second-order time series models look like

$$x_{t+1} = \theta_1 x_t + \theta_2 x_{t-1} + \epsilon_{t+1}, \quad (11)$$

and such models are capable of describing damped oscillations. They are more versatile in describing dynamics with periodic events. A more versatile class of times series models are ARIMA models, where the capital 'I' stands for integrating. Such models are also able to describe non-stationary behavior. There is a host of literature on how to estimate parameters in AR, MA and ARIMA models, see e.g., Box et al. (1994).

4.3 Correlations

A key feature of multivariate measurements is the covariation of the individual variables, usually measured in terms of covariance or correlation. Covariance is a measure of association of two random variables and appears as a set of parameters in the multivariate distribution function of the two random variables. For a bivariate normal distribution, this comes down to

$$\mathbf{x} \sim N(\mu, \Sigma) \quad (12)$$

with

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}, \quad (13)$$

where σ_{ij} is the covariance (or variance if $i = j$) of variables x_i and x_j. In the context of dynamic metabolomic data, consider the time sequences x_t and y_t. If both metabolites are driven by, or probing, the same underlying biological process, then they will show similar behavior. Although this similarity can be described by a covariance measure between x_t and y_t, this is, strictly speaking, not a covariance (see, e.g., Anderson 2003). The correct way of describing their mutual behavior is by writing

$$\begin{aligned} x_t &= \gamma_x \xi_t + v_{x,t} \\ y_t &= \gamma_y \xi_t + v_{y,t}, \end{aligned} \tag{14}$$

where ξ_t represents the underlying dynamic process, γ_x, γ_y are parameters and $v_{x,t}$, $v_{y,t}$ are disturbances. Depending on the variances of these disturbances relative to the variation in ξ_t and the sizes of γ_x, γ_y, the time series x_t and y_t show similar behavior. From now on, we will use the concepts of covariance and correlation for the association between x_t and y_t, although this is a simplification.

4.4 Dimension reduction

When measuring many metabolites, a way to bringing down the complexity of the data is using (linear) dimension reduction of which there are essentially two classes of methods: (common) factor analysis and principal component analysis. The factor analysis model for the vector $\mathbf{x}$ ($J \times 1$) containing the measured metabolites can be written as

$$\mathbf{x} = \mathbf{\Lambda y} + \epsilon + \boldsymbol{\mu}, \tag{15}$$

where $\mathbf{\Lambda}(J \times R)$ is a matrix of constants (loadings); $\mathbf{y}$ ($R \times 1$) and $\varepsilon(J \times 1)$ are random vectors. The elements of $\mathbf{y}$ are called common factors and the elements of ε specific or unique factors. The vector $\boldsymbol{\mu}$ is a vector of means of $\mathbf{x}$. This model is a direct extension of model (14). Upon making assumptions regarding distributions and independence of terms, the parameters of model (15) can be estimated (Mardia et al. 1979). In words, the factor analysis model tries to model the covariance structure of the variables in $\mathbf{x}$ by using common factors.

Principal component analysis (PCA) can be interpreted in different ways: a transformation of the original variables or as a subspace approximation method (see, Smilde et al. (2004) for an extensive discussion). The transformation comes down to $\mathbf{z}' = \mathbf{x}'\mathbf{P}$ where $\mathbf{z}$ is an ($R \times 1$) vector of scores and $\mathbf{P}$ a ($J \times R$) matrix of loadings. This equation is invertible for $J = R$ resulting in $\mathbf{x}' = \mathbf{z}'\mathbf{P}'$ or $\mathbf{x} = \mathbf{Pz}$, and upon deciding on the value of R (usually smaller than J,) $\mathbf{x} = \mathbf{Pz}$ becomes an approximation of $\mathbf{x}$. This is usually expressed in the equation

$$\mathbf{X} = \mathbf{ZP}' + \mathbf{E}, \tag{16}$$

where $\mathbf{X}$ is the $T \times J$ matrix containing the measured time series; $\mathbf{Z}$ ($T \times R$) contains the time series component scores; $\mathbf{P}$ ($J \times R$) is the loading matrix and $\mathbf{E}$ ($T \times J$) the matrix of residuals. The loading matrix $\mathbf{P}$ maximizes the variance of the scores and minimizes the sum of squared residuals. Hence, PCA focusses on the variance of $\mathbf{x}$.

Both PCA and factor analysis models reduce the dimensionality of the original problem (J) to R, where R is usually much smaller than J. There are differences between the models (Mardia et al. 1979; Jolliffe 1986), e.g, PCA does not provide facilities for the unique factors and the factors $\mathbf{y}$ are no linear combinations of the $\mathbf{x}$-variables (note that the $\mathbf{z}$ variables are indeed linear combinations of the variables in $\mathbf{X}$). If the unique factor contributions are small relative to the common factor contributions, PCA and factor analysis give similar results.

The factors $\mathbf{y}$ and scores $\mathbf{z}$ are sometimes called *latent* variables to distinguish them from the the *manifest* variables $\mathbf{x}$. Although this nomenclature is somewhat sloppy in the case of PCA, the term nicely illustrates the basic assumption underlying the PCA and factor analysis models: the variation in $\mathbf{x}$ is summarized by a small set of underlying and unobservable variables.

4.5 What are dynamic methods?

It is useful to give a precise definition of a *dynamic* metabolomics data analysis method. This will be done with the example of PCA which is not a dynamic method in our definition.

Suppose a metabolomics data set is available where ten metabolite concentrations are measured at five time points. The resulting matrix has five rows and ten columns representing the measured metabolite values and can be decomposed using PCA. For simplicity, only the first principal component is considered. The PCA results in score vector $\mathbf{z}_{orig}$ and loading vector $\mathbf{p}_{orig}$. Next, the original data are shuffled, such that the time evolution between the rows is broken. The subsequent PCA of the data the gives scores $\mathbf{z}_{shuffled}$ and loadings $\mathbf{p}_{shuffled}$. After this PCA, the $\mathbf{z}_{shuffled}$ can be reshuffled thereby undoing the initial shuffling. The resulting scores $\mathbf{z}_{reshuffled}$ are exactly equal to the original scores $\mathbf{z}_{orig}$, and it also holds that $\mathbf{p}_{orig} = \mathbf{p}_{shuffled} = \mathbf{p}_{reshuffled}$. Hence, PCA is insensitive to the evolutionary nature of the time axis and is thus not a dynamic metabolomics method. The definition of a dynamic metabolomic data analysis method is now simple: that method should be sensitive to the evolutionary nature of the time axis.

5 Dynamic metabolomic data analysis

5.1 Fundamental models

When a time series for a single metabolite is measured (denoted as x_t) and the form of the difference equation is known, then finding dynamics comes down to estimating unknown parameters α in the difference equation

$$x_{t+1} = f(x_t, \alpha), \tag{17}$$

where an autonomous system is assumed (no explicit t in (17)). Several methods exists for estimating the parameters α. One of those methods is using least squares (or nonlinear least squares), however, such problems can be very complicated in terms of irregular error surfaces and very correlated parameter estimates. Moreover, they have the risk of getting stuck in local minima. This can be avoided to some extent by using natural computational methods such as genetic algorithms or simulated annealing (Apostu and Mackey 2008). A viable alternative is to use smoothness constraints for regularization, thereby making the error surface less rugged and the problem better solvable (Ramsay et al. 2007).

An example of using a difference equation in practice are hormone dynamics. A simple model describing measured dynamic hormone behavior in human blood is

$$x_t = x_{t-1}e^{-k} + \phi_t + \epsilon_t, \tag{18}$$

where t is the index for time points, the parameter k is the first-order decay constant, ϕ_t is the pulsatility function and ε_t is the measurement error. The pulsatility function ϕ_t is nonlinear and represents the secreted hormone. This function can be defined in different ways (Vis et al. 2009). Hence, Eq. 18 is an example of a nonlinear non-autonomous difference equation.

The measured hormone levels are shown in Fig. 3 as dots in the upper panel. The pulsatility function was constrained to have only a limited number of pulses (bars in middle panel). The decay is clearly visible in the slopes of the drawn line (upper panel) and the residual ε is presented in the lower panel. The model fits the data well, and gives a decay rate and information about pulsatile behavior important for endocrinologists to study normal physiology and pathophysiology (including diseases).

The idea of using difference equations can be generalized to multiple metabolite measurements. When measurements are available on J metabolites as a function of time, then these can be symbolized as $\mathbf{x}_t$ $(J \times 1)$. A second-order nonlinear difference equation for such a vector is then

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t, \mathbf{x}_{t-1}; \alpha), \tag{19}$$

where an autonomous system is assumed, $F : R^{2J} \rightarrow R^J$ is a vector valued function and α is a set of parameters. The underlying biology or physiology dictates the specific form of the function F and it comes down to estimating parameters α. In principle, the same methods can be used as in the univariate case, but the multivariate problem is usually much harder to solve. An example for a system of two genes can be found in the gene-expression literature (Cao and Zhao 2008).

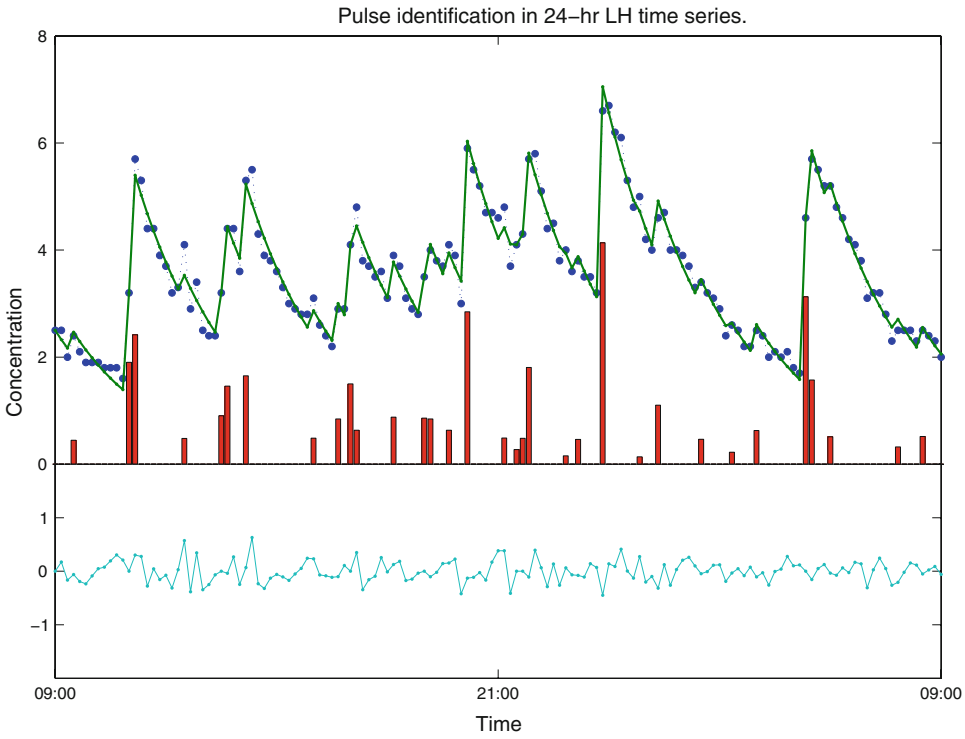

Fig. 3 Measured and fitted luteinizing hormone (*LH*) during 1 day. Legend: *dots upper panel* are original data; *drawn line upper panel* is fitted model; *bars middle panel* are estimated hormone pulses; line lower panel are residuals

For a large number of metabolites and for high dimensional α fitting a set of difference equations is difficult. Moreover, in most cases the exact form of F is unknown and alternative models have to be discriminated with a limited set of samples. This poses a challenging experimental design question: at which time points should the samples been taken to optimally discriminate between competing models?

An alternative to modeling simultaneously all measured metabolites in one single set of equations is to first select the most important metabolites and model only those. This route was taken in the nutrikinetics example, where data analysis preselected three metabolites of potential interest. Subsequently, for each metabolite and each subject a set of two coupled fundamental models were used: one describing the behavior under placebo conditions and one the behavior under treatment conditions. The power of this approach is that each subject serves as her/his placebo thereby reducing the inter-person variability dramatically. This is especially important in nutritional studies because the effect sizes are usually small (van Velzen et al. 2008). The equations to describe the behavior of the cumulative excreted metabolite in the placebo (x^{pla}) and treatment (x^{tea}) period are

$$x_{np}^{pla} = \alpha^{pla} + \beta t_{np} + \epsilon_{np}^{pla} \tag{20}$$

$$x_{nt}^{tea} = \alpha^{tea} + \beta t_{nt} + x_{max}^{tea}(1 - e^{-k_e(t_{nt}-\tau)}) + \epsilon_{nt}^{tea}; \quad t_{nt} \geq \tau \tag{21}$$

$$x_{nt}^{tea} = \alpha^{tea} + \beta t_{nt} + \epsilon_{nt}^{tea}; \quad t_{nt} < \tau$$

where 'pla' abbreviates placebo; 'tea' abbreviates the treatment with tea; t_{np}, t_{nt} indicate the time points of measurements for placebo and treatment periods, respectively (these are not equal). The parameters to estimate are τ (lag time); α_{pla}, α_{tea} (off sets); β (linear cumulative increase); $x^{tea}{}_{max}$ (maximum output of metabolite); and k_e (first-order rate constant). The sums of squared values of $\varepsilon^{pla}{}_{np}$ and $\varepsilon^{tea}{}_{nt}$ are simultaneously minimized using a least squares fit. The working of these equations is shown in Fig. 4. After fitting the data, the estimated parameters can

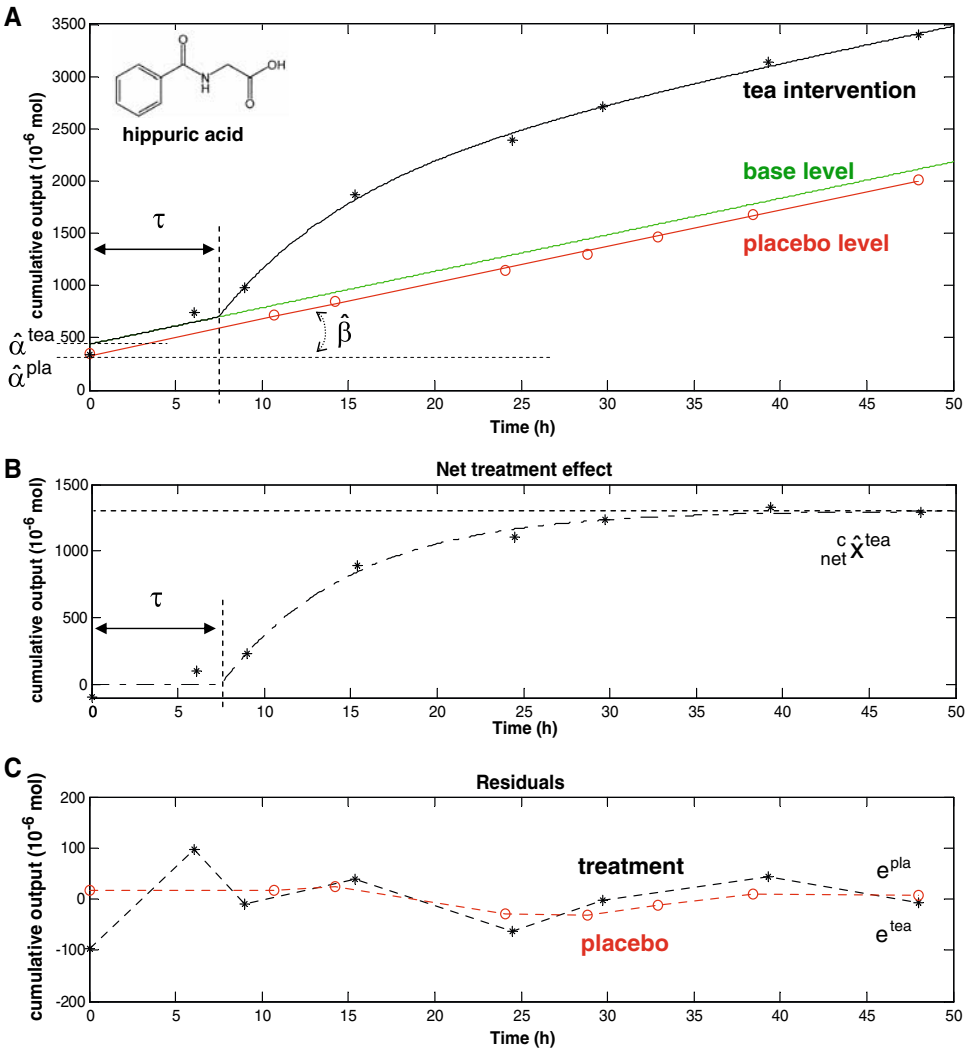

Fig. 4 Nutrikinetic modeling. The *dots* and *stars* represent measured metabolite concentrations at different time points. Shown are the original data (**a**); the net treatment effect (**b**) and the model residuals (**c**). Some of the parameters, explained in the text, are also indicated

be used for phenotyping. For instance, the net cumulative urinary excretion after 48 h can be calculated as $\widehat{x}_{net}^{tea} = \widehat{x}_{max}^{tea}(1 - e^{-\widehat{k}_e(48-\widehat{\tau})})$

and, subsequently, can be used to characterize individual metabolic status (van Velzen et al. 2009).

5.2 Basis functions

If the underlying biology dictates a certain preset form or basis function of the dynamics, then this form can be fitted to the data. Some basis functions (e.g., monotonic decreasing, monotonic increasing and unimodal profiles) are shown in Fig. 5. Examples of the use of splines as basis functions for dynamic data can be found in the gene-expression literature (de Hoon et al. 2002; Storey et al. 2005). No examples are available for similar approaches in metabolomics. One of the reasons may be that postulating basis functions for dynamic metabolomic data is not that easy. Once the basis functions are chosen, the approach is simple because it comes down to simple regression steps. Usually, only few parameters are required and hence the sample sizes can be small. After fitting the individual metabolomic time profiles on the basis functions, the best fitting ones are selected and metabolites with a similar time course behavior are clustered. Hence, a special type of correlation is found, namely the covariation with basis functions. These basis functions are guesses of underlying functions ξ_t and, hence, fit in the framework of (15). This procedure will automatically give a dimension reduction because the basis functions serve as 'latent' variables.

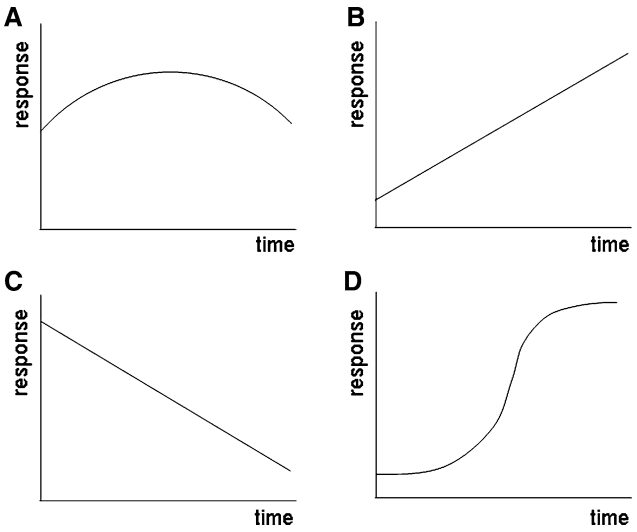

Fig. 5 Examples of simple basis functions

For periodic or oscillating time series, Fourier Analysis or Wavelet transformations can be used. Fourier Analysis requires large sample sizes and repeated patterns; in that respect Wavelets are more flexible. For the analysis of high-dimensional metabolomics data, forcing the latent time variables to follow a wavelet or Fourier transform structure is worthwhile. Combining wavelets with principal component analysis is already done in chemical engineering (Bakshi 1998). A sophisticated way of using basis functions is by means of hidden Markov models (Schliep et al. 2003). The basis functions are implicitly defined in the emission densities of the hidden nodes, thereby also allowing for some flexibility and adaptation of the functions.

5.3 Dimension reducing methods

Usually in metabolomics, many variables are measured, this can range from 100 to 1000 (Bijlsma et al. 2006). Clearly, finding underlying dynamics in such data has to be simplified by reducing the number of variables. This can be done in several ways: by selecting important variables or by dimension reducing methods. The latter class of methods is very broad and versatile: principal component analysis, factor analysis, including all their lagged and dynamic versions. Those will be discussed in some detail.

Variable selection can be done in various ways. If biological knowledge is available, then this should drive the selection. However, in most metabolomics applications discovery of new biology is the goal and hence prior information for selecting the most important variables is by definition not available. Then data driven variable selection techniques have to be used which is a risky undertaking. Although there exist many methods for variable selection in 'classical' statistics (e.g. for regression problems forward selection, backward elimination and stepwise regression), the main problem in high-dimensional data sets is overfitting. By testing (almost) all combinations of variables in a high-dimensional problem, this number becomes so high that overfitting cannot be avoided. Hence, such a selection has always to be accompanied with a good validation strategy to avoid the so-called selection bias (Ambroise and McLachlan 2002). The whole topic of variable selection, including proper validation, deserves a critical review in itself. Obviously, upon assuming that we have selected a number of relevant variables, preferably using *a priori* biological knowledge, we can use some of the dynamic methods as exemplified in this paper.

Combining *a priori* knowledge of the underlying dynamics with a dimension reduction approach is best explained by using the factor analysis framework:

$$\begin{aligned} \mathbf{x}_t &= \Lambda \mathbf{y}_t + \epsilon_t + \mu \\ \mathbf{y}_{t+1} &= f(\mathbf{y}_t, t; \alpha), \end{aligned} \tag{22}$$

where, again, α contains the (unknown) parameters. The issue is the form of the functional relationship f. Either this function is known or it has to be estimated from the data. To the best of our knowledge, models like (22) have not been explored in X-omics data analysis. Model (22) can be simplified (dropping the term μ for simplicity) by postulating

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Lambda} \mathbf{y}_t + \epsilon_t \\ \mathbf{y}_{t+1} &= \mathbf{\Theta} \mathbf{y}_t, \end{aligned} \tag{23}$$

where the dynamics are in the latent variables $\mathbf{y}_t$ in a simple way. This is a combination of dimension reduction and time series analysis.

The idea of making factor analysis models dynamic can also be implemented differently (dropping again the term μ). Dynamic factor analysis (Molenaar 1985) models the data as

$$\mathbf{x}_t = \sum_{l=0}^{L} \mathbf{\Lambda}_l \mathbf{y}_{t-l} + \epsilon_t, \tag{24}$$

where $\mathbf{y}_t$ contains the R factor scores at time t and these factors scores are assumed to be generated by a white noise process (uncorrelated). The index l stands for lag and, hence, lags upto and including L are considered. All the dynamics in $\mathbf{x}_t$ is captured by the lagged loading matrices $\mathbf{\Lambda}_l$.

Component models can also be made dynamic. Besides the obvious extension of (16) where the scores $\mathbf{z}_t$ are forced to follow a predefined dynamic model, there are two alternative ways of constructing dynamic PCA models, called lagged-PCA and dynamic PCA. Lagged-PCA is a simplified version of the more general Lagged Simultaneous Component Analysis (Timmerman 2001) for analyzing multiple data sets simultaneously. To explain the idea of lagged-PCA, it is convenient to introduce the backshift matrix $\mathbf{B}_l$ - where $l = 0,\ldots, L$ defines the time lags - which is defined as follows

$$\mathbf{B}_l = [0_{T\times(L-l)}|I_T|0_{T\times l}]. \tag{25}$$

Using the scores $\mathbf{Z}$ ($(T + L) \times R$), the loadings $\mathbf{P}$ ($J \times R$), and residuals $\mathbf{E}$ ($T \times J$), the lagged-PCA model becomes

$$\mathbf{X} = \sum_{l=0}^{L} \mathbf{B}_l \mathbf{Z} \mathbf{P}_l' + \mathbf{E}. \tag{26}$$

A small numerical example for $L = 2$ is given to illustrate the working of this model:

$$\mathbf{B}_0\mathbf{Z} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 9 & 10 \\ 11 & 12 \\ 13 & 14 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \\ 11 & 12 \\ 13 & 14 \end{pmatrix}, \tag{27}$$

which shows the implicitly defined zero shift matrix $\mathbf{B}_0$ and scores $\mathbf{Z}$. The first lag is modelled as

$$\mathbf{B}_1\mathbf{Z} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 9 & 10 \\ 11 & 12 \\ 13 & 14 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix}, \tag{28}$$

and the second lag is modelled as

$$\mathbf{B}_2\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 9 & 10 \\ 11 & 12 \\ 13 & 14 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix}, \tag{29}$$

which results in the model of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{B}_0\mathbf{Z}\mathbf{P}_0' + \mathbf{B}_1\mathbf{Z}\mathbf{P}_1' + \mathbf{B}_2\mathbf{Z}\mathbf{P}_3' + \mathbf{E}. \tag{30}$$

Clearly, the lagged-PCA model has three sets of loadings ($\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$) representing the relationships between the variables in $\mathbf{X}$ and the scores $\mathbf{Z}$ at different lag-times.

The backshift operator works on the scores in lagged-PCA, this operator can also work directly on the $\mathbf{X}$ matrix, resulting in dynamic-PCA (Ku et al. 1995). The idea is shown using a simple example for $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} 1 & 8 \\ 2 & 9 \\ 3 & 10 \\ 4 & 11 \\ 5 & 12 \\ 6 & 13 \\ 7 & 14 \end{pmatrix} \tag{31}$$

in which the rows represent time points and columns metabolites. Upon using backshift operators, the submatrices

$$\mathbf{B}_0\mathbf{X} = \begin{pmatrix} 3 & 10 \\ 4 & 11 \\ 5 & 12 \\ 6 & 13 \\ 7 & 14 \end{pmatrix}, \mathbf{B}_1\mathbf{X} = \begin{pmatrix} 2 & 9 \\ 3 & 10 \\ 4 & 11 \\ 5 & 12 \\ 6 & 13 \end{pmatrix}, \mathbf{B}_2\mathbf{X} = \begin{pmatrix} 1 & 8 \\ 2 & 9 \\ 3 & 10 \\ 4 & 11 \\ 5 & 12 \end{pmatrix}, \tag{32}$$

can be concatenated to form $\widetilde{\mathbf{X}}$

$$\widetilde{\mathbf{X}} = [\mathbf{B}_0\mathbf{X}|\mathbf{B}_1\mathbf{X}|\mathbf{B}_2\mathbf{X}] \tag{33}$$

which can then be subjected to an ordinary PCA. Hence, the dynamics are in the manifest variables and subsequently, a PCA captures these dynamics. Note that the covariance matrix of $\widetilde{\mathbf{X}}$ contains three types of covariances: (i) those between variables (lag = 0), (ii) those within different time points of the same variables (auto-covariance) and (iii) those between different time points of different variables (cross-covariance). Hence, the subsequent PCA-scores capture a mixture of these three, obscuring the individual contributions.

In fact, matrix $\widetilde{\mathbf{X}}$ is a matricized three-way array (Kiers (2000)) $\underline{\mathbf{X}}$ of size $T \times J \times L$. Hence, an alternative would be to analyze this array with PARAFAC or Tucker3 models (Smilde et al. 2004). It is difficult to say how many samples are needed for stable results for the estimates of the different dynamic factor analysis, lagged-PCA and dynamic-PCA models. A disadvantage of dynamic-PCA is that it 'cuts off' parts of $\mathbf{X}$ (the higher L the more severe this cut-off is) thereby reducing the number of samples in the time direction. This stability will depend on the measurement error, the intrinsic dynamics and the complexity (i.e., intrinsic rank) of $\mathbf{X}$.

Dynamic-PCA also has a close cousin called dynamic-PLS. There are three alternatives for dynamic-PLS. The first version takes lagged x-variables and performs then an ordinary PLS between the expanded (and lagged) $\mathbf{X}$ matrix and the phenotype $\mathbf{y}$. This procedure is based on Finite Impuls Response models as used in systems identification (Ljung 1987). An extension of this is to incorporate also lagged y-variables in the new X-block (Qin and McAvoy 1996).This is a direct generalization of the ARMA modeling strategy (see below). The drawback of both methods is that the $\mathbf{X}$ matrix (which is already huge) is even expanded with many lagged variables thereby aggravating the problem of low sample-to-variables ratio's. Hence, this does not seem to be a viable route to take, despite the dimension reduction capability of PLS.

An alternative is presented in the process control literature and consists of defining a dynamic filter to account for the dynamics in $\mathbf{X}$ and, subsequently, building an (static) model between the filtered $\mathbf{X}$ and $\mathbf{y}$ with PLS (Kaspar and Ray 1993). Stated otherwise, the dynamics in $\mathbf{X}$ are 'whitened' and then related to $\mathbf{y}$. Although this approach does not have the drawback of 'blowing-up' the dimensions of $\mathbf{X}$ unfavorably, it is sensitive to the specified form the filter. Tuning such a filter might not be a trivial task in dynamic metabolomics.

Another way to account for dynamics is to use time as an external variable. This is the approach taken by batch modelling (Wold et al. 1998). The idea is building a PLS model between $\mathbf{X}$ ($T \times J$) and a y-vector containing either a maturity variable or the time corresponding to the sampling of the rows of $\mathbf{X}$ (see Fig. 6).

A maturity variable is a measured variable indicating the progress of the biological process. In both cases, the PLS model finds features in $\mathbf{X}$ related to a time axis and as such is a dynamic approach. This approach has been used in several metabolomics applications (Antti et al. 2002; Jonsson et al. 2006), but has also some drawbacks. One of the problems of this approach is that it will poorly describe features in $\mathbf{X}$ that do not align with the imposed time axis (Westerhuis et al. 1999). A new method has been published based on OPLS models to describe successive differences between two adjacent time points (Rantalainen et al. 2008). The drawback of this method is that the time trajectory information is given in a set of models hampering interpretation.

5.4 Time series analysis

There exists a multivariate extension of time series models:

$$\mathbf{x}_{t+1} = \Theta\mathbf{x}_t + \epsilon_t, \tag{34}$$

which is a multivariate AR(1) model (or a Vector AR(1) model) with Θ an $J \times J$ matrix of coefficients and ε_t a J-valued vector of random shocks. Again, the matrix Θ has to fulfill some regularity conditions. Extension to second-order systems and moving average models also exist. Estimating the model parameters Θ is possible, but requires a lot of samples for stable results, especially if J is rather large (Holtz-Eakin et al. 1988).

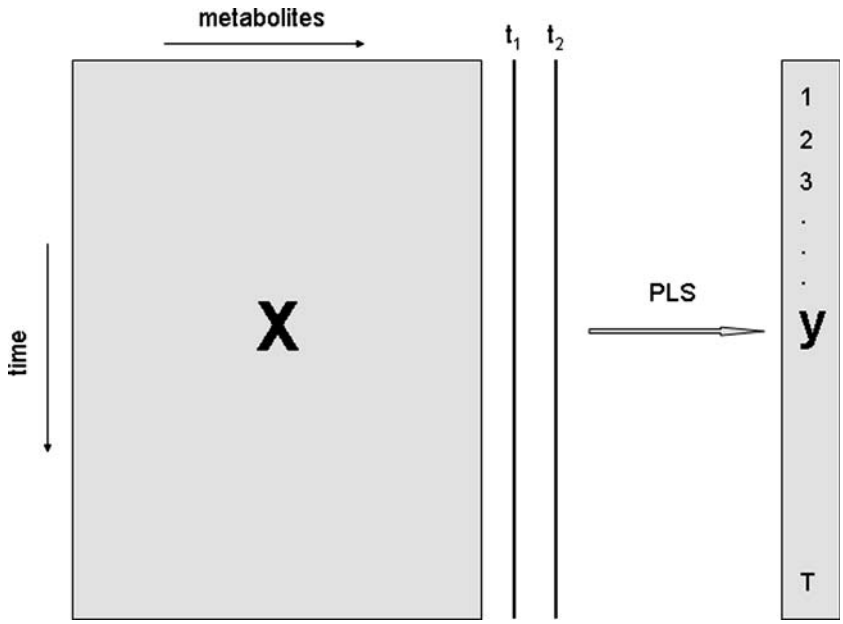

Fig. 6 Batch process modeling: a PLS model is built between time-resolved measured metabolites (collected in **X** and a y-variable measuring time. The score vectors t_1 and t_2 are obtained from the PLS modeling

For regression type problems, ARMAX models can be used. An example of an ARMAX (1,1) model is

$$y_{t+1} = \theta y_t + \varphi_1 x_{t+1} + \varphi_2 x_t + \epsilon_{t+1}, \tag{35}$$

where θ, φ_1 and φ_2 are parameters to be estimated. The time lags used for x and y are both 1, hence the notation ARMAX (1,1) model.

An alternative is to write the set of difference equations in state-space notation,

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \epsilon_t, \end{aligned} \tag{36}$$

where **A** is the $J \times J$ system matrix, **C** the $K \times J$ measurement matrix, $\mathbf{u}_t$ a $M \times 1$ vector of inputs and **B** an $M \times J$ input transfer matrix and **y** ($k \times 1$) the vector of measurements (Fortmann and Hitz 1977; Ljung 1987). For generality, the forcing term $\mathbf{B}\,\mathbf{u}_t$ is introduced. In the case of metabolomics experiments, this forcing term is usually complicated. This forcing term can be a diet, a toxic compound or an administered drug. Then **B** represents the influence of such an intervention directly on the metabolites, which is hard to estimate and is usually not explicitly considered. Since we consider all measured metabolites, **C** = **I** (the identity matrix). Then, by rearranging (36) and solving for $\mathbf{y}_t$, we get

$$\mathbf{y}_{t+1} = \mathbf{A}\mathbf{y}_t + \epsilon_{t+1} - \mathbf{A}\epsilon_t, \tag{37}$$

which is a multivariate ARMA (1,1) model, showing the intimate relationship between time series models and state-space models. Some application of state-space models are reported in the gene-expression literature (Wu et al. 2004) but to our knowlegde, no applications have been reported in metabolomics.

State-space models can also be combined with dimension reduction when the state variables $\mathbf{x}_t$ are regarded as underlying (latent) variables and $\mathbf{y}_t$ represent the measured metabolites. The matrix **C** then relates the manifest to the latent variables and the dimensionality of $\mathbf{x}_t$ can be much lower than $\mathbf{y}_t$. The dynamics are now imposed on the latent variables. This approach differs from dynamic factor analysis (see Eq. 3) because in the latter case the time instances of the latent variables are considered as independent gaussian.

5.5 ANOVA

A different way of tackling dynamic data is by using analysis of variance (ANOVA) models (Searle 1971). In such models, the factor time can be accounted for in both a qualitative and quantitative way. The qualitative analysis pertains to modeling the factor 'time' at its different levels, whereas the quantitative analysis models the factor 'time' in terms (of mixtures) of linear, quadratic and/or cubic trends (depending on the number of time points available). A convenient way of quantitative modeling is by using orthogonal polynomials, since the consecutive terms in such polynomials are orthogonal thereby facilitating the estimation process. In the gene-expression literature, there are examples of both qualitative modeling (Storey et al. 2005), as well as quantitative modeling (Conesa et al. 2006).

The multivariate extension of ANOVA is called Multivariate Analysis of Variance. This extension is not straightforward (e.g., which test statistic to use (Stahle and Wold 1990) and it is not clear how to use it for multivariate time-resolved data (e.g., how to treat the factor time). Moreover, high-dimensional data gives singular covariance

matrices. One way to generalize ANOVA to the high-dimensional case is by performing separate ANOVA's on the individual metabolite profiles thereby partitioning the data according to sources of variation. Subsequent simultaneous component analysis (SCA) models on the different parts of the data perform then the necessary dimension reduction. These approaches, called multilevel-SCA (MSCA) and ANOVA-SCA (ASCA), have been used succesfully in psychometrics (Timmerman and Kiers 2003), metabolomics (Jansen et al. 2004, 2005; Smilde et al. 2005; Vis et al. 2007), proteomics (Harrington et al. 2005), geneexpression (Nueda et al. 2007) and also in process chemometrics (de Noord and Theobald 2005). There exists also a multiway version: PARAFASCA (Jansen et al. 2008). The ASCA methods do not assume linear time behavior and is therefore a general method for capturing nonlinear time behavior (Smilde et al. 2008).

An alternative - called SMART - is to apply special preprocessing steps of the metabolomics data and perform a subsequent PCA (Keun et al. 2004). SMART has some drawbacks, notably its lack of orthogonal partitioning hampering interpretation (Jansen et al. 2005). Moreover, SMART is not a dynamic method according to our definition. ASCA is only a dynamic method if the factor time is treated in a quantitative way in the ANOVA model.

A route taken in the gene-expression literature is to perform single ANOVA's per gene and then cluster the results afterwards (Conesa et al. 2006). It is also possible to combine both steps by using mixture modeling to find genes with a similar time behavior and estimate the dynamic behavior then for the whole cluster simultaneously (Rodriguez-Zas et al. 2006). This amounts to a considerable reduction of parameters to estimate. This procedure can also be used in metabolomics.

5.6 Smoothness

A very general approach to account for the consecutiveness of time evolving processes is by using smoothness constraints. In terms of curve fitting of a single metabolite time profile, this approach can be described as

$$\min_{\mathbf{y}} \left[||\mathbf{x}-\mathbf{y}||^2 + \lambda ||\mathbf{D}\mathbf{y}||^2 \right], \tag{38}$$

where $\mathbf{x} = (x_1, ..., x_T)'$ is the vector of original time series measurements, $\mathbf{y}$ contains the fitted (or smoothed) values $\mathbf{y} = (y_1, ..., y_T)'$, $\mathbf{D}$ is a matrix differentiating consecutive elements of $\mathbf{y}$ and $\lambda \geq 0$ is a metaparameter regulating the constraint $\|\mathbf{D}\mathbf{y}\|^2$ (Eilers 2003; Ramsay and Silverman 1997). There are also other constraints possible, e.g., penalties on second-order differences in (38)

A way to combine ideas of consecutiveness with dimension reduction in time-resolved high-dimensional metabolomics data is to make the estimated principal component scores 'as autocorrelated as possible'. The method Maximum Autocorrelation Factors (MAF) does this for spatially resolved data, and can be used directly for the time-resolved data (Larsen 2002). This method calculates components $\mathbf{z}_r$, $r = 1,\ldots, R$ for which the lag l entries have a maximum autocorrelation, while being mutually orthogonal across $r = 1,\ldots, R$. The lag l has to be chosen by the user. Interestingly, the MAF method is equivalent to the Molgedy-Schuster version of Independent Component Analysis (Larsen 2002). Hence, using Molgedy-Schuster Independent Component Analysis on the time-resolved data would also invoke consecutiveness.

Combining smoothness with dimension reduction can be done by applying smooth-PCA (Westerhuis et al. 2008). The smooth-PCA method models the data by solving

$$\min_{\mathbf{Z},\mathbf{P}} \left[||\mathbf{X}-\mathbf{Z}\mathbf{P}'||^2 + \lambda ||\mathbf{D}\mathbf{Z}||^2 \right], \tag{39}$$

where again $\mathbf{D}$ is a first-order or second-order difference matrix and $\lambda \geq 0$ is a penalty parameter. The higher the value of λ, the smoother the scores in $\mathbf{Z}$ will become.

An example will be shown for the monkey data and, for simplicity, results will only be shown for a typical female monkey, although an analysis using all ten monkeys (i.e., smooth-Simultaneous Component Analysis) would also be possible. Prior to analysis, the data were mean-centered across the time-mode. A smooth-PCA is compared to a normal (non-smooth) PCA. To calculate the smooth-PCA, a second-order penalty was used in (39) and a special arrangement has to be made to accommodate for the non-equidistant sampling scheme (Westerhuis et al. 2008).

The first score vectors are shown for different values of λ, see Fig. 7.

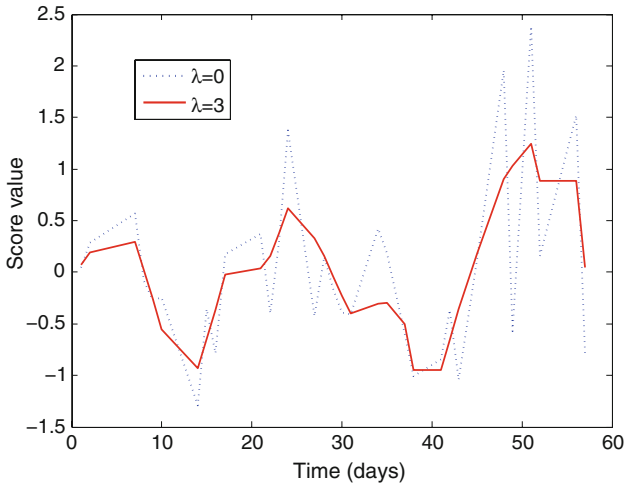

Fig. 7 Scores of the first PC and smooth-PC. Legend: the numbers *0* and *3* refer to the value of λ

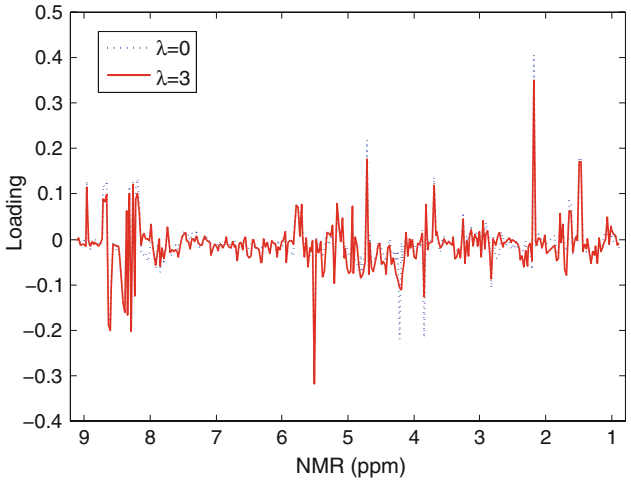

Fig. 8 Loadings of the first PC and smooth-PC. Legend: *drawn line* is PCA and *dotted line* is smooth-PCA

For $\lambda = 0$, the PCA solution is obtained. Raising λ penalizes roughness more and makes the scores smoother. It is hard to give objective criteria to select λ, but a value of 3 seems reasonable, whereas a value of 30 (result not shown) gives a too smooth solution. The first score vector shows a rhythm with a period of 27–28 days which may be due to the oestric cycle (Xu et al. 2007). The corresponding loadings are shown in Fig. 8, and do not differ much between normal PCA and smooth-PCA. The first score explains 21.9% of the variation in the data, whereas the solution with $\lambda = 3$ explains 13.8%. Hence, the loss of variance explained should be compensated by a better interpretability. A detailed interpretation of these results is outside the scope of this paper.

6 Discussion

It would be nice to end this paper with giving a scheme on when to apply which method in what situation. As in all statistical modeling situations, the dynamic modeling process is also an art without predefined rules.

First, the type of biological question, the amount of knowledge of the system and the availability of data is important. If a phenotypic variable is available, then this variable might steer the unravelling of the dynamics in the metabolome data. Looking for specific rhythms with known dynamics calls for different methods than exploring dynamic patterns in metabolomics data of relatively unknown organisms.

Second, the experimental design is important, the number of time points, their spacing in time and the number of metabolites measured. The design puts restrictions on the methods to use. Some of the methods require many time-resolved samples (time series methods) whereas other methods can do with a limited number (basis functions). With high-dimensional data, it is worthwhile to consider methods involving dimension reduction.

Third, the type of measurements performed is important. For example, NMR and MS data have different characteristics and these should be kept in mind when using dynamic methods. Some of the methods are easily adapted to accommodate nonhomogeneous errors. Such an adaption might be profitable in terms of the quality of the estimated parameters.

Preferably, the choices for measurement and data analysis are driven by the biological question, the data generating process, the experimental design and the assumptions of the data analysis methods.

Notation

$\mathbf{x}$ (vector)	bold lowercase
$\mathbf{X}$ (matrix)	bold uppercase
$t = 1,\ldots, T$	time index (T = number of time points)
$j = 1,\ldots, J$	variable index (J = number of variables)
$l = 0,\ldots, L$	lagging index (L = number of time lags)
$r = 1,\ldots, R$	principal component index (R = number of principal components)
$\mathbf{B}_l$ $(T \times (T + L))$	back-shift operator for lag l
$\mathbf{U}$ $(I \times I)$	left singular vectors of $\mathbf{X}$
$\mathbf{S}$ $(I \times I)$	diagonal matrix containing (in descend-ingorder) the singular values of $\mathbf{X}$
$\mathbf{V}$ $(J \times I)$	right singular vectors of $\mathbf{X}$
$\mathbf{U}_1$ $(I \times R)$	R 'largest' left singular vectors of$\mathbf{X}$
$\mathbf{S}_1$ $(R \times R)$	R largest singular values of $\mathbf{X}$
$\mathbf{V}_1$ $(J \times R)$	R 'largest' right singular vectors of$\mathbf{X}$
$\mathbf{Z}$ $(T \times R)$	scores in a PCA model of $\mathbf{X}$
$\mathbf{P}$ $(J \times R)$	loadings in a PCA model of $\mathbf{X}$
$\mathbf{A}$ $(J \times J)$	system matrix
$\mathbf{D}$ $(T \times T)$	difference matrix
$\Lambda(T \times R$	loading matrix in factor analysis
θ	parameter(s) in AR models
ϕ	parameter(s) in MA models
φ	parameter(s) in the X part of ARMAX models
α	parameter(s) in DE's and DFE's
ε_t (scalar)	random shock, random error or specific factor
λ (scalar)	penalty parameter in penalized methods

References

Ambroise, C., & McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America, 99*(10), 6562–6566.

Anderson, T. (2003). *An introduction to multivariate statistical analysis*. New York: Wiley.

Antti, H., Bollard, M. E., Ebbels, T., Keun, H., Lindon, J. C., Nicholson, J. K., et al. (2002). Batch statistical processing of h-1 nmr-derived urinary spectral data. *Journal of Chemometrics, 16*, 461–468.

Apostu, R., & Mackey, M. C. (2008). Understanding cyclical thrombocytopenia: A mathematical modeling approach. *Journal of Theoretical Biology, 251*(2), 297–316.

Bakshi, B. R. (1998). Multiscale pca with application to multivariate statistical process monitoring. *AIChE Journal, 44*, 1596–1610.

Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., et al. (2006). Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry, 78*(2), 567–574.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis, forecasting and control*. Englewood Cliffs: Prentice Hall.

Cao, J., & Zhao, H. (2008). Estimating dynamic models for gene regulation networks. *Bioinformatics, 24*(14), 1619–1624.

Conesa, A., Nueda, M. J., Ferrer, A., & Talon, M. (2006). Masigpro: A method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics, 22*(9), 1096–1102.

de Hoon, M. J. L., Imoto, S., & Miyano, S. (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines. *Bioinformatics, 18*(11), 1477–1485.

de Noord, O. E., & Theobald, E. H. (2005). Multilevel component analysis and multilevel pls of chemical process data. *Journal of Chemometrics, 19*(5–7), 301–307.

Eilers, P. H. C. (2003). A perfect smoother. *Analytical Chemistry, 75*(14), 3631–3636.

Fortmann, T. E., & Hitz, K. L. (1977). *An introduction to linear control systems*. New York: Marcel Dekker Inc.

Glass, L., & Mackey, M. C. (1988). *From clocks to chaos: "The rythms of life"*. Princeton: Princeton University Press.

Harrington, P. D., Vieira, N. E., Espinoza, J., Nien, J. K., Romero, R., & Yergey, A. L. (2005). Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta, 544*(1–2), 118–127.

Heijne, W. H. M., Lamers, R.-J. A. N., van Bladeren, P. J., Groten, J. P., van Nesselrooij, J. H. J., & van Ommen, B. (2005). Profiles of metabolites and gene expression in rats with chemically induced hepatic necrosis. *Toxicologic Pathology, 33*(4), 425–433.

Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica, 56*(6), 1371–1395.

Hood, L. (2003). Systems biology: Integrating technology, biology, and computation. *Mechanisms of Ageing and Development, 124*(1), 9–16.

Jansen, J. J., Hoefsloot, H. C. J., Boelens, H. F. M., van der Greef, J., & Smilde, A. K. (2004). Analysis of longitudinal metabolomics data. *Bioinformatics, 20*(15), 2438–2446.

Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., Westerhuis, J. A., & Smilde, A. K. (2005). Asca: Analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics, 19*, 469–481.

Jansen, J. J., Bro, R., Hoefsloot, H. C. J., van den Berg, F. W. J., Westerhuis, J. A., & Smilde, A. K. (2008). Parafasca: Asca combined with parafac for the analysis of metabolic fingerprinting data. *Journal of Chemometrics, 22*(1–2), 114–121.

Jolliffe, I. T. (1986). *Principal component analysis*. Berlin: Springer Verlag.

Jonsson, P., Stenlund, H., Moritz, T., Trygg, J., Sjostrom, M., Verheij, E. R., et al. (2006). A strategy for modelling dynamic responses in metabolic samples characterized by gc/ms. *Metabolomics, 2*(3), 135–143.

Kaspar, M. H., & Ray, H. (1993). Dynamic pls modelling for process control. *Chemical Engineering Science, 48*, 3447–3461.

Keun, H. C., Ebbels, T. M., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E., et al. (2004). Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chemical Research inToxicology, 17*(5), 579–587.

Kholodenko, B., & Westerhoff, H. (Eds.). (2004). *Metabolic engineering in the post genomic era*. Wymondham, UK: Horizon Bioscience.

Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics, 14*(3), 105–122.

Kleemann, R., Verschuren, L., van Erk, M., Nikolsky, Y., Cnubben, N., Verheij, E., et al. (2007). Atherosclerosis and liver inflammation induced by increased dietary cholesterol intake: A combined transcriptomics and metabolomics analysis. *Genome Biology, 8*(9), R200.

Kok, P., Roelfsema, F., Frolich, M., van pelt, J., Stokkel, M. P. M., Meinders, A. E., et al. (2006). Activation of dopamine d2 receptors simultaneously ameliorates various metabolic features of obese women. *American Journal of Physiology- Endocrinology and Metabolism, 291*(5), 1038–1043.

Kok, S. W., Roelfsema, F., Overeem, S., Lammers, G. J., Frohlich, M., Meinders, A. E., et al. (2004). Pulsatile lh release is dimished, whereas fsh secretion is normal, in hypocretin-deficient narcoleptic men. *American Journal of Physiology-Endocrinology and Metabolism, 287*, 630–636.

Ku, W. F., Storer, R. H., & Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 30*, 179–196.

Larsen, R. (2002). Decomposition using maximum autocorrelation factors. *Journal of Chemometrics, 16*(8–10), 427–435.

Ljung, L. (1987). *System identification*. New Yersey: Prentice Hall.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press.

Molenaar, P. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181–201.

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C. J., Smilde, A. K., Talon, M., et al. (2007). Discovering gene expression patterns in time course microarray experiments by anova-sca. *Bioinformatics, 23*, 1792–1800.

Qin, S. J., & McAvoy, T. J. (1996). Nonlinear fir modeling via a neural net pls approach. *Computers and Chemical Engineering, 20*(2), 147–159.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional Data Analysis*. Heidleberg: Springer.

Ramsay, J.O., Hooker, G., Campbell, D., & Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 69*, 741–770.

Rantalainen, M., Cloarec, O., Ebbels, T. M. D., Lundstedt, T., Nicholson, J. K., Holmes, E., et al. (2008). Piecewise multivariate modelling of sequential metabolic profiling data. *BMC Bioinformatics, 9*, 105.

Rodriguez-Zas, S. L., Southey, B. R., Whitfield, C. W., & Robinson, G. E. (2006). Semiparametric approach to characterize unique gene expression trajectories across time. *BMC Genomics, 7*, 233.

Rubingh, C. M., Bijlsma, S., Jellema, R. H., Overkamp, K. M., van der Werf, M. J., & Smilde, A. K. (2009). Analyzing longitudinal microbial metabolomics data. *Journal of Proteome Research* (accepted).

Samoilov, M., Arkin, A., & Ross, J. (2001). On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos, 11*(1), 108–114.

Schliep, A., Schonhuth, A., & Steinhoff, C. (2003). Using hidden markov models to analyze gene expression time course data. *Bioinformatics, 19*, i255–i263.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Smilde, A. K., Bro, R., & Geladi, P. (2004). *Multi-way analysis. Applications in the chemical sciences*. Chichester: Wiley.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R. A. N., van der Greef, J., & Timmerman, M. E. (2005). Anova-simultaneous component analysis (asca): A new tool for analyzing designed metabolomics data. *Bioinformatics, 21*(13), 3043–3048.

Smilde, A. K., Hoefsloot, H. C. J., & Westerhuis, J. A. (2008). The geometry of asca. *Journal of Chemometrics, 22*(7–8), 464–471.

Stahle, L., & Wold, S. (1990). Multivariate-analysis of variance (manova). *Chemometrics and Intelligent Laboratory Systems, 9*(2), 127–141.

Stephanopoulos, G., Aristidou, A., & Nielsen, J. (1998). *Metabolic engineering. Principles and methodologies*. San Diego: Academic Press.

Storey, J. D., Xiao, W. Z., Leek, J. T., Tompkins, R. G., & Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America, 102*(36), 12837–12842.

Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Cambridge: Perseus Books

Timmerman, M. E. (2001). *Component analysis of multisubject multivariate longitudinal data*. Ph.D. thesis, University of Groningen.

Timmerman, M. E., & Kiers, H. A. L. (2003). Four simultaneous component models of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika, 86*(1), 105–122.

van Berlo, R. J. P., van Someren, E. P., & Reinders, M. J. T. (2003). Studying the conditins for learning dynamic bayesian networks to discovery genetic regulatory networks. *Simulation, 79*(12), 689–702.

van der Greef, J., Hankemeier, T., & McBurney, R. N. (2006). Metabolomics-based systems biology and personalized medicine: Moving towards n = 1 clinical trials? *Pharmacogenomics, 7*(7), 1087–1094.

van der Greef, J., Martin, S., Juhasz, P., Adourian, A., Plasterer, T., Verheij, E. R., et al. (2007). The art and practice of systems biology in medicine: Mapping patterns of relationships. *Journal of Proteome Research, 6*(4), 1540–1559.

van Velzen, E., Westerhuis, J., van Duynhoven, J., van Dorsten, F., Grun, C., Jacobs, D., et al. (2009). Phenotyping tea consumers by nutrikinetic analysis of polyphenolic end-metabolites. *Journal of Proteome Research, 8*(7), 3317–3330.

van Velzen, E., Westerhuis, J., van Duynhoven, J., van Dorsten, F., Hoefsloot, H., Smit, S., et al. (2008) Multilevel data analysis of a cross-over design human nutritional study. *Journal of Proteome Research, 7*(10), 4483–4491

Vis, D., Westerhuis, J., Hoefsloot, H., Pijl, H., Roelfsema, F., van der Greef, J., et al. (2009). Endocrine pulse identification using penalized methods and a minimum set of assumptions. *American Journal of Physiology* (accepted).

Vis, D. J., Westerhuis, J. A., & Smilde, A. K. (2007). Statistical validation of megavariate effects in asca. *BMC Bioinformatics, 8*, 322.

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1999). Comparing alternative approaches for multivariate statistical analysis of batch process data. *Journal of Chemometrics, 13*, 397–413.

Westerhuis, J. A., Hoefsloot, H. C. J., & Smilde, A. K. (2008). Smooth-pca (submitted).

Wold, S., Kettaneh, N., Friden, H., & Holmberg, A. (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and intelligent laboratory systems, 44*, 331–340.

Wolkenhauer, O. (2002). Mathematical modelling in the post-genome era: Understanding genome expression and regulation–a system theoretic approach. *Biosystems, 65*(1), 1–18.

Wu, F., Zhang, W., & Kusalik, A. (2004). State-space model with time delays for gene regulatory networks. *Journal of Biological Systems, 12*(4), 483–500.

Xu, J., Xu, F., Hennebold, J. D., Molskness, T. A., & Stouffer, R. L. (2007). Expression and role of the corticotropin-releasing hormone/urocortin-receptor-binding protein system in the primate corpus luteum during the menstrual cycle. *Endocrinology, 148*(11), 5385–5395.