Application of rotated PCA models to facilitate interpretation of metabolite profiles

Lawaetz, Anders Juul; Schmidt, Bonnie; Stærk, Dan; Jaroszewski, Jerzy W.; Bro, Rasmus

Published in:
Planta Medica

DOI:
10.1055/s-0028-1112194

Publication date:
2009

Document version
Publisher's PDF, also known as Version of record

Application of Rotated PCA Models to Facilitate Interpretation of Metabolite Profiles: Commercial Preparations of St. John's Wort

Author **Anders Juul Lawaetz**[1], **Bonnie Schmidt**[1], **Dan Staerk**[2], **Jerzy W. Jaroszewski**[3], **Rasmus Bro**[1]

Affiliation
[1] Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Copenhagen, Denmark
[2] Department of Basic Sciences and Environment, Faculty of Life Sciences, University of Copenhagen, Copenhagen, Denmark
[3] Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Copenhagen, Copenhagen, Denmark

Key words
- *Hypericum perforatum* L.
- St. John's wort
- Clusiaceae
- principal component analysis (PCA)
- orthogonal rotation
- metabolite profiling

Abstract

This paper describes the application of orthogonal rotation of models based on principal component analysis (PCA) of ^{1}H nuclear magnetic resonance (NMR) spectra and high-performance liquid chromatography-photo diode array detection (HPLC-PDA) profiles of natural product mixtures using extracts of antidepressive pharmaceutical preparations of St. John's wort as an example. ^{1}H-NMR spectroscopy of complex mixtures is often used in metabolomic, metabonomic and metabolite profiling studies for assessment of sample composition. Interpretation of the derived chemometric models may be complicated because several sample properties often contribute to each principal component and because the influence of individual metabolites may be shared by several principal components. Furthermore, extensive signal overlap in ^{1}H-NMR spectra poses additional challenges to the interpretation of PCA models derived from such data. Orthogonal rotation of PCA models derived from ^{1}H-NMR spectra and HPLC-PDA profiles of the extracts of St. John's wort preparations facilitate interpretation of the model. Using the varimax criterion, rotation of loadings provides simpler conditions for understanding the influence of individual metabolites on the observed clustering. Alternatively, rotation of scores simplifies the understanding of the influence of whole metabolite profiles on the clustering of individual samples.

Introduction

^{1}H-NMR spectroscopy is an attractive analytical technique for assessment of samples of biological origin, i. e., biofluids and plant extracts. The technique is non-destructive, applicable to intact biomaterial and information-rich with regard to molecular structure elucidation. Thus, the technique has been widely used as the analytical platform to generate information-dense data in metabonomic, metabolomic, and metabolite profiling studies. However, ^{1}H-NMR spectra of biological samples can be extremely complex as they may contain thousands of distinctive resonances. Therefore, visual inspection of a series of such spectra may only release a small percentage of the total information available.

Computer-based methods are often used to reduce the complexity of data to a suitable level. In ^{1}H-NMR-based metabolite profiling studies, principal component analysis (PCA) is often used [1]. Graphical outputs from PCA enable researchers across disciplines to discuss detailed facets of conceivably complex mathematical models. A PCA model uses orthogonal and intrinsically abstract latent variables. This means that interpretation of the model in terms of finding the connection between loadings and the variables used in the analysis can be difficult. Even though a PCA bi-plot of scores and loadings provides insight into the structure of the data, it can still be difficult to interpret the many correlations occurring in NMR-based metabonomic studies.

In this study we explore a route to simplify the interpretation of complex PCA models with respect to the influence of individual compounds on the observed clustering of samples. ^{1}H-NMR spectra and HPLC-PDA profiles of extracts of 24 commercially available preparations of St. John's wort, a popular herbal medicine, are used as model data sets. Metabolite profiles based on ^{1}H-NMR spectroscopy have previously proven useful for assessment of herbal medicines or plant extracts using different two-way chemometric methods [2], [3], [4], [5], [6], [7], [8], [9].

received June 4, 2008
revised September 26, 2008
accepted October 27, 2008

Bibliography
DOI 10.1055/s-0028-1112194
Planta Med 2009; 75: 271–279

Published online December 18, 2008
ISSN 0032-0943

Correspondence
Anders Juul Lawaetz
Department of Food Science
Faculty of Life Sciences
University of Copenhagen
Rolighedsvej 30
1958 Frederiksberg C
Denmark
Tel.: +45-3533-3254
Fax: +45-3533-3245
ajla@life.ku.dk

In an earlier study comprising commercial preparations of St. John's wort obtained from retail stores in Denmark, interpretation of the full-resolution ^{1}H-NMR data was based on separate PCA models derived from samples formulated as tablets and capsules, respectively [8]. Moreover, another data set derived from St. John's wort preparations originating from several continents and based on HPLC-PDA profiles has been analyzed by applying parallel factor (PARAFAC) analysis [10]. This provided relative concentrations of individual compounds, which were used to facilitate comparison of samples by PCA. Interpretation of the PCA model in terms of constituents responsible for the differences and similarities in composition between preparations was straightforward, because the analysis focused on well-characterized compounds. However, the influence of each compound was shared by several components, complicating the interpretation of the PCA model, because more components had to be interpreted to understand the interrelationship between individual compounds and the samples. In the present study the data set size of the originally investigated ^{1}H-NMR data set [8] has been extended to include St. John's wort preparations from several continents, previously investigated [10] by HPLC-PDA. The aim of this study is to be able to interpret the full-resolution ^{1}H-NMR data as well as HPLC-PDA data from PCA models based on the entire collection of samples.

To simplify complex PCA models of data sets based on ^{1}H-NMR spectra and HPLC-PDA profiles, rotations of loadings and scores have been performed. Such rotations can lead to model representations where individual variables are more exclusively related to distinct components rather than being shared across many. Rotations can be performed with techniques such as varimax and quartimax rotation [11]. The use of rotations in multivariate data analysis is not a new approach, and it has been used for decades in some areas, e.g., in psychometrics [12]. However, in natural sciences in general and metabolomics and metabonomics in particular the use of rotations of PCA models is far more limited, although a few recent examples can be found [13], [14], [15], [16]. Traditionally, the use of rotations in PCA modelling has not been strictly needed, because multivariate modelling has usually been performed on fairly simple data. Even though data sets with hundreds or thousands of variables have often been used in chemometrics, the real underlying complexity of the data was usually low, involving either a few independent components, or many but highly correlated variables as in analysis of profiles of electronic or vibrational spectra. Nowadays, data sets such as those arising in metabolomic, metabonomic and metabolite profiling studies have a much higher complexity. Thus, rather than analyzing, e.g., UV spectra profiles spanning the whole variable domain, it is common to study variables represented by separate narrow peaks, like those present in ^{1}H-NMR and mass spectra. This creates the need for additional mathematical tools to simplify interpretation.

In this paper, we present an application of rotated PCA models of ^{1}H-NMR and HPLC-PDA data representing complex natural mixtures.

Materials and Methods

Extracts of St. John's wort preparations

The extracts of twenty-four different commercial preparations of St. John's wort were the same as described elsewhere [10]. Thirteen preparations were formulated as tablets (preparations 1 – 4, 11, 12, 14, 16, 17, 21 – 24), and the remaining as capsules (preparations 5 – 10, 13, 15, 18 – 20). Preparations 1 – 10, 23, and 24 originated from Europe, preparations 11 and 17 from Asia, preparations 12, 13, 15, 16, 18, and 20 – 22 from North America and preparations 14 and 19 from Africa. For one of the brands, two different batches were obtained (preparations 7 and 8). For acquisition of NMR data, the extracts were lyophilized twice with 2 mL of methanol and 90 mL of water and once with 2 mL of D_2O.

NMR experiments

NMR experiments were performed on a Bruker Avance spectrometer (^{1}H resonance frequency of 600.13 MHz) (Bruker BioSpin) using standard Bruker library pulse sequences. 1 D ^{1}H-NMR spectra were recorded using a 5 mm TXI probe. Samples of the extracts (15 mg) were dissolved in 700 μL of DMSO-d_6 (99.8 atom% of deuterium) and transferred into 5 mm NMR tubes. Each sample was prepared in triplicate. For each sample 128 transients were collected using 64 k data points with a spectral width of 16 ppm, using 30° pulses and inter-pulse delay of 4.41 s in order to obtain practically fully relaxed spectra. The spectra were Fourier-transformed to 128 k data points, using line broadening of 0.1 Hz, and referenced to internal TMS.

HPLC-PDA data

The HPLC-PDA data, aligned using an extended algorithm of correlation optimized warping (COW), were the same as described previously [10].

Pre-treatment of ^{1}H-NMR data

^{1}H-NMR data were phased and referenced in Xwin-nmr ver. 3.1 (Bruker BioSpin) and imported into MATLAB ver. 7.0.1 software (MathWorks) for further data pre-treatment and data analysis. A cubic polynomial baseline correction was applied and regions corresponding to residual solvents (δ = 2.48 – 2.54), water (δ = 3.16 – 3.52), residual extraction solvents (δ = 7.36 – 7.40, 7.76 – 7.82, and 8.56 – 8.60), and compounds not considered interesting in relation to this work (fatty acids δ = 0.82 – 0.88, 1.20 – 1.30, and 2.14 – 2.22) were excluded. This exclusion resulted in a data set containing 77,717 variables (NMR descriptors). After exclusion of the specified regions, data were autoscaled using an offset of 500,000 with the following equation:

$$x_{ij}^{autoscaled\ \ with\ offset} = \frac{x_{ij} - \bar{x}_j}{s_j + offset}$$

This offset was chosen based on a visual assessment of the magnitude of variables containing only noise or baseline. The use of an offset prevents these noise areas to have too much influence on the models.

Software for rotation

The rotations were performed using a general rotation tool for PCA models made in MATLAB. The tool is available from www.models.life.ku.dk (September, 2008).

Results and Discussion

Principles of rotation of PCA models

Given a data matrix **X** of size $I \times J$ (objects × variables), principal component analysis is a way of partitioning **X** into a systematic part and a residual (noise) part. The systematic part consists of possibly a few latent variables, i.e., principal components that summarize the largest variance in the data. The projection of I objects in **X** onto the first loading vector $\mathbf{p}_1$ provides the score values of the first component, $\mathbf{t}_1$, which describes maximal variation in the data. Subsequent components are found similarly and describe as much as possible of yet unexplained variation. PCA can be described by $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$ where **T** is the score matrix holding the above score vectors as columns, $\mathbf{P}^T$ is the transposed loading matrix, and **E** is the residuals. The scores and loadings are determined so as to minimize the residuals in a least-squares sense.

For a given PCA model, it is possible to rotate the scores or loadings in the model without affecting the overall fit of the solution if either the loadings or the scores are similarly counter-rotated [17]. Hence, rotation is simply a way to represent the systematic variation differently. The actual variation described by the overall PCA model is not changed.

If **Q** is an $m \times m$ orthogonal matrix, i.e., $\mathbf{Q} \times \mathbf{Q}^T = \mathbf{I}$ and we define $\mathbf{S} = \mathbf{TQ}$ and $\mathbf{M}^T = \mathbf{Q}^T\mathbf{P}^T$ then $\mathbf{TP}^T = \mathbf{TQQ}^T\mathbf{P}^T = \mathbf{SM}^T$. Hence, $\mathbf{X} = \mathbf{SM}^T + \mathbf{E}$ is a model with scores **S** and loadings **M** that are rotated versions of the original ones, and which represents exactly the same fit to the data.

The major challenge when applying a rotation to a PCA model is how to choose the rotation matrix **Q**. From a mathematical point of view, there is an infinite number of ways to define **Q** and different criteria for its choice have been developed [11], [12], [17], [18], [19]. The main principle of these criteria is to rotate towards a simpler structure, i.e., the rotation procedure seeks to establish a simpler relationship within the individual loadings so that these become easier to interpret [20]. An example of a simple structure could be the largest possible loading of a variable in one component, resulting in diminished loadings of the same variable for other components. Hence, samples in the direction of this loading vector can be clearly associated with distinct variables.

As an example, consider a loading matrix, which reads:

$$\mathbf{P} = \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ -\sqrt{2} & \sqrt{2} \end{bmatrix}$$

This is a complicated structure because both original variables contribute significantly in both components (columns of **P**). Rotating this model by a rotation matrix **Q** (which in this case happens to be equal to **P**) yields

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This is a loading matrix with an obviously simple structure, because now every manifest or measured variable is only associated with one latent variable. Thus, rotations are used to obtain another view of the model in which each variable is maximally correlated with one component and reaches a near-zero correlation with other components. The fit of the overall explained variance of the model is unchanged upon the rotation, but the scores and the contribution of explained variance of each component in the PCA model as well as the loadings will inevitably change. In a PCA model, the first component explains the largest fraction of variance and the subsequent components describe progressively smaller fractions. Upon rotation, this is no longer the case.

Rotating the PCA model towards simplicity of scores rather than simplicity of loadings is equally feasible, as follows from the symmetry of the PCA model. However, most studies published so far have used rotation of the PCA model for obtaining simpler loadings [15], [16], [17]. Rotation of scores can be particularly useful when a certain clustering is expected among the samples, as shown in the following paragraphs.

Two general categories of rotations are available, orthogonal and oblique rotations. In the first category, the angular dependence between the original set of loading vectors is preserved (as in the simple example stated above), whereas in the latter category, the angles between loading vectors are not necessarily preserved. The quartimax and varimax criteria are orthogonal rotations, whereas criteria such as oblimin, promax and simplimax represent oblique rotations [19], [21].

One advantage of the orthogonal rotations is that orthogonality makes the numerical approaches simpler and better behaved. A potential drawback could be that orthogonality between loadings is seldom the reality of the underlying features. However, the aim of rotations as presented here is not to find the `true' profiles, but rather to find a mathematical representation that can simplify interpretation. None of the traditional rotation methods, be they orthogonal or not, can provide estimates of real profiles in normal situations. Hence, the choice of rotation method should generally not be guided by a quest for true profiles. If such estimates are sought, then the family of curve-resolution methods is useful. In this study we focus on orthogonal rotations.

Among orthogonal rotations, the quartimax criterion described by Ferguson [11], [17], [22], as well as the varimax criterion described by Kaiser [11], [23] have been commonly described under the orthomax criterion [24]. The principle of the orthomax rotations is to maximize the orthomax criterion given by:

$$V = \sum_{j=1}^{J}\sum_{f=1}^{F} p_{jf}^4 - \frac{\gamma}{J}\sum_{f=1}^{F}\left(\sum_{j=1}^{J} p_{jf}^2\right)^2$$

where pj_f is the loading value for variable j on component f, $j = 1,\ldots,J$ represent variables, and $f = 1,\ldots,F$ represent components; $0 \le \gamma \le 1$. If $\gamma = 0$ the equation becomes the quartimax criterion and if $\gamma = 1$ it becomes the varimax criterion.

The varimax criterion is by far the most often applied method among the orthogonal rotations [13], [14], [15], [16]. Maximizing the varimax criterion provides a solution where the variance of the squared loading elements is maximized. For two competing solutions, the one having a higher varimax criterion value will have optimized loading values in each principal component, i.e., values that are either high (in absolute value) or close to zero. This is a solution that fulfils the definition of a simple structure [11]. On the other hand, maximizing the quartimax criterion

maximizes the variance of each (squared) variable, i.e., optimizes loadings for each variable to a high value in one component and low or zero values in other components. Hence, quartimax minimizes the number of components needed to explain each variable. Kaiser stated that there is a possible bias with the quartimax, as it tends to give one general factor [11]. The varimax rotation principle is the rotation principle applied in this study.

Interpretation of rotated PCA models based on ^{1}H-NMR spectra of St. John's wort extracts

^{1}H-NMR spectroscopy is a non-selective technique that gives unique signals for each hydrogen-containing secondary metabolite above a certain concentration limit. Preparations of St. John's wort are complex mixtures containing many different metabolites, and ^{1}H-NMR spectra of these preparations are very complex and show hundreds of signals.

Interpretation of the derived PCA models at the individual compound level requires assignment of individual resonances of these compounds. Assignment of ^{1}H-NMR spectra of major constituents of extracts of commercial preparations of St. John's wort was performed using 2 D NMR experiments (COSY, TOCSY, *J*-resolved, HSQC and HMBC) with reference to data reported by Bilia et al. [25]. Due to the complexity of the ^{1}H-NMR spectra of the extracts, complete assignments were limited to major constituents. The 2 D NMR results guided the choice of authentic samples used for spiking, performed in order to confirm identification, especially the differentiation between closely related compounds. This led to the assignment of all resonances of chlorogenic acid, rutin, hyperoside, isoquercetin, quercetrin, and quercetin.

PCA models presented in this study are all based on data sets obtained from the full-resolution ^{1}H-NMR spectra (77,717 variables). Using the full spectral resolution rather than binned (integrated) data enhances the interpretation possibilities of derived models. ^{1}H-NMR spectra of natural product extracts often contain signals from several closely related compounds, and the use of integrated data may lead to loss of identity of individual signals and hence to loss of important information [8].

The PCA model of the preprocessed ^{1}H-NMR data used 12 components to explain 93 % of the total variance in data. The number of components was chosen based on the explained variance, and on evaluation of loadings and residuals. 2 D score plots of the first six components are shown in ◉ **Fig. 1**. An excellent separation according to supplier was achieved indicating that considerable differences between the preparations exist. This is likely due to the fact that standardizations according to procedures described in relevant pharmacopoeias [26], [27] only require standardization of a few among many constituents present. The score plots shown in ◉ **Fig. 1** clearly illustrate that it is hardly possible to find any exclusive preparation, i.e., none of the preparations is completely differentiated from the others by means of specific scores and loadings. The individual clustering of preparations shows that the content of all detected hydrogen-containing compounds is different between suppliers. Interpretation of the contributions of individual plant metabolites to the observed clustering is of utmost importance for understanding the patterns displayed in the score plots.

The loadings of the PCA model were subsequently rotated using the varimax criterion, while the scores were counter-rotated. Score plots of the first six components of the rotated PCA model are shown in ◉ **Fig. 2**. It is apparent that the first five components mainly describe features in individual preparations (preparations 14, 9, 4, 17, and 15, respectively), whereas the sixth component describes features in several preparations. Thus, rotation of loadings enabled exclusive clustering of individual preparations by means of specific loadings. This simplifies interpretation, because individual metabolites only influence a few components in the rotated PCA model as opposed to the non-rotated model, where the influences of individual metabolites are partitioned over several components. Moreover, it is also apparent from ◉ **Fig. 2** that the explained variance of each component has changed upon rotation and that the explained variance does not follow component number in a descending order. Nevertheless, the total variance explained by the rotated and the original model is exactly the same.

The loadings derived from the non-rotated as well as the rotated PCA model have been transformed using the reciprocal of the scaling factor for each variable to be able to interpret the loadings of autoscaled data. In ◉ **Fig. 3**, the back-scaled loadings corresponding to the first six components are shown for both models. It is apparent that the loadings of the rotated PCA model are more simple to interpret, e.g., the signal at $\delta = 5.18$ [H-1 of glucose (Glc) in sucrose (Suc), a pharmaceutical excipient] almost exclusively influences the fourth component, and the resonan-

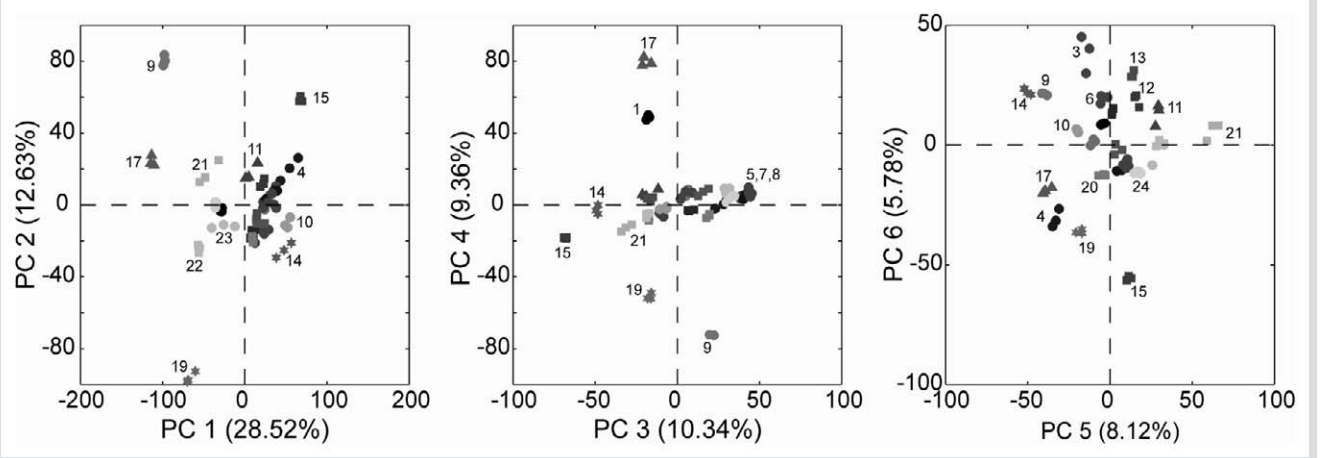

Fig. 1 Score plots of the first six components derived from a PCA model based on ^{1}H-NMR spectra of 24 preparations of St. John's wort. All samples were measured in triplicate.

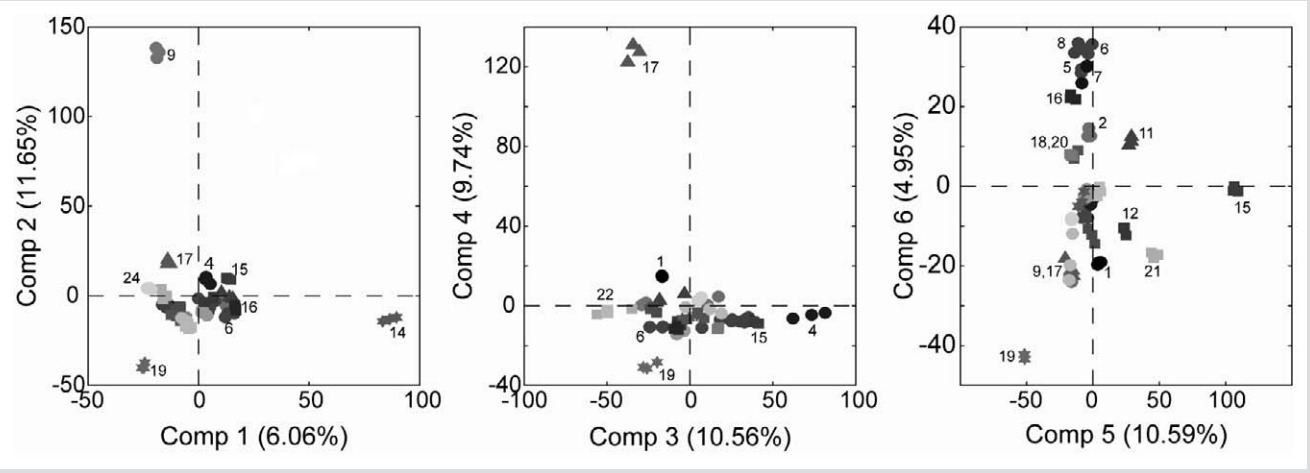

Fig. 2 Score plots of the first six components derived from a rotated PCA model (loadings rotated) based on ^{1}H-NMR spectra of 24 preparations of St. John's wort.

ces of the pharmaceutical excipients, α- and β-lactose (α- and β-Lac) at $\delta = 4.89$ (H-1 Glc, α-Lac), 4.32 (H-1 Glc, β-Lac), 4.20 (H-1 Gal, β-Lac), and 4.18 (H-1 Gal, α-Lac) almost solely influence the second component. Analysis of the same signals in the loadings derived from the non-rotated PCA model reveals that the resonance signal of Suc ($\delta = 5.18$) influences the first six components. The resonance signals of α- and β-Lac ($\delta = 4.89$, 4.32, 4.20, and 4.18) influence the first, second, fourth, fifth, and the sixth component. Thus, interpretation of these resonances only requires analysis of two components when the rotated PCA model is used for interpretation, whereas interpretation of six components is necessary when the non-rotated PCA model is used. In fact, the loadings corresponding to the second and fourth component of the rotated PCA model provide good approximations of real ^{1}H-NMR spectra of Suc and α- and β-Lac, respectively.

Suc and α- and β-Lac are primarily pharmaceutical excipients and not constituents of St. John's wort. The clustering of extracts of commercial preparation of St. John's wort due to the excipients may seem uninteresting. However, this example illustrates the possibility of using rotations of PCA models to obtain unique loadings for outlying samples, which may be a valuable tool for identifying causative sources for outliers. Moreover, in the specific case of medicinal products, this example illustrates that rotations of PCA models can be used to separate clustering due to excipients from that due to genuine constituents of the plant.

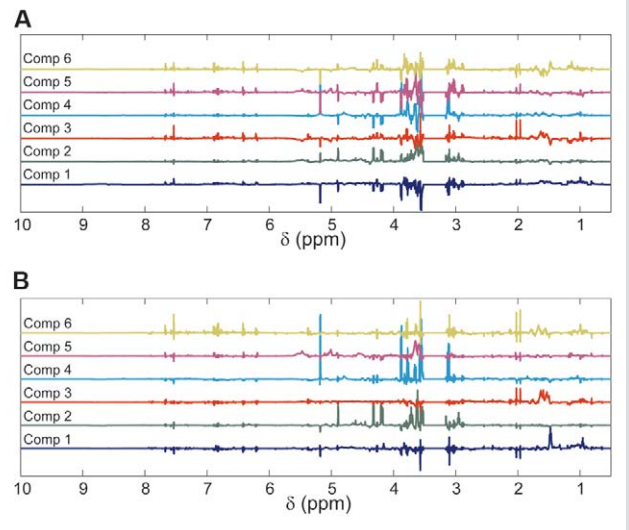

Fig. 3 Back-transformed loadings corresponding to the first six components derived from the PCA model (**A**) and the rotated PCA model (loadings rotated) (**B**).

An interesting observation in ▶ **Fig. 3** is the distribution of the influence of signals in the region around $\delta = 2$. The influence of these signals is distributed over the first, third, fifth, and the sixth component in the original as well as the rotated PCA model. The interpretation of the influence of these signals seems more straightforward using the loadings derived from the original PCA model, since mostly the third component is influenced by these signals, whereas the influence of these signals is equally distributed over the above-mentioned four components in the rotated PCA model. As already mentioned, no change in the overall fit of the model occurs upon rotation and the aim is to obtain a more simple structure with a few high loading values and many small (ideally zero) loading values. The cost can be that some loading elements do not change at all or become even more complex upon rotation, even though the overall representation is simpler. Thus, it is not possible to obtain a perfect description of every element in the matrix without changing the overall fit. Therefore, the application of rotated PCA models should be seen as an additional opportunity rather than a replacement of the original PCA model.

To be able to interpret the influence of individual plant metabolites on the observed clustering, a closer look at the loadings is necessary. The influence of quercetin on the observed clustering has been further analyzed by looking at the H-5' signal of quercetin ($\delta = 6.89$). Loadings corresponding to this signal are shown in ▶ **Fig. 4** for the non-rotated PCA model (▶ **Fig. 4A**) as well as the rotated PCA model (loadings rotated) (▶ **Fig. 4B**).

Comparison of the loadings corresponding to H-5' of quercetin clearly illustrates that interpretation of the influence of quercetin on the observed clustering is facilitated using the rotated loadings (▶ **Fig. 4B**) as compared to the non-rotated loadings (▶ **Fig. 4A**). Interpretation is aided since the influence of quercetin is partitioned over many components in the non-rotated PCA model, whereas in the rotated PCA model the influence of quercetin is described mainly by the sixth and seventh components.

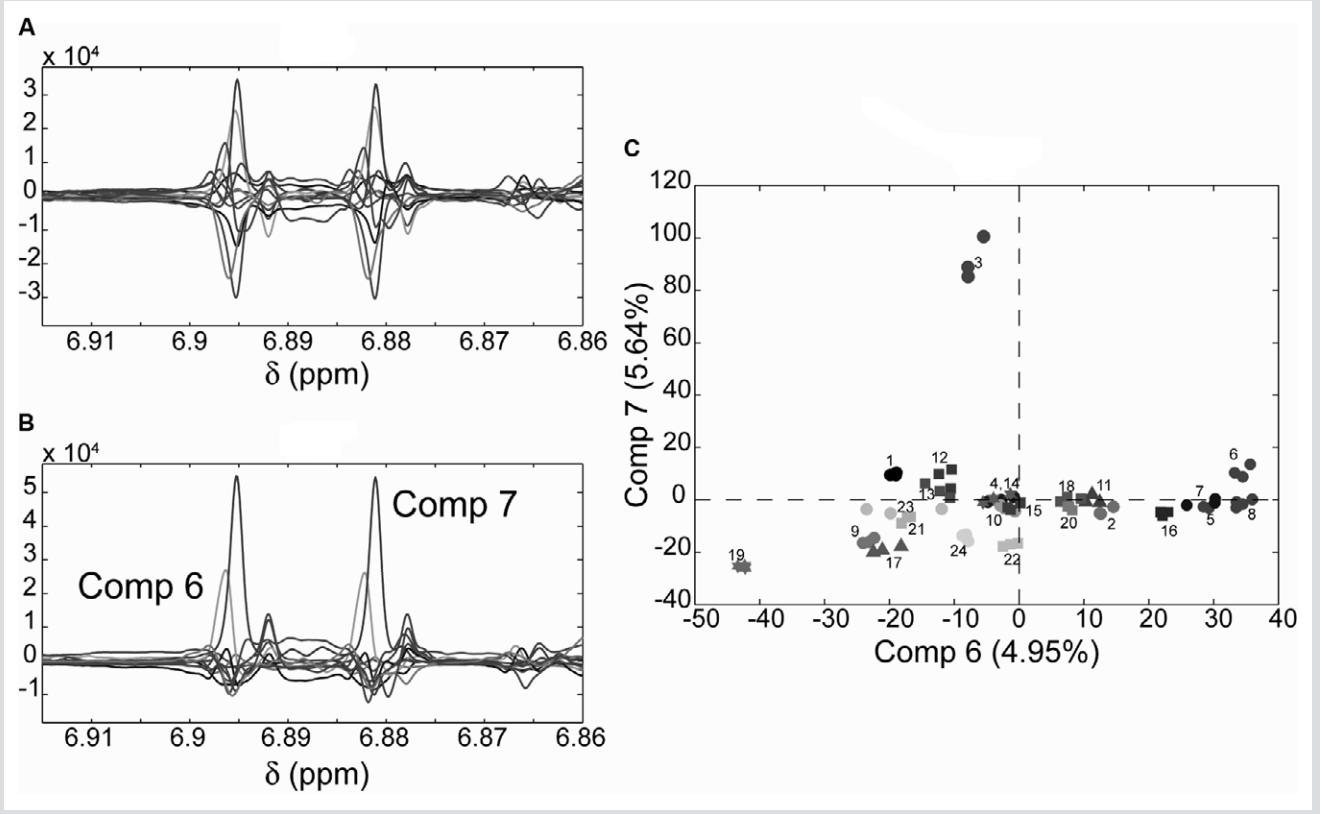

Fig. 4 Back-transformed loadings corresponding to all twelve components derived from the PCA model (**A**) and the rotated PCA model (loadings rotated) (**B**). The score plot of the sixth and seventh component of the rotated PCA model shows a tight clustering of preparations.

Further analysis of the loadings corresponding to the sixth and seventh components reveals that the sixth component is also positively influenced by other flavonoid glycosides (rutin, hyperoside, isoquercetin, quercetrin) and chlorogenic acid. The seventh component only describes the influence of quercetin on the clustering in the positive direction of this component as disclosed by H-5' shown in **Fig. 4** and other resonances of quercetin at $\delta = 7.67$ (H-2'), 7.55 (H-6'), 6.42 (H-8), and 6.19 (H-6) (data not shown). A score plot of the sixth and seventh component of the rotated PCA model is shown in **Fig. 4C**. Analysis of this score plot and the corresponding loadings reveals that the clustering of preparations 2, 5, 6, 7, 8, and 16 in the positive direction of the sixth component of the rotated PCA model is due to higher levels of rutin, hyperoside, isoquercetin, quercetrin, quercetin, and chlorogenic acid, whereas the clustering of preparation 3 in the positive direction of the seventh component is due only to higher levels of quercetin as compared with other preparations.

Interpretation of rotated PCA models based on HPLC-PDA profiles of St. John's wort extracts

To illustrate the simplified interpretation provided by rotated PCA models, an example using an extremely condensed yet comprehensive data matrix will follow. The data matrix contains relative concentrations of St. John's wort plant metabolites derived from PARAFAC analysis of HPLC-PDA profiles. Identification of the plant metabolites represented by the chromatographic peaks was provided by HPLC-PDA-SPE-NMR-MS experiments [10].

Interpretation of the first three components in the derived PCA model has been described in detail in previous work; however, the interpretation of the influence of several plant metabolites involved analysis of several components [10]. The loadings of the derived PCA model were therefore rotated. Score plots of the first four components of the rotated PCA model in association with a loading bar plot are shown in **Fig. 5**. It is apparent that each component explains the influence of individual plant metabolites, and the rotated PCA model facilitates interpretation of the observed clustering.

The first four components of the rotated PCA model describe the influence of guaijaverin, quercetrin, miquelianin, and quercetin 3-*O*-*β*-D-(2-*O*-acetyl)galactoside, respectively (**Fig. 5C**). Thus, the clustering of preparations 14, 17, and 23 in the negative direction of the first component is caused by a higher content of guaijaverin. Preparations 12, 13, 21, and 22, all originating from North America, contain higher levels of quercetrin as compared with other preparations, which cause their separation in the positive direction of the second component (**Fig. 5A**). Higher levels of miquelianin cause the separation of preparations in the positive direction of the third component. In agreement with earlier results [10], higher levels of miquelianin as compared with other preparations influence the clustering of preparations 9 and 23. Preparation 23 displays a more distinct discrimination in the positive direction of the third component (**Fig. 5B**) due to a higher level of miquelianin in this preparation as compared with preparation 9. The influence of quercetin-3-*O*-*β*-D-(2-*O*-acetyl)galactoside on the observed clustering is described in the fourth component. The presence of higher levels of this plant metabolite in preparations 5, 7, 8, 14, and 23

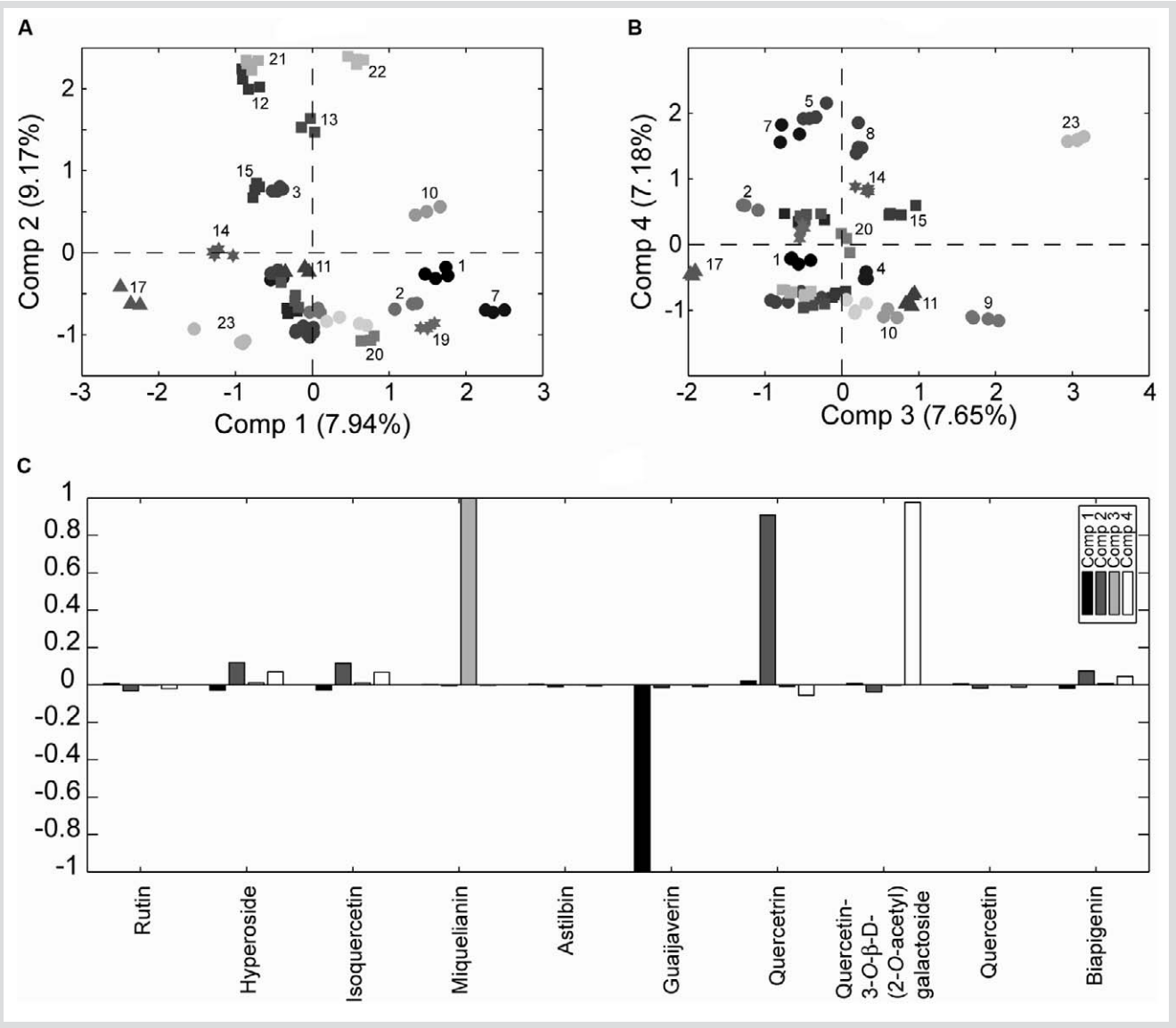

Fig. 5 Score plots of the first four components derived from the rotated PCA model (loadings rotated) (**A** and **B**). Loading bar plot of the corresponding four components is shown in (**C**). The PCA model is based on HPLC-PDA profiles of extracts of preparations of St. John's wort [10].

explains their discrimination from other preparations in the positive direction of the fourth component (**Fig. 5B**).

Rotation of loadings thus eases interpretation of the influence of individual plant metabolites on the observed clustering. If, on the other hand, the aim of the study is to gain knowledge about the plant metabolites influencing the clustering of individual samples, rotation of scores can be a valuable tool. To interpret the influence of metabolic profiles on the clustering of individual samples, rotation of scores from the derived PCA model based on the condensed data matrix with relative concentrations has been performed. This aids interpretation of plant metabolites influencing the clustering of individual samples.

Score plots of the first four components of the rotated PCA model (scores rotated) in association with a bar plot of the corresponding loadings are shown in **Fig. 6**. It is apparent that rotation of scores provides discrimination of individual preparations or closely related preparations. Thus, preparation 17 is discriminated in the negative direction of the first component (**Fig. 6A**), and therefore interpretation of the loadings corresponding to this component provides information about plant metabolites influencing the clustering of this preparation. From the loading bar plot it is seen that guaijaverin and biapigenin influence the clustering of preparation 17 (**Fig. 6C**), in agreement with earlier results [10]. As opposed to the non-rotated PCA model, which also provided discrimination of preparation 17 in the first component, the loadings derived from rotated PCA model (scores rotated) is not confounded by the influence of other preparations. Interpretation of plant metabolites influencing the clustering of preparation 23 required analysis of several components of the non-rotated PCA model [10]. Rotation of scores provides easier interpretation of plant metabolites influencing the clustering of this preparation by analysis of a single component – the third component of the rotated PCA model (scores rotated) (**Fig. 6B**). Thus, plant metabolites influencing the clustering of preparation 23 in the positive direction of the third component are directly seen in the loading bar plot corresponding to this component (**Fig. 6C**). This shows that higher levels of miquelianin, guaijaverin, and quercetin 3-*O*-*β*-D-(2-*O*-acetyl)-

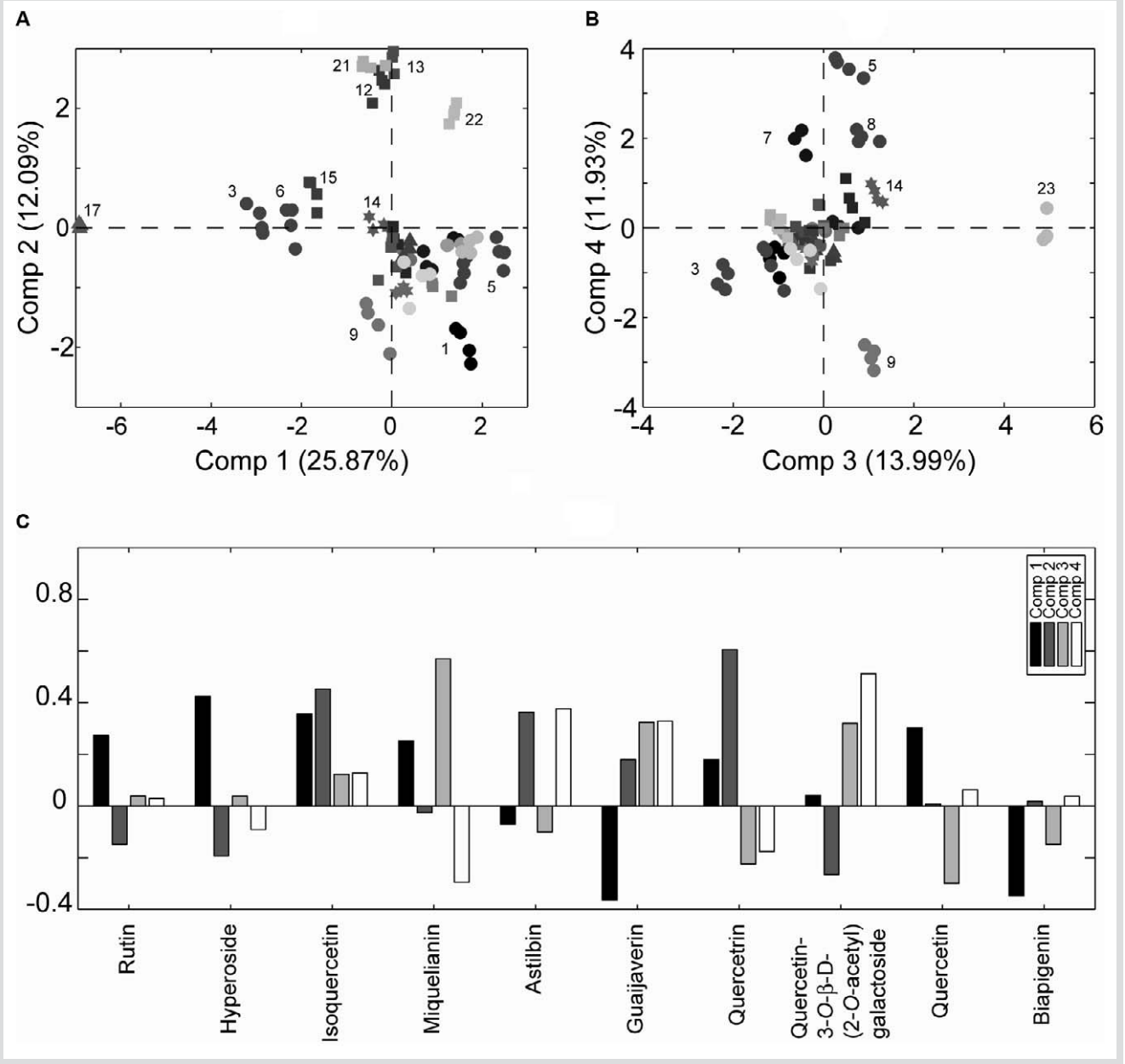

Fig. 6 Score plots of the first four components derived from the rotated PCA model (scores rotated) (**A** and **B**). Loading bar plot of the corresponding four components is shown in (**C**). The PCA model is based on HPLC-PDA profiles of extracts of preparations of St. John's wort [10].

galactoside, and to a minor degree also higher levels of rutin, hyperoside, and isoquercetin, are responsible for the observed clustering of preparation 23, in agreement with earlier results [10].

In conclusion, this study has illustrated the advantages of using rotated PCA models for aiding interpretation of PCA models derived from ^{1}H-NMR spectra as well as from HPLC-PDA profiles of herbal remedies. Rotation of loadings led to simpler visualizations in terms of interpretation of the influence of individual metabolites on the observed clustering, since the number of components influenced by individual metabolites was reduced as compared to the non-rotated PCA model. For the ^{1}H-NMR data, only a few components of the rotated PCA model described the influence of quercetin, whereas for the HPLC-PDA data each component of the rotated PCA model described the influence of an individual plant metabolite. Rotation of scores of the PCA model derived from the HPLC-PDA data set led to conditions, where the whole plant metabolite profiles that are characteristic for individual preparations could be derived from the rotated PCA model. This approach is especially valuable for understanding the clustering of individual preparations or groups of clusters. Rotation of PCA models illustrated in this study is believed to have general applicability in metabonomic, metabolomic, and metabolite profiling studies.

Acknowledgements

Bruker Avance 600 spectrometer used in this work was acquired through a grant from "Apotekerfonden af 1991" (Copenhagen). The technical assistance of Ms. Birgitte Simonsen and Ms. Dorte Brix is gratefully acknowledged.

References

1 *Martens H, Næs T.* Multivariate calibration. New York: Wiley; 1989
2 *Bailey NJ, Sampson J, Hylands PJ, Nicholson JK, Holmes E.* Multi-component metabolic classification of commercial feverfew preparations via high-field ^{1}H-NMR spectroscopy and chemometrics. Planta Med 2002; 68: 734 – 8
3 *Choi YH, Kim HK, Hazekamp A, Erkelens C, Lefeber AWM, Verpoorte R.* Metabolomic differentiation of *Cannabis sativa* cultivars using ^{1}H NMR spectroscopy and principal component analysis. J Nat Prod 2004; 67: 953 – 7
4 *Frédérich M, Choi YH, Angenot L, Harnischfeger G, Lefeber AWM, Verpoorte R.* Metabolomic analysis of *Strychnos nux-vomica*, *Strychnos icaja* and *Strychnos ignatii* extracts by ^{1}H nuclear magnetic resonance spectrometry and multivariate analysis techniques. Phytochemistry 2004; 65: 1993 – 2001
5 *Wang Y, Tang H, Nicholson JK, Hylands PJ, Sampson J, Whitcombe I.* Metabolomic strategy for the classification and quality control of phytomedicine: A case study of chamomile flower (*Matricaria recutita* L.). Planta Med 2004; 70: 250 – 5
6 *Kim HK, Choi YH, Erkelens C, Lefeber AWM, Verpoorte R.* Metabolic fingerprinting of *Ephedra* species using ^{1}H NMR spectroscopy and principal component analysis. Chem Pharm Bull 2005; 53: 105 – 9
7 *Holmes E, Tang HR, Wang YL, Seger C.* The assessment of plant metabolite profiles by NMR-based methodologies. Planta Med 2006; 72: 771 – 85
8 *Rasmussen B, Cloarec O, Tang HR, Stærk D, Jaroszewski JW.* Multivariate analysis of integrated and full-resolution ^{1}H-NMR spectral data from complex pharmaceutical preparations: St. John's wort. Planta Med 2006; 72: 556 – 63
9 *Seger C, Sturm S.* Analytical aspects of plant metabolite profiling platforms: Current standings and future aims. J Proteome Res 2007; 6: 480 – 97
10 *Schmidt B, Jaroszewski JW, Bro R, Witt M, Stærk D.* Combining PARAFAC analysis of HPLC-PDA profiles and structural characterization using HPLC-PDA-SPE-NMR-MS experiments: Commercial preparations of St. John's wort. Anal Chem 2008; 80: 1978 – 87
11 *Kaiser HF.* The Varimax criterion for analytic rotation in factor-analysis. Psychometrika 1958; 23: 187 – 200
12 *Kiers HAL.* A comparison of techniques for finding components with simple structure. In: Cuadras CMC, Rao CR, editors. Multivariate analysis: future directions 2 Amsterdam: Elsevier; 1993: 67 – 86
13 *Bundy JG, Sidhu JK, Rana F, Spurgeon DJ, Svendsen C, Wren JF.* "Systems toxicology" approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm, *Lumbricus rubellus.* BMC Biol 2008; 6: 25
14 *Rasmussen S, Parsons AJ, Fraser K, Xue H, Newman JA.* Metabolic profiles of *Lolium perenne* are differentially affected by nitrogen supply, carbohydrate content, and fungal endophyte infection. Plant Physiol 2008; 146: 1440 – 53
15 *Soares PK, Bruns RE, Scarminio IS.* Statistical mixture design – Varimax factor optimization for selective compound extraction from plant material. Anal Chim Acta 2008; 613: 48 – 55
16 *Stojanovic K, Jovancicevic B, Vitorovic D, Golovko Y, Pevneva G, Golovko A.* Evaluation of saturated and aromatic hydrocarbons oil-oil maturity correlation parameters (SE Pannonian Basin, Serbia). J Serb Chem Soc 2007; 72: 1237 – 54
17 *Kiers HAL.* Simple structure in component analysis techniques for mixtures of qualitative and quantitative variable. Psychometrika 1991; 56: 197 – 212
18 *Cattell RB.* "Parrallel proportional profiles" and other principles for determining the choice of factors by rotation. Psychometrika 1944; 9: 267 – 83
19 *Malinowski ER.* Factor analysis in chemistry, 3 rd edition. New York: Wiley-Interscience; 2002
20 *Dien J, Beal DJ, Berg P.* Optimizing principal components analysis of event-related potentials: Matrix type, factor loading weighting, extraction, and rotations. Clin Neurophysiol 2005; 116: 1808 – 25
21 *Kiers HAL.* Simplimax – oblique rotation to an optimal target with simple structure. Psychometrika 1994; 59: 567 – 79
22 *Ferguson GA.* The concept of parsimony in factor analysis. Psychometrika 1954; 19: 281 – 90
23 *ten Berge JMF.* Suppressing permutations or rigid planar rotations: A remedy against nonoptimal varimax rotations. Psychometrika 1995; 60: 437 – 46
24 *Harman HH.* Modern factor analysis. 3 rd edition. Chicago: University of Chicago Press; 1976
25 *Bilia AR, Bergonzi MC, Mazzi G, Vincieri FF.* Analysis of plant complex matrices by use of nuclear magnetic resonance spectroscopy: St. John's wort extract. J Agric Food Chem 2001; 49: 2115 – 24
26 European Pharmacopoeia, 5th edition. Strassbourg: Council of Europe; 2005: 2485 – 6
27 United States Pharmacopeia, USP 30: The National Formulary 25. Rockville: The United States Pharmacopeial Convention; 2007: 978 – 9