Talanta 115 (2013) 590–599

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

High-dimensional nested analysis of variance to assess the effect of production season, quality grade and steam pasteurization on the phenolic composition of fermented rooibos herbal tea

I. Stanimirova [a], M. Kazura [a], D. de Beer [b], E. Joubert [b,c], A.E. Schulze [c], T. Beelders [c], A. de Villiers [c], B. Walczak [a,*]

[a] Department of Analytical Chemistry, Chemometric Research Group, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Post-Harvest and Wine Technology Division, Agricultural Research Council (ARC), Infruitec-Nietvoorbij, Private Bag X5026, Stellenbosch 7599, South Africa
[c] Departments of Food Science, and Chemistry and Polymer Science, Stellenbosch University, Private Bag X1, Matieland, Stellenbosch 7602, South Africa

ARTICLE INFO

Article history:
Received 27 February 2013
Received in revised form
10 June 2013
Accepted 14 June 2013
Available online 25 June 2013

Keywords:
Chemometrics
Analysis of variance-simultaneous component analysis
Mixed-effects modeling
Antioxidant capacity
Phenolic content

ABSTRACT

A nested analysis of variance combined with simultaneous component analysis, ASCA, was proposed to model high-dimensional chromatographic data. The data were obtained from an experiment designed to investigate the effect of production season, quality grade and post-production processing (steam pasteurization) on the phenolic content of the infusion of the popular herbal tea, rooibos, at 'cup-of-tea' strength. Specifically, a four-way analysis of variance where the experimental design involves nesting in two of the three crossed factors was considered.

For the purpose of the study, batches of fermented rooibos plant material were sampled from each of four quality grades during three production seasons (2009, 2010 and 2011) and a sub-sample of each batch was steam-pasteurized. The phenolic content of each rooibos infusion was characterized by high performance liquid chromatography (HPLC)-diode array detection (DAD). In contrast to previous studies, the complete HPLC-DAD signals were used in the chemometric analysis in order to take into account the entire phenolic profile.

All factors had a significant effect on the phenolic content of a 'cup-of-tea' strength rooibos infusion. In particular, infusions prepared from the grade A (highest quality) samples contained a higher content of almost all phenolic compounds than the lower quality plant material. The variations of the content of isoorientin and orientin in the different quality grade infusions over production seasons are larger than the variations in the content of aspalathin and quercetin-3-*O*-robinobioside. Ferulic acid can be used as an indicator of the quality of rooibos tea as its content generally decreases with increasing tea quality. Steam pasteurization decreased the content of the majority of phenolic compounds in a 'cup-of-tea' strength rooibos infusion.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Typically, the analysis of food and beverages involves a qualitative or quantitative determination of specific chemical components in order to explain changes that occur during processing and loss of quality, amongst others. Recently, the analysis of herbal chromatographic fingerprints and specifically a non-targeted approach has been advocated and applied [1]. The main aims are generally quality control of herbal products and elucidation of the role of various factors on composition taking into account minor and/or unidentified constituents.

The aim of this work was to evaluate the significance of the effect of production season, quality grade and post-production processing (steam pasteurization) on the phenolic content of the infusion of the popular herbal tea, rooibos [2], prepared at 'cup-of-tea' strength using chromatographic fingerprints. Several chemometric methods that combine the statistical properties of analysis of variance (ANOVA) and the advantages of a dimensionality reduction technique like principal component analysis (PCA), ANOVA–PCA or simultaneous component analysis (SCA), ASCA for handling highly correlated multivariate data have been described in the literature [3–6]. To date, all of these supervised chemometric methods have mainly been used to analyze data with fixed effects factors [7]. In this paper, we present a

* Corresponding author. Tel.: +48 32 359 1219; fax: +48 32 259 9978.
E-mail address: beata.walczak@us.edu.pl (B. Walczak).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.06.023

methodology based on the ASCA method which allows us to deal with a high dimensional mixed-effects analysis of variance with nesting. Furthermore, a permutation strategy for a test the effect of each particular factor is also presented.

The phenolic composition of rooibos plays an important role in its sensory (color, taste and astringency) and health-promoting properties [8–11]. Joubert et al. [12] have studied the effect of production season and quality grade on the content of 15 selected phenolics of rooibos infusions, which were prepared from steam-pasteurized fermented plant material and showed that the content of some individual phenolic compounds significantly depends on the production season and quality grade. Koch et al. [9] determined that steam pasteurization reduced the color and astringency of fermented rooibos infusions, yet no significant change in the major phenolic compounds with the exception of aspalathin, was observed, indicating that other compounds such as polymers may be important. In the present study, a more comprehensive set containing data from the same rooibos samples used in Joubert et al. [12] and data from sub-samples that were not pasteurized was analyzed using the methodology based on the ASCA approach in order to consider all of the possible sources of variations related to production season, quality grade and steam pasteurization. Specifically, a four-way ANOVA where the experimental design involves nesting in two of the three crossed factors is considered. In contrast with the previous study, which focused on selected compounds, the presented methodology allowed us to work with data in which each sample was characterized by its complete HPLC-DAD signal in order to take into account the entire phenolic profile.

2. Experimental section

2.1. Design of experiments

Individual production batches of unrefined, fermented rooibos plant material, originating from several locations (i.e. farms and areas), were sampled during three production seasons (2009, 2010 and 2011) and classified into four quality grades, A, B, C and D. The quality grading, based on sensory properties, was performed in-house by a rooibos processing and marketing company in accordance with their quality grading system as described by Koch et al. [9]. Samples of 10 batches were then randomly selected from each quality grade material in order to guarantee the representativeness of the samples. Samples were treated as described by Joubert et al. [12]. In short, the samples were sieved to obtain the refined plant material that is sold to the consumer, whereafter a portion of the refined fraction >40 mesh and <10 mesh from each sample was subjected to steam pasteurization following basically the same conditions as those used by the industry (steam exposure ± 96 °C for 2 min followed by drying to ca 10% moisture content). Next, duplicate infusions filtered through Whatman no. 4 filter paper (Whatman International, Ltd., Maidstone, U.K.) were prepared from each sample (both the unpasteurized and pasteurized material) at 'cup-of-tea' strength and aliquots of ca. 1.5 mL were stored at −20 °C. Finally, the HPLC-DAD analysis of each infusion was performed in duplicate.

The design of the experiment is presented schematically in Fig. 1.

A total of 920 HPLC-DAD measurements were recorded for all samples. Each 'X' in Fig. 1 represents one HPLC-DAD measurement registered for one of the two injections from one of the two infusions prepared for each pasteurized or unpasteurized sub-sample. Specifically, 320 measurements were performed during each of the first two production seasons (2009 and 2010). Only 5 sample batches of grade D rooibos plant material were collected during 2011, and therefore, 40 measurements for those infusions are missing (those marked with 'o' in Fig. 1). For the first two seasons, 160 measurements were characteristic for pasteurized plant material and 160 measurements were recorded for infusions of unpasteurized plant material, while only 140 measurements were performed for each of the two processing conditions for the last production season. Considering two replicate measurements for each rooibos infusion originating from each of the ten sample batches and a definite quality grade (A, B, C or D), a total of 40 measurements were recorded for each type (unpasteurized or pasteurized) of plant material per year, with the exception of grade D samples in 2011 for which a total of 20 measurements per type were recorded.

Such a design of the experiments allowed us to evaluate the effects of three main factors–the production season, pasteurization and quality grade–on the phenolic content of rooibos infusion. Sample batches are nested within all of the combinations of the levels of the quality grade and production season factors. Steam pasteurization and quality grade are fixed factors, while the production season and sample batches are considered to be random factors.

2.2. HPLC-DAD methodology

HPLC analysis was carried out as described by Beelders et al. [13] using an Agilent 1200 system (Agilent, Santa Clara, CA) equipped with DAD (standard 13 μL flow cell and 10 mm path length) to record the UV spectra between 200 and 700 nm. Separation was performed at 37 °C on a 100×4.6 mm^2, 1.8 μm Agilent Zorbax SB-C18 column protected with an Acquity ultra-performance liquid chromatography (UPLC) in-line filter (Waters, Milford, MA) and a 5.0 μm SB-C18 guard

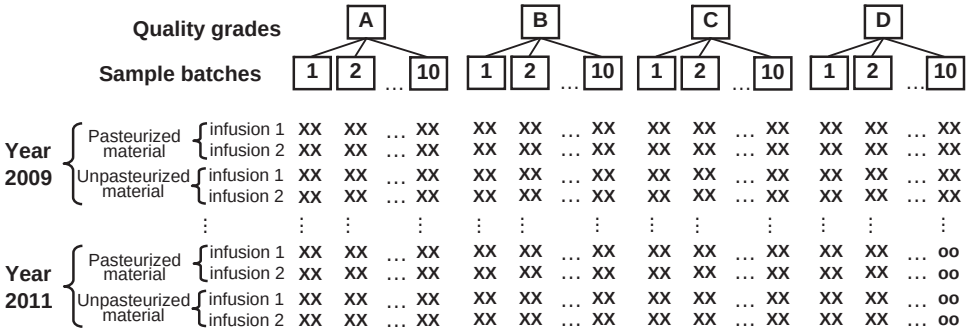

Fig. 1. Scheme of the experimental design with four factors, e.g. production season, steam pasteurization, quality grade (A, B, C and D) and sample batches (1, 2, …, 10) that are nested within the levels of the quality grade and production season factors. Each 'X' represents one HPLC-DAD measurement, while 'o' represents a missing measurement.

column (Agilent). The flow rate was 1.0 mL/min, and a multilinear gradient was performed as follows: 10% B (0–2 min), 10–14.8% B (2–19 min), 14.8–36.8% B (19–34 min), 36.8–100% B (34–37 min), 100% B isocratic (37–42 min), 100–10% B (42–45 min) and 10% B (45–50 min), with two solvents, 2% (*m*/*v*) acetic acid in water and acetonitrile, respectively. The validation and optimization of the HPLC-DAD method, the details of which are presented elsewhere [12], showed that the values obtained for the relative standard deviation (RSD) for intra- and inter-day precision, as well as sample stability tests, were lower than 5%, which were deemed to be acceptable for this study.

3. Theory

The general framework for the analysis of highly structured multidimensional data consists of (i) selecting a model in accordance with the provided design of the experiments for the problem being studied, (ii) evaluating factor effects, collecting these in respective factor matrices and significance testing of the effects, and (iii) use of a dimensionality reduction technique to interpret and visualize the information in the individual factor matrices. This general scheme described by Smilde et al. [14] for a high dimensional fixed-effects analysis of variance was generally followed in this study for four-way analysis with nesting in two of the three crossed factors.

3.1. Selection of a model in accordance with the provided design of experiments

To simplify the description, let us assume a balanced experimental design with four factors (Fig. 1) in which there are I ($i=1, 2, \ldots, I$) independent rooibos tea samples, each of which is described by a chromatogram with J ($j=1, 2, \ldots, J$) elution time measurement points. The chromatogram for the i-th sample of the original data organized as matrix $\mathbf{X}$ ($I \times J$) is recorded at level k ($k=1, 2, \ldots, K$) of the first factor (production season) with effect α, at level l ($l=1, 2, \ldots, L$) of the second factor (quality grade) with effect β, at level n ($n=1, 2, \ldots, N$) of the third factor (sample batches) with effect γ and at level t ($t=1, 2, \ldots, T$) of the fourth factor (steam pasteurization) with effect δ. The first, second and fourth factors are crossed because every level in one of the factors occurs at every level of the other two factors. The third factor of sample batches is nested within all of the combinations of the levels of the 'production season' and 'quality grade' factors, but crossed with the 'steam pasteurization' factor. The first and third factors are random, which implies that the interactions containing these factors are also considered to be random. Thus, the linear four-way nested analysis of variance (ANOVA) model for the j-th elution time measurement recorded for the i-th rooibos tea sample at $kln t$ levels of the first, second, third and fourth factors, $x_{i_{kln\ t}j}$ of $\mathbf{X}$, can be written as follows:

$$x_{i_{kln\ t}j} = \mu_j + \alpha_{kj} + \beta_{lj} + (\alpha\beta)_{klj} + \gamma_{n(kl)j} + \delta_{tj} + (\alpha\delta)_{ktj} + (\beta\delta)_{ltj} + (\alpha\beta\delta)_{kltj} + (\gamma\delta)_{n(kl)tj} + \varepsilon_{i_{kln\ t}j} \quad (1)$$

where μ_j is the mean value calculated for the j-th elution time point and $\varepsilon_{i_{kln\ t}j}$ is the error term. In this model, there are three main one-factor effects, three simple two-factor interaction effects and one simple three-factor interaction effect. These seven effects are the most interesting in the study. The third factor that is nested within the levels of the first and second factors additionally adds one nested interaction effect to the model. Thus, in comparison with the full four-way factorial model, the effects $\gamma_{nj}, (\alpha\gamma)_{knj}, (\beta\gamma)_{\ln j}$ and $(\alpha\beta\gamma)_{kln\ j}$ are replaced by the nested main effect $\gamma_{n(kl)j}$ and the effects $(\gamma\delta)_{ntj}, (\beta\gamma\delta)_{nltj}, (\alpha\gamma\delta)_{kntj}$ and $(\alpha\beta\gamma\delta)_{knltj}$ are replaced by the nested interaction effect $(\gamma\delta)_{n(kl)tj}$. This means that the fifteen effects in the factorial model are reduced to nine effects including the nested sample batches effect.

The model described by Eq. (1) is not uniquely defined without considering some constraints. In this study, we assumed an unrestricted mixed model, which means that the sum of the individual effects over the levels of a fixed factor are equal to zero, while the sum of the mixed interaction effects over the levels of the respective fixed effects are not equal to zero.

3.2. Evaluation of factor effects and significance testing

Next, the contributions to the variation of all I samples described by J elution time points can be collected in the respective effect matrices of dimensions $I \times J$ and this can be presented as:

$$\mathbf{X} = \mathbf{M} + \mathbf{X}_\alpha + \mathbf{X}_\beta + \mathbf{X}_{\alpha\beta} + \mathbf{X}_\gamma + \mathbf{X}_\delta + \mathbf{X}_{\alpha\delta} + \mathbf{X}_{\beta\delta} + \mathbf{X}_{\alpha\beta\delta} + \mathbf{X}_{\gamma\delta} + \mathbf{X}_\varepsilon \quad (2)$$

The decomposition defined by Eq. (1) is applied to each elution time point (column) of the original $\mathbf{X}$ data matrix. Hence each matrix contains as many groups of identical mean vectors as the levels of the respective factor. The mean matrix, $\mathbf{M}$, comprises I identical mean vectors, the elements of which are the respective column means of $\mathbf{X}$. After subtracting the corresponding column means from each original element, e.g. after performing centering ($\mathbf{X}-\mathbf{M}$), the elements corresponding to the first factor matrix can be estimated as mean vectors of the K groups. This means that the members of the k-th group in $\mathbf{X}_\alpha$ are represented by the corresponding row mean vector calculated on the centered matrix. After deflation, the elements of the following effect matrix $\mathbf{X}_\beta$ are evaluated as mean vectors for L groups. Similarly, the elements of all effect matrices are calculated. Finally, the individual sample variation is collected in $\mathbf{X}_\varepsilon$ matrix. The partition of the total variation of original data is performed so that the column spaces of all effect matrices are orthogonal. Therefore, the decomposition of the sum of squares of $\mathbf{X}$, SS_{total}, can be expressed as follows:

$$SS_{total} = SS_M + SS_\alpha + SS_\beta + SS_{\alpha\beta} + SS_\gamma + SS_\delta + SS_{\alpha\delta} + SS_{\beta\delta} + SS_{\alpha\beta\delta} + SS_{\gamma\delta} + SS_\varepsilon \quad (3)$$

The key point of this methodology is the significance test of the effect being investigated. A null hypothesis is defined for each effect as for which there are no differences in the chromatograms recorded for different conditions of a factor. Using the statistical terminology, one tests the equality of several population means, where the populations correspond to the groups defined by the levels of a factor. When using the classic fixed-effects ANOVA model [15], the null hypothesis is true when the ratio of the sum of squares of a given factor with the respective degrees of freedom (MS, mean squares) to error mean squares is smaller than the critical value of the F distribution defined for a given level of significance. An exact F-ratio denominator cannot be defined for mixed models and designs with random effects. The F-ratios in such cases are only approximately F distributed and in order to construct an error mean square with the proper components of variation, some mean square terms are added or subtracted. The reason is that when there are random factors, there is no single source of variation to be defined in the denominator, but there are confounded sources of variation contributing to the variation of a specific combination of factor levels. These complex F-ratios are usually referred to as quasi-F-ratios in the literature [16]. In general, to construct a proper test for any effect and interaction, the lowest order interaction effect that contains all of the factors in that effect should be considered as an error variance in the denominator, while the nominator contains the mean squares for the effect being tested. Accordingly, the degrees of freedom for the denominator, $df_{denominator}$, should be adjusted to these additions

and subtractions of m mean squares [17] as follows:

$$df_{denominator} = \frac{(MS_1 \pm \ldots \pm MS_m)^2}{(MS_1^2/df_1) + \ldots + (MS_m^2/df_m)} \quad (4)$$

Considering analysis with classic ANOVA, one assumes that the errors are normally distributed. This assumption is often not fulfilled for real multivariate chemical data as the one presented in this study. Therefore, the use of permutation tests for significance testing of the effect terms is highly valued. Even though normality of the errors is not a required assumption when using a permutation test, the assumption that errors are identically and independently distributed (i.i.d.) is still valid. Reviewing the literature, a few works investigating the relevant strategies for an approximate permutation testing of individual factor effects in multiple linear regressions have been published [18–21]. The results presented therein suggested that permutation of the residuals under the reduced model is the best strategy to be used in comparison with the permutation tests of the full-model residuals and unrestricted permutations of raw data [21]. In our study, a permutation test of the residuals under the reduced model was used for each term of the ANOVA model defined by Eq. (1). The terms of the ANOVA model that should be removed by subtracting the respective group means when testing a particular factor effect are presented in Table 1.

The null hypothesis for the tested effect is accepted when the sum of the squares for the experimental data is smaller than the sum of the squares for the permuted data for most of the cases, indicating that there are no significant differences among group means. The level of significance (P-value) is determined as the number of permutations in which the sum of squares is larger than the true experimental value of the sum of squares.

3.3. Interpretation and visualization of information in the individual effect matrices using a dimensionality reduction technique

After the significance testing, analysis of each effect matrix (Eq. (2)) of a significant factor is required in order to interpret and visualize the relationship between the groups associated with the effect and chemical information obtained from chromatographic analysis. As mentioned earlier, each effect matrix is of dimensions $I \times J$, and contains as many groups of identical row vectors as the levels defined by the factor studied. To observe the distribution of samples and to interpret their relation to variables (elution times), a dimensionality reduction technique like principal component analysis, PCA [22], or simultaneous component analysis, SCA [23], can be employed. Under orthogonality constraints, the results of PCA and SCA are the same. Both methods decompose an effect matrix into two matrices, a scores matrix $\mathbf{T}$ $(I \times f)$ associated with the projections of samples in the space spanned by f orthogonal vectors called principal components, PCs, and a loadings matrix $\mathbf{P}$ $(J \times f)$ associated with the projections of variables onto f PCs. These principal components are obtained by maximizing the variance of the projected data. For example, the bi-linear model for the effect matrix, $\mathbf{X}_\alpha$, of the first factor (production season) can be written as follows:

$$\mathbf{X}_\alpha = \mathbf{T}_\alpha \mathbf{P}_\alpha^T + \mathbf{E}_\alpha \quad (5)$$

In this sub-model $\mathbf{E}_\alpha$ is the error term associated with this factor. Even though $\mathbf{T}_\alpha$ is of dimensions $I \times f$, it contains K groups of identical projections for K production seasons. The rank of $\mathbf{X}_\alpha$ matrix is equal to the number of levels of the first factor minus 1, $K-1$, indicating the number of PCs in this model. Bi-linear models for the effect matrices of the other significant factors can be constructed in the same manner and the projections of group samples and variables are displayed on the so-called score and loading plots.

Table 1
Four-way ANOVA where the experimental design involves nesting in two of the three crossed factors.

Effect in the ANOVA model to be tested	Type of the effect	Permutation test on reduced model
		Terms of the model that should be removed by subtraction of group means
α	Random	β, δ, $\beta\delta$
β	Fixed	α, δ, $\alpha\delta$
$\alpha\beta$	Random	α, β, δ, $\beta\delta$
γ	Random	α, β, $\alpha\beta$, δ, $\alpha\delta$, $\beta\delta$, $\alpha\beta\delta$, $\delta\gamma$
δ	Fixed	α, β, $\alpha\beta$, γ
$\alpha\delta$	Random	α, β, $\alpha\beta$, γ, δ, $\beta\delta$
$\beta\delta$	Fixed	α, β, $\alpha\beta$, γ, δ, $\alpha\delta$
$\alpha\beta\delta$	Random	α, β, $\alpha\beta$, γ, δ, $\alpha\delta$, $\beta\delta$, $\delta\gamma$
$\gamma\delta$	Random	α, β, $\alpha\beta$, γ, δ, $\alpha\delta$, $\beta\delta$, $\alpha\beta\delta$

4. Results and discussion

The first step undertaken before the analysis of the chromatographic data was the preprocessing of all 920 HPLC-DAD signals. The intensity of the 2D signal for each sample recorded at 102 wavelengths (from 298 to 500 nm) and 12,000 elution time measurement points (from 1.70 min to 40.00 min) was considered. An example of a HPLC-DAD signal is presented in Fig. 2a.

A single mean chromatogram for each sample was then constructed by averaging the intensities across the wavelengths as shown in Fig. 2b. Thus, each sample of the whole data set of 920 was described by a mean chromatogram with 12,000 elution time points. From Fig. 2c which displays all 920 mean chromatograms, it is evident that the peaks are shifted along the elution time axis. Therefore, a peak shift correction was required. The Automated Alignment (AA) algorithm as described in [24] and Correlation Optimized Warping (COW) [25] were adopted for this purpose. The main idea in both algorithms is to warp all signals by maximizing the correlation coefficient between a selected target mean chromatogram and each of the remaining mean chromatographic profiles. First, the target mean chromatogram was found among all mean chromatograms as the one having the largest mean value of the correlation coefficient with all of the studied mean chromatograms [26]. Then the remaining signals were warped using the AA algorithm optimizing the number of spline basis functions in the range of 10–20. Even though most of the signals were warped well with the AA algorithm, some of the signals required further warping. COW warps two signals by piecewise linear stretching and compression. Using COW, the number of segments (or the section length of a signal), N, and the slack parameter, s, should be optimized. The positions of the end points of each section are in the range of $-s$ and $+s$ for a given s value. Next, the asymmetric penalized least squares algorithm described by Eilers [27] was applied for background estimation and elimination. The first derivatives of signals were used to trace signal smoothness and the λ parameter was set to 10,000.

In order to avoid an analysis of unbalanced data, the missing 40 mean chromatographic signals for five sample batches of grade D collected in 2011 were estimated using the mean chromatograms of the respective replicate measurements from the five sample batches and assuming the same level of noise (uncertainty estimation). The elution time points between 1.70 and 5.00 min were removed from the warped mean chromatograms. Within this time range the elution of ascorbic acid, added to all samples to

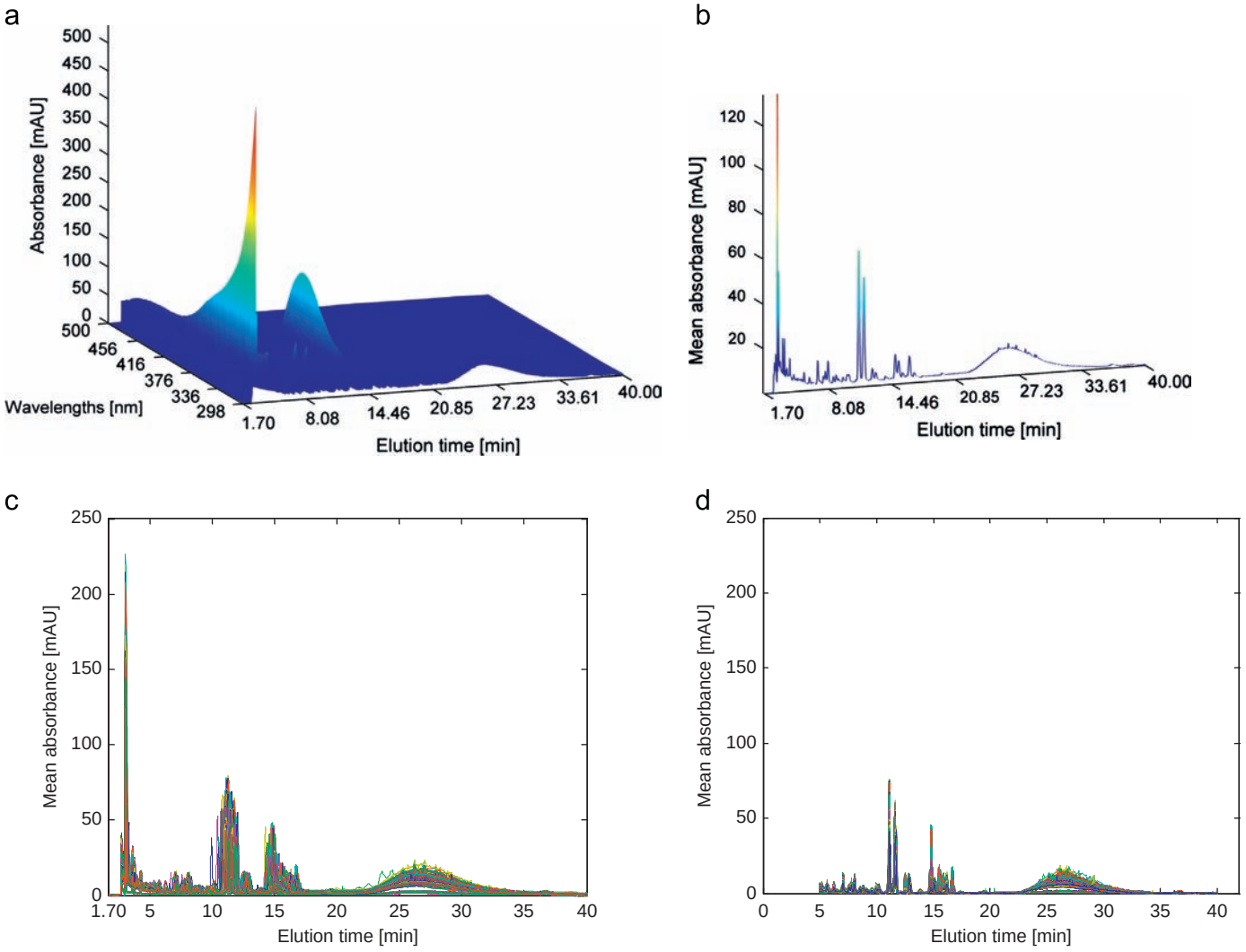

Fig. 2. Data representations (a) A HPLC-DAD signal for one sample, (b) mean chromatogram in which the mean absorbance is obtained by averaging the absorbance over the wavelengths, (c) a total of 920 unaligned mean chromatograms and (d) all 960 mean chromatograms (elution time from 5.00 to 40.00 min.) after warping and baseline correction.

prevent phenolic compound degradation during the time in the HPLC autosampler, is observed and thus, this range was considered as containing unimportant information for this study. The final data set is of the dimensions 960 × 10,967 (Fig. 2d).

As described earlier, the experiments were designed in order to evaluate the effects of production season ($K=3$), quality grade ($L=4$) and steam pasteurization ($T=2$) on the phenolic content of rooibos infusion at 'cup-of-tea' strength. The third factor, e.g. sample batches, is nested within the levels of the production seasons and quality grades. A total of seven effects, e.g. the effects of the three main factors and their interactions are the most interesting. Therefore, firstly the permutation test on the residuals under the reduced model was performed to check the statistical significance of the effects for those factors. The group means that should be subtracted from the centered experimental data are presented in Table 1.

For example, to test the effect of production season, the effect matrices for β, δ, $\beta\delta$ were subtracted from the centered data and the reduced data were used to estimate the true sum of squares. The true sum of squares was then compared with the distribution of the sums of squares obtained from the reduced data with a permuted order of objects. The null hypothesis, defined as that for which there is no effect of the factor being studied, is true when the sum of squares for the original data is smaller than the sum of squares for the permuted reduced data for most of the cases. In other words, there are no significant differences in the original group means associated with the particular factor in comparison with the group means calculated after the random assignment of objects to these groups. The histograms constructed for 10,000 permutations for the effects of all factors and their interactions (Eq. (1)) are shown in Fig. 3. The sum of squares for the true reduced data is displayed as a vertical line in each of the plots.

The null hypotheses for these are defined as those for which there are no significant effects of the factors and their interactions. The true sums of squares for all factors tested are much larger than the sums of squares for the permuted data. The P-values are 1/10,000 (Table 2) and hence, the null hypotheses can be rejected indicating significant effects.

In other words, the content of phenolic compounds related to the antioxidant properties of rooibos tea significantly depends on the production season, the use of steam pasteurization during the production process, the quality and homogeneity of the plant

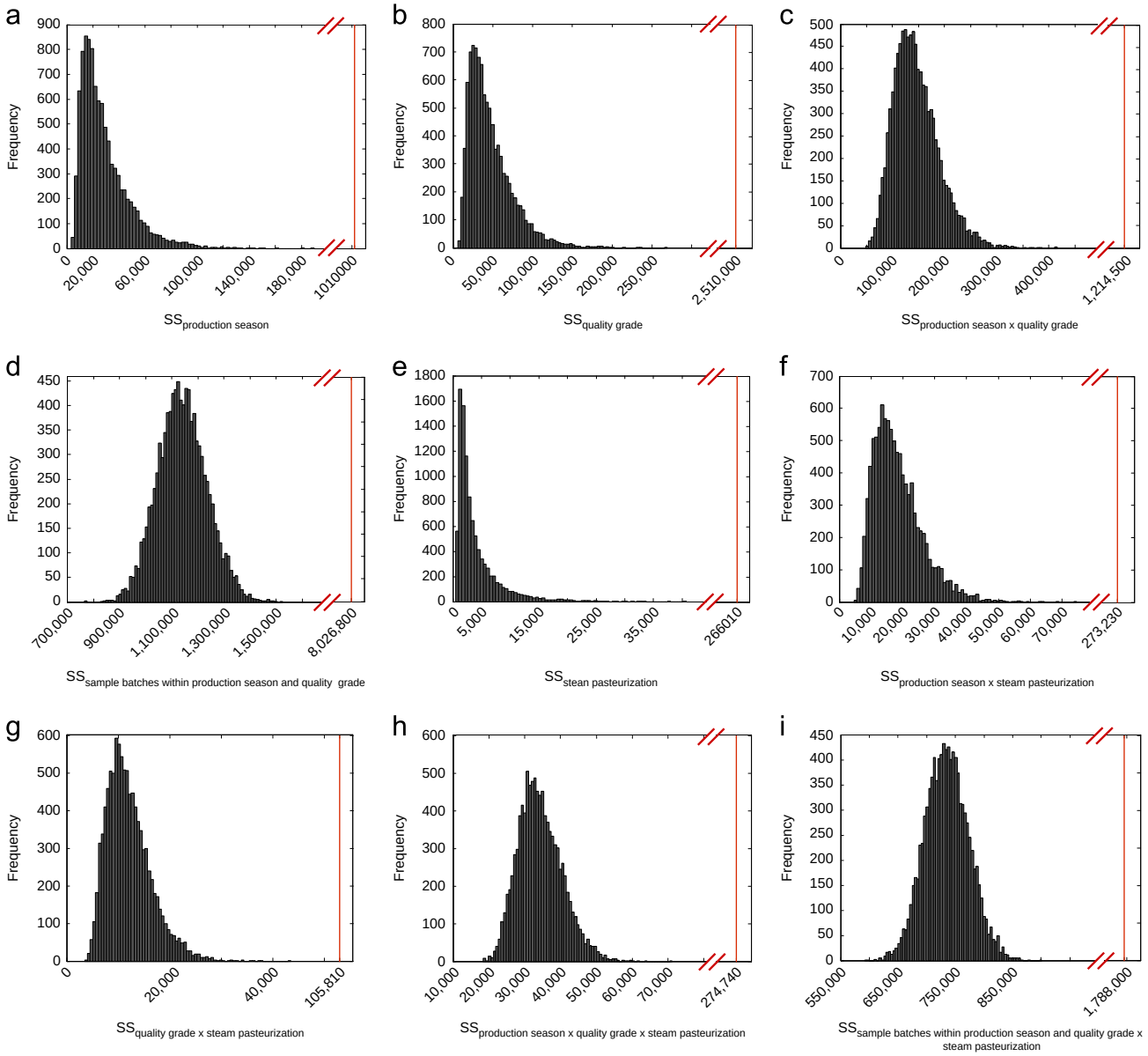

Fig. 3. Histograms constructed for 10,000 permutations for (a) 'production season' effect, α, (b) quality grade' effect β, (c) 'production season × quality grade' effect, $\alpha\beta$, (*d*) 'sample batches within production season and quality grade' effect, γ, (e) 'steam pasteurization' effect, δ, (*f*) 'production season × steam pasteurization' effect, $\alpha\delta$, (g) 'quality grade × steam pasteurization' effect, $\beta\delta$, (*h*) 'production season × quality grade × steam pasteurization' effect, $\alpha\beta\delta$ and (i) 'sample batches within production season and quality grade × steam pasteurization' effect, $\gamma\delta$.

material of a certain quality that is collected from individual farmers as well as on the effects of some combinations of those factors.

The percentages of variance for each factor estimated with respect to the corrected total sum of squares are listed in Table 2. The corrected total sum of squares is calculated using the centered data matrix. Since a balanced design of experiments was considered, the sum of individual parts of variation in percentages is equal to 100. From the values presented in Table 2, the larger percentage of variation among the individual factors is related to the tea quality factor (β, 15.09%) followed by the production season (α, 6.07%). The 'steam pasteurization' factor, δ, contributes much less (1.60%) to the total percentage of variance. The nested factor, γ, which is related to the homogeneity of the plant material from each quality grade and each production season, has a variance of 48.29%. Variation of 10.76%, is associated with the random effect of the two-factor interaction between the sample batches within grades and

Table 2
Results of the four-way mixed model ANOVA. *P*-values were obtained from permutation tests.

Source of variation	Degrees of freedom for the effect	Sum of Squares **x10^4**	Percentage of variance [%]	*P*-Value
α	2	101	6.07	<0.001
β	3	251	15.09	<0.001
$\alpha\beta$	6	121	7.28	<0.001
γ	108	803	48.29	<0.001
δ	1	26.6	1.60	<0.001
$\alpha\delta$	2	27.3	1.64	<0.001
$\beta\delta$	3	10.6	0.64	<0.001
$\alpha\beta\delta$	6	27.5	1.65	<0.001
$\gamma\delta$	108	179	10.76	<0.001
ε	720	116	6.98	
Total corrected	959	1663.0		

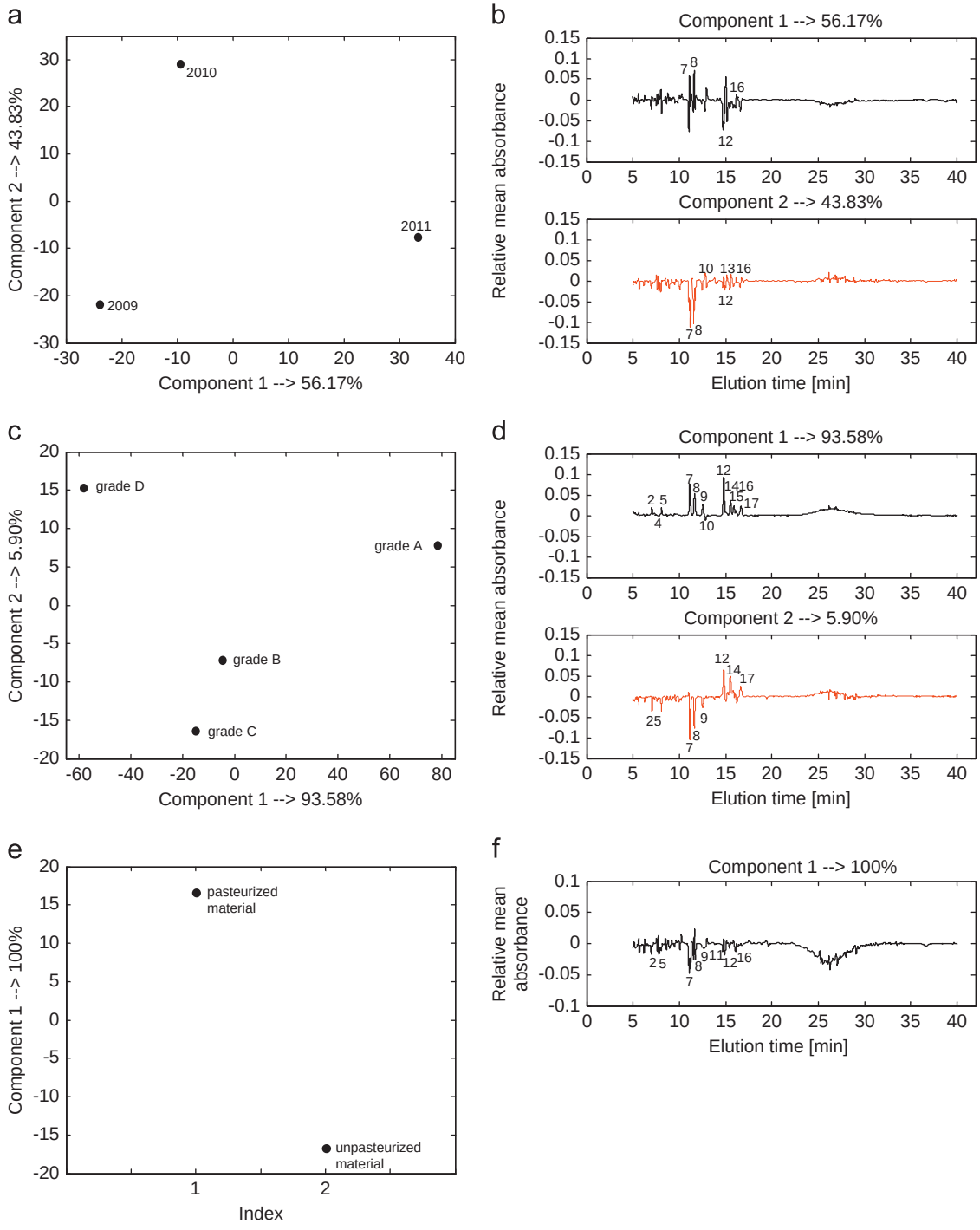

Fig. 4. Simultaneous component analysis sub-models: for the 'production season' effect (a) projections of years on the plane spanned by the first and second components, (b) projections of mean absorbance measurement points on the first and second components; for the 'quality grade' effect (c) projections of quality grades A, B, C and D on the plane spanned by the first and second components, (d) projections of mean absorbance measurement points on the first and second components respectively; for the 'steam pasteurization' effect (e) projections related to the pasteurized and unpasteurized material on the plane spanned by the first and second components and (f) projections of mean absorbance measurement points on the first component.

production seasons, and steam pasteurization ($\gamma\delta$). This is not a surprising observation since the content of the plant material may change from year to year depending on the climate conditions. Furthermore, the plant material was not collected from the same plantations year after year and therefore, the natural variation among plantations as a result of the use of seedlings is large.

The differences in the content of the different groups of tea samples can be further investigated by constructing individual component models for each effect matrix in Eq. (2). To simplify the description, only the sub-models for the main effects and the interaction effect that explain the largest part of the total variation will be presented here. Even though the 'steam pasteurization'

Table 3
Important phenolic compounds identified. Peak numbers are in retention time order.

Peak number	Name of the compound
1	Luteolin-6,8-di-*C*-glucoside
2	(*S*)-eriodictyol-6-*C*-glucoside
3	Carlinoside, isocarlinoside, neocarlinoside or isocarlinoside isomer
4	Vicenin-2
5	(*R*)-eriodictyol-6-*C*-glucoside
6	Phenylpyruvic acid-2-*O*-glucoside (PPAG)
7	Isoorentin
8	Orientin
9	Aspalathin
10	Ferulic acid
11	Aspalalinin
12	Quercetin-3-*O*-robinobioside
13	Vitexin
14	Hyperoside
15	Rutin
16	Isovitexin
17	Isoquercitrin
18	Nothofagin

factor explains only 1.60% of the total percentage of variance, it will be considered in the interpretation. The reason is that only pasteurized tea is sold commercially and it is of a great interest to see what causes differences, if any, in the phenolic content of infusions from samples with and without pasteurization.

The score and loading plots for the simple first factor effect explaining 6.07% of the total model variance are shown in Fig. 4a and b.

The number of components of this model is two because the number of degrees of freedom for the effect is two (Table 2). Similar to PCA, the results are interpreted combining information from both the score and loading plots. The phenolic compounds identified using data presented by Beelders et al. [13] are listed in Table 3.

Knowing retention times for the important phenolic compounds, and looking at the relative mean absorbance (loading values) for the first and second component (Fig. 4b), it can be pointed out that the infusions of samples collected in 2011 have relatively higher contents of orientin, isoorientin and isovitexin as well as somewhat lower contents of quercetin-3-*O*-robinobioside in comparison with the samples from the other two production seasons. The polymeric phenolic compounds, eluting as an undefined "hump" between 22 and 35 min elution time, were also less in samples from the 2011 production season. The rooibos tea infusions from the other two production seasons (2009 and 2010) can also be distinguished along Component 2 with respect to their content of isoorientin, orientin and quercetin-3-*O*-robinobioside. Higher contents of these phenolic compounds are characteristic for the infusions of samples collected in 2009 compared to 2010. A previous study, investigating only infusions of pasteurized fermented rooibos noted higher levels of isoorientin, isovitexin and PPAG with lower levels of nothofagin and rutin for 2011 samples compared to the other production seasons [12].

The effect of quality grade is associated with 15.09% of the total model variance, while the first two factors of the SCA explain 99.48% of the sub-model variance. In agreement with the previous observations [12] that the greater quality of the plant material is associated with the higher phenolic content, we came to a similar conclusion analyzing the score and loading plots are presented in Fig. 4c and d. The first component contrasts the content of almost all phenolic compounds with positive relative mean absorbances ((*S*)-eriodictyol-6-*C*-glucoside, (*R*)-eriodictyol-6-*C*-glucoside, orientin, isoorientin, aspalathin, quercetin-3-*O*-robinobioside, hyperoside, rutin, isovitexin, isoquercitrin and undefined polymeric phenolic compounds) with the content of vicenin-2, and ferulic acid with negative relative mean absorbances. The second component is associated mainly with the content of (*R*)-eriodictyol-6-*C*-glucoside, (*S*)-eriodictyol-6-*C*-glucoside, orientin, isoorientin, and aspalathin (negative relative mean absorbance) as well as vicenin-2 quercetin-3-*O*-robinobioside, hyperoside and isoquercitrin (positive relative absorbances). The infusion of the highest quality rooibos (grade A) has a high content of almost all phenolic compounds, including polymeric phenolic compounds and the potent antioxidant aspalathin, but has a somewhat lower content of ferulic acid in comparison with the lower quality rooibos tea material B, C and D. The infusion of grade D rooibos is also much lower in orientin, isoorientin and aspalathin in comparison with the grade C rooibos. The higher ferulic acid content of the lower quality grade material could potentially be explained in terms of the release of ferulic acid from lignocellulosic structures during over-fermentation, although both over- and under-fermentation could be a reason for a sample obtaining of a low quality grading. In addition, the ferulic acid content present in the plant material prior to fermentation will also play a role.

One principal component from PCA is enough (explaining 100% of the sub-model variation) to explain the differences in the phenolic compositions of infusions of pasteurized and unpasteurized samples. Fig. 4e and f present the score and loading plots obtained for this sub-model. This principal component is associated with the content of (*S*)-eriodictyol-6-*C*-glucoside, isoorientin, orientin, aspalathin, aspalalinin, quercetin-3-*O*-robinobioside, isovitexin and polymeric phenolic compounds (negative relative mean absorbance). Compared to the infusion of pasteurized plant material, a 'cup-of-tea' infusion from the unpasteurized material has a higher content of these phenolic compounds. This indicates that post-fermentation steam pasteurization causes a decrease in the amount of phenolic compounds extracted from the tea leaves during the infusion preparation. Previously, similar results were shown by Koch et al. [9] and Beelders [28].

The following sub-model was constructed for the interaction effect between production seasons and quality grades explaining 7.28% of the total model variance. The rank of this sub-matrix is six (Table 2) indicating the maximum number of components. The first two components explain 95.06% of the sub-model variation. The score and the respective loading plots are presented in Fig. 5.

Each of the two score sub-plots (Fig. 5a) shows the changes in phenolic composition of the tea depending on the quality grade over the different production seasons. The first component can be associated mainly with the content of isoorientin and orientin, while the second component is mainly related to aspalathin and quercetin-3-*O*-robinobioside. Compared to the grade A infusions obtained from the plant material of production seasons of 2009 and 2011, the grade A infusions of 2010 are richer in isoorientin and orientin. The content of these two compounds increases in the infusions of grade B and C in the last two years (2010 and 2011) and even a higher content than for the grade A infusions are observed in 2011. The content of isoorientin and orientin decreases in the infusions of grade D rooibos over the years and the lowest content is observed in 2011. The variation of aspalathin and quercetin-3-*O*-robinobioside content (along Component 2) in the different grade infusions over the production seasons is much smaller in comparison to the variations along component 1. Compared to the grade B infusions of 2009 and 2011, a larger content of aspalathin and quercetin-3-*O*-robinobioside is found in 2010.

Finally, the score and loading plots for the error matrix ($\mathbf{X}_e$) were inspected and no particular residual structure was observed in the space spanned by the first two components.

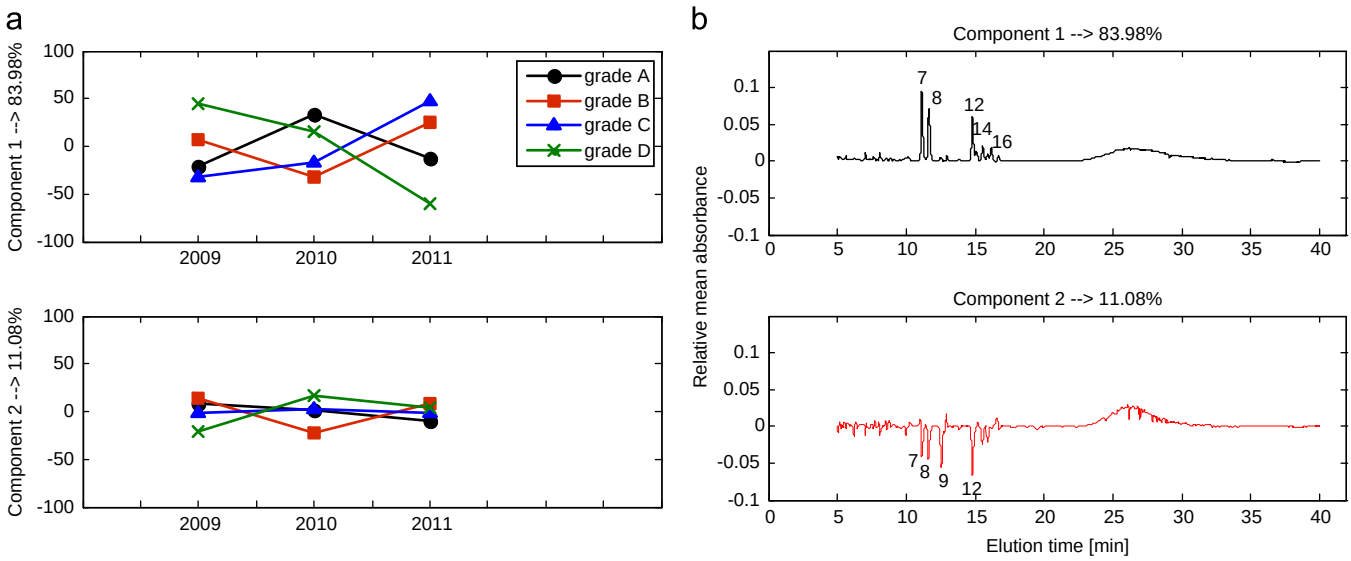

Fig. 5. Simultaneous component analysis sub-model for the 'production season × quality grade' effect (a) projections of sample batches of a particular quality A, B, C or D over time on the first and second components and (b) projections of the respective mean absorbance measurement points on the first and second components.

5. Conclusions

In this paper, a mixed-effect model of ASCA that enables the modeling and interpreting of high-dimensional data obtained from a design in which one factor is nested within two of the three crossed factors was presented. A scheme for the permutation testing of each term in the analysis of variance model was also shown.

Using the nested ASCA model, we came to the conclusion that production season, steam pasteurization and the quality grade all have a significant influence on the phenolic content of a rooibos infusion at 'cup-of-tea' strength. By analyzing the individual sub-models for different effect factors and the respective loadings, the interpretation of the effects based on individual phenolic compounds that were identified was possible. A major advantage of analyzing the entire phenolic profile, instead of quantified individual phenolic compounds, is that individual phenolic compounds that could not be quantified and/or identified are included in the data analysis. The effect of these factors on the polymeric phenolic content, represented by an undefined "hump" at the end of the chromatogram, could also be determined. Specifically, the infusions from the grade A samples have the highest content of most phenolic compounds in comparison with the infusions from lower quality plant material. The variations in the content of isoorientin and orientin in the different quality grade infusions over production seasons are larger than the variations in the content of aspalathin and quercetin-3-*O*-robinobioside. It seems that ferulic acid may be used as an indicator of rooibos tea quality as infusions from lower quality tea material are associated with a higher ferulic acid content. Another important finding is that a 'cup-of-tea' strength rooibos infusion from pasteurized plant material has a lower content of phenolic compounds, including polymeric phenolic compounds.

Acknowledgments

Rooibos samples were provided by Rooibos Ltd, Clanwilliam, South Africa. Funding for sample preparation and HPLC analysis was provided by the South African Rooibos Council and THRIP (Grant 72065 to EJ), an initiative of the Department of Trade and Industry, South Africa. AS was awarded a post-graduate bursary by the National Research Foundation of South Africa, under the Indigenous Knowledge Systems Programme, sponsored by the Department of Science and Technology (Grant 70525 to EJ).

We are grateful to the anonymous reviewers of this article for their insightful comments.

References

[1] C. Tistaert, B. Dejaegher, Y. Vander Heyden, Anal. Chim. Acta 690 (2011) 148.
[2] M. Carocho, I.C.F.R. Ferreira, Food Chem. Toxicol. 51 (2013) 15.
[3] J.J. Jansen, H.C.J. Hoefsloot, J. Van der Geert, M..E. Timmerman, J.A. Westerhuis, A.K. Smilde, J. Chemom. 19 (2005) 469.
[4] A.K. Smilde, H.C.J. Hoefsloot, J.A. Westerhuis, J. Chemom. 22 (2008) 464.
[5] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.A.N. Lamers, J. Van der Greef, M. Timmerman, Bioinformatics 21 (2005) 3043.
[6] P.B. Harrington, N.E. Vieira, P. Chen, J. Espinoza, J. Kae Nien, R. Romero, A. Yergey, Chemom. Intelligent Lab. Syst. 82 (2006) 283.
[7] I. Stanimirova, K. Michalik, Z. Drzazga, H. Trzeciak, P.D. Wentzell, B. Walczak, Anal. Chim. Acta 689 (2011) 1.
[8] E. Joubert, D. De Beer, S. Afr. J. Bot. 77 (2011) 869.
[9] I. Koch, M. Muller, E. Joubert, M. Van der Rijst, T. Naes., Food Res. Int. 46 (2012) 217.
[10] T. Heinrich, I. Willenberg, M.A. Glomb, J. Agric. Food Chem. 60 (2012) 5221.
[11] E. Joubert, D. de Beer, C.J. Malherbe, N. Muller, S.L. Bonnet, J.H. van der Westhuizen, D. Ferreira, Food Chem. 136 (2013) 1078.
[12] E. Joubert, T. Beelders, D. De Beer, C.J. Malherbe, A.J. de Villiers, G.O. Sigge, J. Agric. Food Chem. 60 (2012) 9171.
[13] T. Beelders, G.O. Sigge, E. Joubert, D. De Beer, A.De Villiers, J. Chromatogr. A 1219 (2012) 128.
[14] A.K. Smilde, M.E. Timmerman, M.M.W.B. Hendriks, J.J. Jansen, H.C.J. Hoefsloot, Brief. Bioinf. 13 (2011) 524.
[15] D.C. Montgomery, Design and Analysis of Experiments, John Wiley & Sons In., The USA, 2005.
[16] C.P. Doncaster, A. Davey, Analysis of Variance and Covariance. How to Choose and Construct Models for the Life Sciences, Cambridge University Press, UK, 2007.
[17] R.E. Kirk, Expermental Design: Procedures for the Behavioural Sciences, Wadsworth, Belmont CA, 1994.
[18] D.J. Vis, J.A. Westerhuis, A.K. Smilde, J. van der Greef, BMC Bioinf. 8 (2007) 322.
[19] M.J. Anderson, P. Legendre, J. Statist. Comput. Simulation 62 (1999) 271.
[20] M.J. Anderson, Can. J. Fish. Aquat. Sci. 58 (2001) 626.
[21] M.J. Anderson, C.J.F. Ter Braak, J. Statist. Comput. Simulation 73 (2003) 85.
[22] S. Wold, K. Esbensen, P. Geladi, Chemom. Intelligent Lab. Syst. 2 (1987) 37.
[23] J.J. Jansen, H.C.J. Hoefsloot, J. van der Greef, M.E. Timmerman, A.K. Smilde, Anal. Chim. Acta 530 (2005) 173–183.

[24] M. Daszykowski, Y. Vander Heyden, C. Boucon, B. Walczak, J. Chromatogr. A 1217 (2010) 6127.
[25] N. Nielsen, J. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[26] M. Daszykowski, B. Walczak, J. Chromatogr. A 1176 (2007) 1.
[27] P.H.C. Eilers, Anal. Chem. 75 (2003) 3631.
[28] T. Beelders, HPLC method development for the characterisation of the flavonoid and phenolic acid composition of rooibos (*Aspalathus linearis*) infusions, Stellenbosch University, Stellenbosch, South Africa, 2011, M.Sc. (Food Science) thesis.