



Article

WE-ASCA: The Weighted-Effect ASCA for Analyzing Unbalanced Multifactorial Designs—A Raman Spectra-Based Example

Nairveen Ali ^{1,2}, Jeroen Jansen ³, André van den Doel ^{3,4}, Gerjen Herman Tinnevelt ³ and Thomas Bocklitz ^{1,2,*}

¹ Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Helmholtzweg 4, D-07743 Jena, Germany; nairveen.ali@uni-jena.de

² Leibniz Institute of Photonic Technology (Leibniz-IPHT), Member of Leibniz Research Alliance Health Technologies, Albert-Einstein-Strasse 9, D-07745 Jena, Germany

³ Institute for Molecules and Materials (IMM), Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands; jj.jansen@science.ru.nl (J.J.); h.vandendoel@science.ru.nl (A.v.d.D.); G.Tinnevelt@science.ru.nl (G.H.T.)

⁴ TI-COAST, Science Park 904, 1098 XH Amsterdam, The Netherlands

* Correspondence: thomas.bocklitz@uni-jena.de; Tel.: +49-3541-948328

Abstract: Analyses of multifactorial experimental designs are used as an explorative technique describing hypothesized multifactorial effects based on their variation. The procedure of analyzing multifactorial designs is well established for univariate data, and it is known as analysis of variance (ANOVA) tests, whereas only a few methods have been developed for multivariate data. In this work, we present the weighted-effect ASCA, named WE-ASCA, as an enhanced version of ANOVA-simultaneous component analysis (ASCA) to deal with multivariate data in unbalanced multifactorial designs. The core of our work is to use general linear models (GLMs) in decomposing the response matrix into a design matrix and a parameter matrix, while the main improvement in WE-ASCA is to implement the weighted-effect (WE) coding in the design matrix. This WE-coding introduces a unique solution to solve GLMs and satisfies a constrain in which the sum of all level effects of a categorical variable equal to zero. To assess the WE-ASCA performance, two applications were demonstrated using a biomedical Raman spectral data set consisting of mice colorectal tissue. The results revealed that WE-ASCA is ideally suitable for analyzing unbalanced designs. Furthermore, if WE-ASCA is applied as a preprocessing tool, the classification performance and its reproducibility can significantly improve.

Keywords: ASCA; unbalanced experimental design; general linear model; weighted-effect coding; biomedical Raman spectra



Citation: Ali, N.; Jansen, J.; van den Doel, A.; Tinnevelt, G.H.; Bocklitz, T. WE-ASCA: The Weighted-Effect ASCA for Analyzing Unbalanced Multifactorial Designs—A Raman Spectra-Based Example. *Molecules* **2021**, *26*, 66. <https://dx.doi.org/10.3390/molecules26010066>

Academic Editor: Ewa Sikorska

Received: 11 November 2020

Accepted: 22 December 2020

Published: 25 December 2020

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An essential part of statistical analysis is the extraction of informative features that describe a specific phenomenon based on a limited number of samples. These samples are mostly collected by conducting either experiments or surveys [1,2]. In survey studies, a large number of individuals are involved to collect information without changing the existing conditions of the studied phenomenon. The other type of sampling is to conduct an experiment that tests the effect of one, or more than one, treatment on selected individuals. This experimental approach is widely applied in the fields of the physical and life sciences. In such studies, an experiment is carefully designed so that the obtained results are objective and valid [3,4]. Here, the term “design of experiment” (DoE) refers to statistical techniques that deal with planning and analyzing controlled tests, which investigate the effect of the studied treatments on selected individuals. There are several techniques for designing the experiments and one of the most common designs is the factorial design [5]. Therein, experiments are planned to extract information based on investigating the effect of at least

two treatments in one experiment [6]. These treatments are termed the experimental factors, and the combinations of these factors define their interactions.

Within biomedical studies, e.g., testing the efficiency of a new drug or checking a new technique for disease detection, the effect of experimental factors and their interactions are translated into different types of variations that can be categorized into two main groups: informative (or interesting) variations and disturbing (or unwanted) variations. The informative variations highlight the differences between different states like sample properties or disease states. In contrast, disturbing variations may be assigned to systematic perturbations within the experiment, which might negatively affect the results of further analyses. The later variation is very difficult to be controlled, and it mostly arises when many devices or different individuals are considered to perform an experiment. In this discourse, multifactorial analysis methods were introduced as powerful techniques to understand and analyze the variations within the factorial experimental design. The basic idea here is built upon hypothesis testing of more than two groups referring to factor levels. These factorial analysis tests were established quite well for univariate data, and they are known as analysis of variance (ANOVA) tests [1,7,8]. In one of their classical versions, namely the one-way ANOVA test, the effect of one factor on selected observations is studied based on testing the mean differences between the factor levels. The multi-way ANOVA tests search in a multifactorial design for significant effects based on checking the differences between the levels of each factor and each factor interaction. However, if the response data set is described by multiple features, only a few methods of multifactorial design were developed, which typically feature some limitations. For instance, in the basic form of multivariate-ANOVA (MANOVA) tests, an ANOVA test is performed for several response variables, which allows for studying the effect of one or more than one factor on these response variables [9,10]. The main restriction here is that performing MANOVA tests require a large number of measurements, e.g., sample size, compared to the number of variables (features). Such data are often not available, especially in modern technologies which introduce high dimensional measurements like spectra or images. Therefore, combining principal component analysis (PCA) models with ANOVA tests provided a solution to deal with the high dimensionality of response matrices in multifactorial designs [11]. The PC-ANOVA starts by fitting the response matrix with a PCA model, then the obtained principal components (PCs) are analyzed using ANOVA tests. Although the PC-ANOVA does not have the limitation respecting the sample size and number of response variables, some information related to factor contributions might be missed during the PCA projection. Besides, ANOVA simultaneous component analysis (ASCA) was presented as a powerful tool to deal with multivariate data in multifactorial designs [12,13]. In brief, ASCA methods decompose the response matrix into different effect matrices, which characterize the contribution of each effect in the designed model. These contributions are measured by the amount of variance explained by each effect. Thereafter, ASCA checks which effect contributes significantly to the considered model, and finally, the dimensions of each effect matrix are reduced based on a PCA model or a SCA model [14,15]. Later improvements of ASCA were introduced based on scaling the response matrix first, then applying the classical ASCA pipeline [16]. In this reference, it was shown how the considered scaling approach can affect the interpretation of the ASCA results. However, the proposed design of ASCA and its improvements are valid only for balanced designs, in which the levels of each factor have equal numbers of measurements. This constraint creates an additional limitation to the application area of ASCA. Thus, the ASCA+ was introduced later as an extension of ASCA to deal with unbalanced designs [17]. It utilizes general linear models (GLMs) to decompose the response matrix into two main terms: The estimated response matrix and the residual matrix, which refers to the estimation error [8,17]. Within ASCA+, the levels of each effect are coded using the deviation coding that has a main advantage concerning the variance maximization produced if the classical ASCA is applied on unbalanced designs [18].

In the paper, we review the implementation of ASCA and ASCA+ in unbalanced designs, and we introduce an updated extension of ASCA based on weighted-effect (WE) coding as a powerful tool to deal with these unbalanced designs. This WE-coding is a type of dummy coding that offers an attractive feature in which the sum of all level effects of a categorical variable is equal to zero [19,20]. Additionally, the results of WE-coding are identical to those obtained by deviation coding in balanced designs. The new ASCA extension, named WE-ASCA, substitutes the deviation coding of the design matrix in the ASCA+ method by the WE-coding in order to estimate the contributions of experiment effects. The performance of WE-ASCA was evaluated based on a Raman spectral data set of 47 individual mice for two different applications. In the first task, we analyzed the complex multifactorial design presented within the Raman data set using WE-ASCA, and we compared its results with the obtained ones by ASCA and ASCA+. Thereafter, the WE-ASCA was implemented as a preprocessing technique to improve the tissue classification. This was accomplished by applying the WE-ASCA to exclude disturbing variations from the training set. Later, a classification model was constructed using the new training set only, i.e., without disturbing variations, while the classification results based on WE-ASCA were compared with those obtained by the same classifier trained without using WE-ASCA-based preprocessing.

2. Results

Two applications of WE-ASCA are demonstrated in this section based on an unbalanced multifactorial design of a Raman spectral data set comprising 387 colorectal tissue scans that were collected from 47 mice. Within this study, the activity of the P53 gene (active +, inactive -) and the mice gender (male, female) were recoded while the Raman spectra of each scan were annotated as different tissue “types” representing normal, hyperplasia (HP), adenoma, and carcinoma tissue. Later, the biological variations of these colorectal tissues extracted from mice rectum or colon were evaluated based on the acquired Raman spectral scans [21]. In the presented work, a mean spectrum per tissue type was calculated resulting in 485 Raman spectra acquired from 387 scans. The number of mean spectra and number of scans differ because in one scan multiple tissue types can be present, yielding a higher number of mean spectra per scan. For example, a scan may contain cancer tissue and normal tissue, leading to two mean spectra for this scan. Figure 1 depicts the mean spectra of the tissue types beside the design of our experiment. Therein, the factors exhibit the sample location, the mouse gender, and the activity of the P53 gene. Notably, the number of spectra within the levels of each factor is different; thus, the introduced WE-ASCA is ideal for analyzing this unbalanced design.

In the following, the contributions of experimental factors in addition to their interactions are estimated based on the classical ASCA and its extensions ASCA+ and WE-ASCA. Then, an evaluation for the variance explained by each effect was performed. In the second subsection, the WE-ASCA is applied as a preprocessing technique to assess its performance in improving the classification of colorectal tissues.

2.1. A Comparison between Analyses of Multifactorial Design Using ASCA, ASCA+ and WE-ASCA

Since 47 mice were included to perform this study, an additional variation connected to the biological differences between mice might be produced. We added therefore another factor featuring the individuals (mice) contribution to the experimental design. Consequently, the multifactorial model that describes the considered experiment was built upon the individual factor with 47 levels indicating the mice, the activity of the P53 gene, the mouse gender, and the sample location (colon or rectum) in addition to interactions between the last three factors. This mathematical model can be formulated as:

$$X = M_0 + \text{Con}_{\text{Individual}} + \text{Con}_{\text{Location}} + \text{Con}_{\text{P53}} + \text{Con}_{\text{Gender}} + \text{Con}_{\text{Location:P53}} + \text{Con}_{\text{Location:Gender}} + \text{Con}_{\text{P53:Gender}} + \hat{E}, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{485 \times 696}$ is the Raman spectral matrix, \mathbf{M}_0 denotes a matrix in which mean Raman spectra with respect to the wavenumbers are oriented in rows, and \mathbf{Con}_f represents the matrix of an effect f . Based on this model, we evaluated the results of classical ASCA and its extensions ASCA+ and WE-ASCA using the obtained percentages of variances. As it is displayed in Table 1, the residual matrix of all analyses introduced the largest percentage of variance while the individual factor produced the largest factor contribution among all other effects. Here, the percentages of variance explained by the individuals are 37.97%, 33%, and 33.94% when ASCA, ASCA+, and WE-ASCA are applied, respectively. In contrast, the remaining factors and their interactions showed quite small contributions to the overall variance if any of the three multifactorial analyses were implemented. Nevertheless, the sum of percentages of variances by ASCA, ASCA+, and WE-ASCA is 108.62%, 93.3%, and 96.01%, respectively. This means that the classical ASCA maximized the effect contributions while the ASCA+ analysis minimized these contributions; however, the WE-ASCA analysis introduced the best estimations of effect contributions among the other analyses.

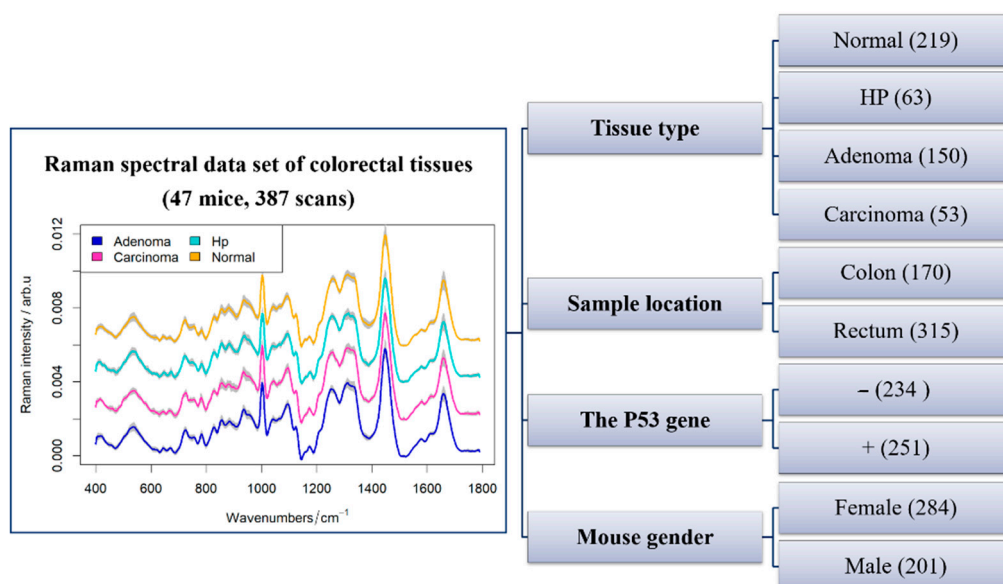


Figure 1. An overview of the experimental design of the Raman spectral data set. The studied data set consists of 47 mice and 387 scans that were collected from four different tissue types. The means spectra of these tissue types are shown in the left side. In this experiment, three factors were investigated: the activity of the P53 gene (active and inactive), the mouse gender, and the sample location (colon and rectum). It is observed that the number of scans of factor levels is different.

Table 1. The variance explained by the experimental effects in percentage (%) using ANOVA-simultaneous component analysis (ASCA), ASCA+ and WE-ASCA. A value of zero represents a percentage smaller than 10^{-9} .

Effect	Individual	Location	P53 Gene	Gender	Location: P53	Location: Gender	P53: Gender	Residuals	Sum (%)
ASCA	37.96	1.47	1.07	1.03	0.25	0.28	0.94	65.06	108.62
ASCA+	33.00	0.53	0	0	0.19	0.18	0	59.38	93.3
WE-ASCA	33.94	0.61	0	0	0.20	0.19	0	61.07	96.01

The obtained results in Table 1 show that only the individual factor and the sample location in addition to its interactions with the P53 gene and the mouse gender contributed to explain the variance in the considered design. To interpret the inner variance of these effects, a separate PCA model was fitted to each of previous effect matrices, then the obtained PCA results were summarized in Table 2 and Figure 2. The column named “all data” in Table 2 shows the percentages of variance explained by the first two PCs if the spectral

matrix X was mean centered and projected by a PCA model. These two PCs explained around 57.72% of the overall variance. For the PCA sub-model of the individual effect and the sample location, the three multifactorial analyses estimated approximately the same percentage of variance by the first two PCs. Therein, the first two PCs of the individual sub-models explained between 53.68% and 55.57% of the variance presented in the individual effect matrix, while the first PC estimated almost the whole variance of the sample location effect. Moving to the interaction effects, the first PC could describe almost 100% of the effect's variance if ASCA+ and WE-ASCA were applied. But this variance estimation was different in the case of classical ASCA, where the first PC explained around 3% variance less of this interaction effect. In Figure 2, the score plots of the PCA sub-models extracted from the effect matrix of the sample location and its interaction with the P53 gene and mouse gender are presented. The points in this figure depict the sum of the contribution of each effect and the projection of the residual matrix on the loadings obtained by the PCA sub-models (see [22] for more details). The bold points represent the group means and the ellipses are a representation of the covariance matrix. The WE-ASCA here shows slightly better separation between the group means in comparison to the results obtained by the ASCA and ASCA+.

After calculating the effect contributions and the PCA sub-models, we determined which effects contributed significantly to the experimental design using both ASCA extensions, i.e., ASCA+ and WE-ASCA, based on the described permutation test with $N = 1000$ iterations. The obtained p -values $p(f)$ by these tests were combined and presented in Table 3. Whether ASCA+ or WE-ASCA are applied, the individual factor and the sample location factor caused a significant effect in the design of our experiment. Here, the obtained $p(f)$ of the individual factor is almost zero by both analyses, while the $p(f)$ of the sample location is 0.021% and 0.034% when WE-ASCA and ASCA+ are applied, respectively.

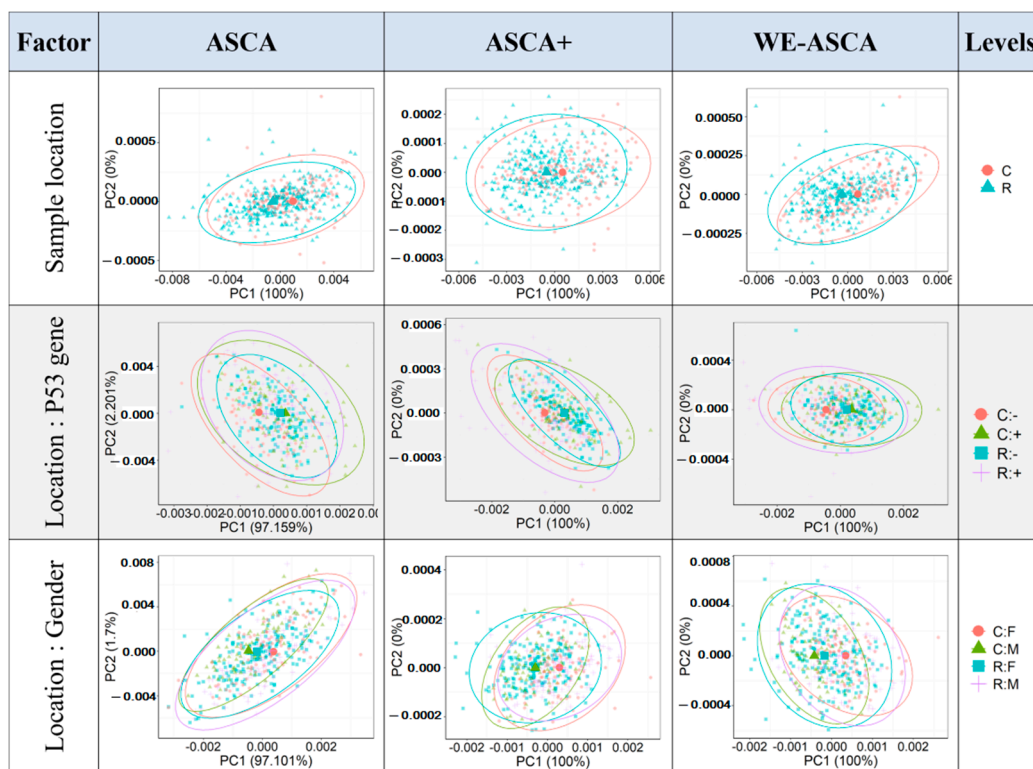


Figure 2. The score plots of first two principal components (PCs) of principal component analysis (PCA) sub-models using ASCA, ASCA+ and WE-ASCA. The WE-ASCA analysis provides better separation between the group means in comparison to the results obtained by the classical ASCA or its extension ASCA+.

Table 2. The percentage (%) of variance explained by the first two principal components of the PCA models. The mice data set and effect matrices obtained by ASCA, ASCA+, and WE-ASCA are fitted with a PCA model.

Effect		All Data	Individuals	Location	Location: P53	Location: Gender
ASCA	PC1 (%)	42.52	36.44	100	97.16	97.10
	PC2 (%)	15.20	17.74	0	2.20	1.7
ASCA+	PC1 (%)	42.52	38.78	100	100	100
	PC2 (%)	15.20	17.40	0	0	0
WE-ASCA	PC1 (%)	42.52	35.49	100	100	100
	PC2 (%)	15.20	18.19	0	0	0

Table 3. The obtained p -values $p(f)$ based on applying the permutation test on the results of ASCA+ and WE-ASCA analyses.

Effect		Individuals	Location	Location: P53	Location: Gender
$p(f)$	ASCA+	0.000	0.034	0.399	0.453
	WE-ASCA	0.000	0.021	0.383	0.424

To conclude, the WE-ASCA improved the analysis of unbalanced multifactorial design within the considered Raman data set. This improvement was detected by comparing the sum of explained variance obtained by the classical ASCA and its extension ASCA+ with the results of the presented WE-ASCA. The analysis of this experiment yielded that the effect of the individual factor produced the highest variation and the most significant effect within the studied data set.

2.2. A WE-ASCA as Preprocessing Technique in Classification Models

The goal of this subsection is to assess the performance of the ASCA analyses as a preprocessing technique. Therein, the WE-ASCA was implemented within the cross-validation as a preprocessing tool in order to exclude disturbing variations. In our work, a leave-one-mouse-out cross-validation was performed to check the results of two classifiers, namely the combination of a principal component analysis with a linear discriminant analysis (PCA-LDA) and the combination of a partial least square regression with a linear discriminant analysis (PLS-DA). The classification and validation procedure starts by fixing spectra of a specific mouse $T_i : i = 1, 2, \dots, 47$ as a test set and training the classifiers using the Raman spectra of the remaining mice, e.g., $\mathbf{X}(-T_i)$. This procedure is iterated until spectra of all mice are predicted once while the WE-ASCA is always applied on the training set of each cross-validation iteration (see Figure 3). It was shown in Section 2.1 that the individual factor produced the highest percentage of variance, which might negatively affect the classification results. Therefore, a new training set $\mathbf{X}_{\text{ASCA}}(-T_i)$ is estimated by excluding the variation of these individuals, then the training set $\mathbf{X}_{\text{ASCA}}(-T_i)$ can be defined as:

$$\mathbf{X}_{\text{ASCA}}(-T_i) = \mathbf{X}(-T_i) - \mathbf{Con}_{\text{Individual}}(-T_i); (i = 1, 2, \dots, 47), \quad (2)$$

where $\mathbf{Con}_{\text{Individual}}(-T_i)$ denotes to the individual effect matrix obtained by applying the WE-ASCA on $\mathbf{X}(-T_i)$. Using $\mathbf{X}_{\text{ASCA}}(-T_i)$ and $\mathbf{X}(-T_i)$, the PCA-LDA model and PLS-DA model are constructed. Then the tissue labels of each scan are predicted by both classifiers.

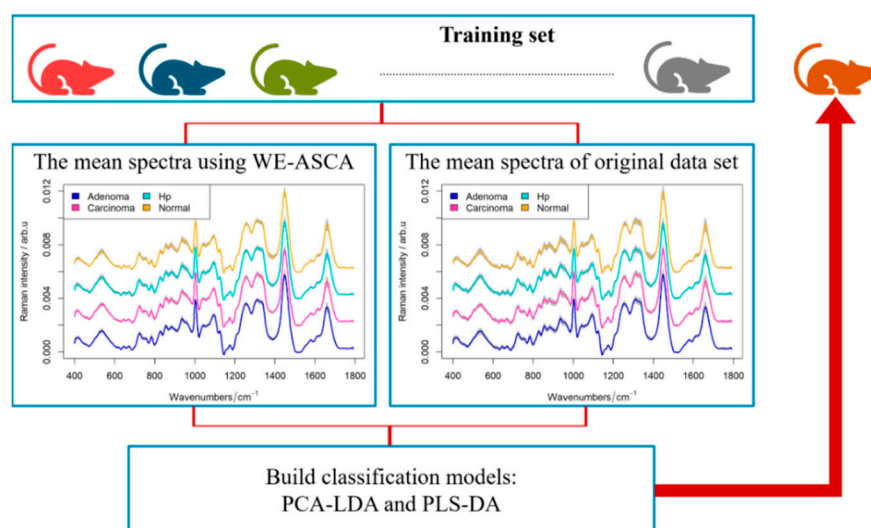


Figure 3. Overview of classification pipelines using WE-ASCA as a preprocessing step. Herein, a leave-one-mouse-out cross-validation (LOMO-CV) is implemented, while WE-ASCA is performed for each iteration within the CV loop. We can see that the variation of mean spectra after applying the WE-ASCA is smaller than the variation of mean spectra of the tissue types without preprocessing the spectral data. Based on training sets with or without applying WE-ASCA as preprocessing step, a PCA-LDA and PLS-DA are constructed, and their performance was determined.

The previous procedure was tested on four classification tasks, and the results are presented in Figure 4. The columns of this figure describe the results of PCA-LDA and PLS-DA models while the rows show the cross-validation results of each task for a different number of PCs or latent variables. If an LDA was trained on $\mathbf{X}_{ASCA}(-T_i)$, the standard deviation and the mean sensitivity were presented by the yellow regions and the yellow lines. In the other case, the blue regions and blue lines indicate the standard deviation and the mean sensitivity of LDA models constructed by $\mathbf{X}(-T_i)$ without preprocessing by WE-ASCA. In Figure 4A, the classification means sensitivities of PCA-LDA and PLS-DA are presented in order to differentiate between the scans of normal, HP, adenoma, and carcinoma tissues. It is observed that both classifiers produced better results when they were trained by the $\mathbf{X}_{ASCA}(-T_i)$. Here, the maximum mean sensitivity of the PCA-LDA model trained on $\mathbf{X}_{ASCA}(-T_i)$ is 50.67%. If the same classifier was built on the training set $\mathbf{X}(-T_i)$, the mean sensitivity decreased to 42.93%. Constructing a PLS-DA model on a $\mathbf{X}_{ASCA}(-T_i)$ or on $\mathbf{X}(-T_i)$ presented almost the same mean sensitivity, but it showed narrower standard deviation regions. For classifying adenoma and carcinoma tissues, it is clear in Figure 4B that using the WE-ASCA in preprocessing Raman spectra improved the LDA performance. While the maximum mean sensitivity of PCA-LDA and PLS-DA without implementing the WE-ASCA was 62.56% and 66%, respectively, the maximum mean sensitivity of PCA-LDA and PLS-LDA based on the training sets $\mathbf{X}_{ASCA}(-T_i)$ increased to 67.98% and 68.47%, respectively. Moving to Figure 4C, which represents the classification results of the three suspicious tissues, WE-ASCA significantly improved the classification results, which can be observed as an increase in the mean sensitivity and a decrease of the standard deviation of PCA-LDA and PLS-DA models if they are trained by $\mathbf{X}_{ASCA}(-T_i)$. In this case, the maximum mean sensitivity of PCA-LDA and PLS-LDA models increased at least 5% if they are constructed on the training sets $\mathbf{X}_{ASCA}(-T_i)$. The last classification task aimed to differentiate between normal tissues and tumor tissues combining carcinoma and adenoma spectra in one class. The obtained results are presented in Figure 4D. Both classifiers provided almost the same maximum mean sensitivity based on the training sets $\mathbf{X}_{ASCA}(-T_i)$ and $\mathbf{X}(-T_i)$. However, training the classification models on $\mathbf{X}_{ASCA}(-T_i)$ decreased the standard deviation and required less latent variables for constructing PLS-LDA models. Nonetheless, utilizing WE-ASCA in preprocessing Raman spectra enhanced

the results reproducibility of classification models. This reproducibility improvement is clearly seen in Figure 4 as narrower variation regions when the LDA models were trained by the $X_{ASCA}(-T_i)$. Here, the standard deviations of both classifiers built on $X_{ASCA}(-T_i)$ decreased significantly compared to the standard deviations of the same models trained by $X(-T_i)$.

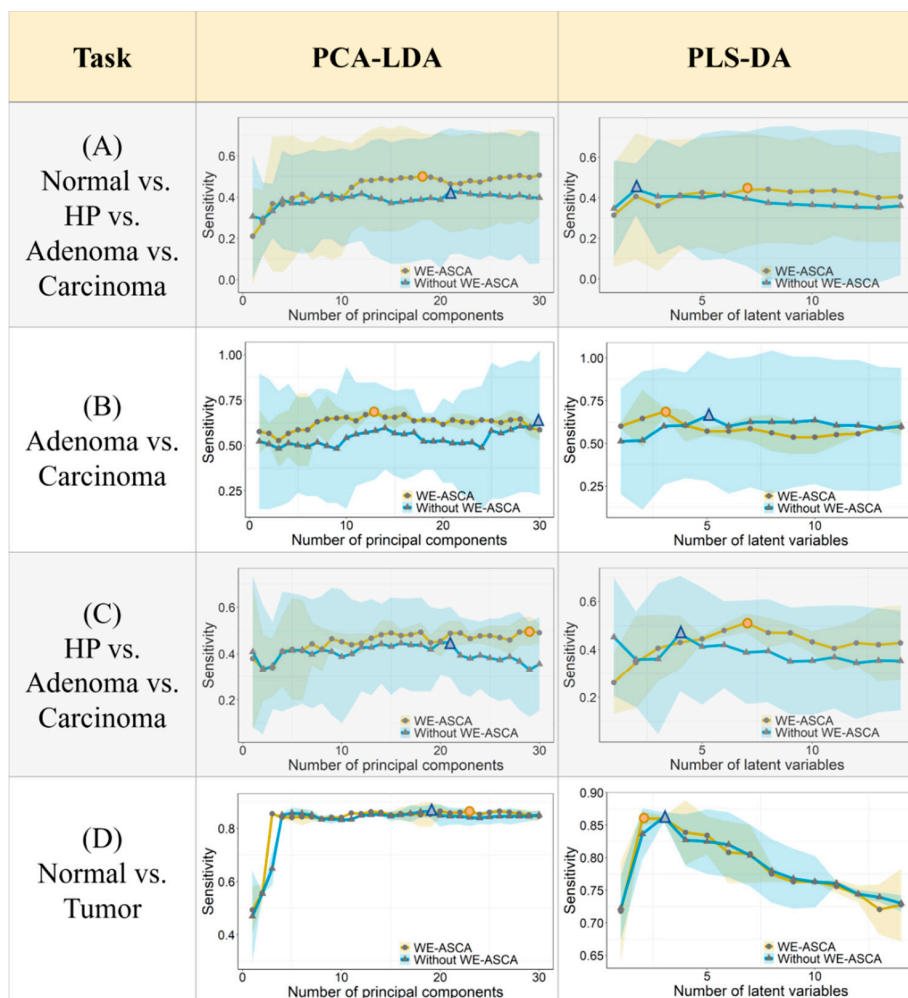


Figure 4. A comparison between the classification results of principal component analysis with a linear discriminant analysis (PCA-LDA) and partial least square regression with a linear discriminant analysis (PLS-DA) models based on leave-one-mouse-out cross-validation. Each classifier was trained twice with and without applying WE-ASCA-based preprocessing. The blue lines and the blue regions show the mean sensitivity and the standard deviation of a classifier constructed on the spectra without applying WE-ASCA on the training set. The yellow lines and the yellow regions depict the mean sensitivity and the standard deviation of a classification model trained on training sets that were preprocessed using WE-ASCA. The elimination of individual variations based on WE-ASCA improved the classification performance, and it significantly reduced the variance within the cross-validation results: **(A)** The maximum mean sensitivity for the differentiation between the scans of normal, HP, adenoma, and carcinoma tissues is 50.67%, and it was reached when training a PCA-LDA model on spectra processed by WE-ASCA. **(B)** For the classification of adenoma and carcinoma tissues, the maximum mean sensitivity of PCA-LDA (PLS-DA) is 67.98% (68.47%). These results were also achieved if the training sets were processed based on the WE-ASCA. **(C)** WE-ASCA-based preprocessing improved the differentiation between the three suspicious tissues. The maximum mean sensitivity of PCA-LDA model and PLS-DA are 49.85% and 51.09%, respectively. **(D)** The results of differentiating the normal and tumor tissue. While, training an PCA-LDA model with or without spectra processed by WE-ASCA provided almost the same classification results, training a PLS-DA model on spectra processed by WE-ASCA improved the mean sensitivity and decreased the standard deviation.

Overall, the WE-ASCA-based spectral preprocessing allowed us to exclude the disturbing variation from the training data set, and it significantly improved the classification performance. This improvement can be detected as an increase in the classification mean sensitivities and a reduction of the variance in the cross-validation results. Furthermore, the WE-ASCA-based spectral preprocessing required a smaller number of principal components (or latent variables) to build the classification models since the data distortion in the training data set was eliminated.

3. Discussion

The weighted-effect ASCA (WE-ASCA) was introduced as a new extension of the classical ASCA to analyze multivariate data in unbalanced multifactorial designs. The core of this WE-ASCA is to use the weighted-effect (WE) coding in designing model matrices of GLMs instead of dummy coding and deviation coding considered in designing schemes of classical ASCA and its extension ASCA+, respectively. The main advantage of implementing the WE-coding is that the sum of all level effects of a categorical variable in the design matrix is equal to zero for unbalanced designs, which is not the case by the other coding schemes. Also, the WE-coding offers a unique estimation of the effect of a specific variable because it always codes a variable with α levels by $\alpha - 1$ columns within the design matrix. The described advantages convinced us to update the coding scheme utilized in the design matrix of ASCA+ with the WE-coding. The response matrix then can be estimated easily by a general linear model and using the new balanced design matrix and the parameter matrix. This estimated response can be decomposed linearly as different effect matrices representing the experimental factors and their interactions. Besides, the significant effects in a particular design are determined based on permutation tests while the dimensions of the effect matrices are reduced using PCA.

Using a Raman spectral data set consisting of four colorectal tissue types that were collected from 47 mice in 387 scans, two possible applications of WE-ASCA were checked. Here, the data set was acquired with respect to four factors describing the experimental design. These factors are the different individuals with 47 levels referring to the mice, the activity of the P53 gene, the mouse gender, and the location of samples (colon or rectum). In the first application, we aimed to understand and analyze the design of our experiment in addition to determining which of experimental factors contributed significantly to the considered experiment. This was achieved by applying ASCA, ASCA+, and WE-ASCA and comparing their results based on the explained variances by all effects. It turned out that the classical ASCA overestimated the effect contributions, while the ASCA+ underestimated these contributions. In contrast, the presented WE-ASCA performed the best in estimating these effect contributions in term of the summation of percentage of explained variances. Nevertheless, the three versions of ASCA proved that the individual factor has the largest effect in our design. Therefore, we studied the influence of excluding such variations on the classification of colorectal tissues. This was demonstrated for four different classification tasks using two classifiers, i.e., PCA-LDA and PLS-DA, and leave-one-mouse-out cross-validation as a validation method. Our results showed that excluding the contribution of the individual factor from the training set introduced more robust classification results, and it improved the mean sensitivity in most classification tasks. Additionally, the training of an LDA model on spectra, when their individual effects were excluded, required a smaller number of principal components (or latent variables) and improved the reproducibility of the results.

4. Methods

4.1. ASCA for Crossed Balanced Design

The typical way of depicting the ASCA starts by introducing the ANOVA decomposition for the response of a single measurement described by a single variable [23]. Therein, the variation of each cell in a response matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, which has n measurements described by m variables (features), can be defined for a two-factor crossed design as:

$$x_{ijpq} = \mu_j + a_{jp} + b_{jq} + ab_{jpq} + \varepsilon_{ijpq}, i = 1, 2, \dots, n; j = 1, 2, \dots, m; p = 1, 2, \dots, \alpha; q = 1, 2, \dots, \beta, \quad (3)$$

where x_{ijpq} is the observed response of a measurement, i , on the variable, j , that has the level p of factor A and the level q of factor B . The parameter μ_j indicates the global mean with respect to the variable j , and ε_{ijpq} refers to the error term which is supposed to be a Gaussian distributed random variable with a mean of 0. Under additional constraints of ANOVA described in [13,17,23], a unique estimation of the previous statistical model is:

$$x_{ijpq} = x_{j..} + (x_{.jp} - x_{j..}) + (x_{.jq} - x_{j..}) + (x_{.jpq} - x_{.jp} - x_{.jq} + x_{j..}) + (x_{ijpq} - x_{.jpq}). \quad (4)$$

The dot-notation in the previous equation's subscripts describes over which index the mean is calculated. Moving to multivariate data, the classical ASCA provides a direct generalization of the ANOVA tests in balanced designs. It calculates the effect contributions to the response matrix \mathbf{X} :

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}, \quad (5)$$

where $\mathbf{M}_0 \in \mathbb{R}^{n \times m}$ represents the global mean matrix (its rows are the means over the variables), $\mathbf{Con}_f \in \mathbb{R}^{n \times m}$ refers to the estimated effect contribution of f where $f \in \{A, B, AB\}$, and $\hat{\mathbf{E}} \in \mathbb{R}^{n \times m}$ estimates the residual matrix. When the experimental design is balanced, each estimated effect matrix in Equation (5) is orthogonal, and the summation of percentage of variances of these matrices equals to 100%. Consequently, the contributions of individual effects in the overall variance can be measured by the partitions of the following sums of squares:

$$\|\mathbf{X}\|^2 = \|\mathbf{M}_0\|^2 + \|\mathbf{Con}_A\|^2 + \|\mathbf{Con}_B\|^2 + \|\mathbf{Con}_{AB}\|^2 + \|\hat{\mathbf{E}}\|^2, \quad (6)$$

where $\|\cdot\|^2$ indicates the squared Frobenius norm.

4.2. General Linear Models and ASCA+

The general linear models (GLMs) usually refer to a multiple linear regression where a continuous response variable is given continuous and (or) categorical predictors. These GLMs are fundamentals for several statistical tests such as ANOVA and ANCOVA (analysis of covariance) [24]. In its multivariate version, GLMs aim to decompose a response matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ linearly into different contributions based on a design matrix $\mathbf{D} \in \mathbb{R}^{n \times p}$ and a parameter matrix $\boldsymbol{\beta} \in \mathbb{R}^{p \times m}$. These contributions are related to the experimental factors and their interactions while the linear decomposition can be formulated as the following equation:

$$\mathbf{X} = \mathbf{D}\boldsymbol{\beta} + \mathbf{E}, \quad (7)$$

where $\mathbf{E} \in \mathbb{R}^{n \times m}$ denotes to the residual matrix. In formula (7), the matrix $\boldsymbol{\beta}$ relies only on the data features like intensity or pixel values while the method of coding the design (model) matrix \mathbf{D} performs a critical part in GLMs decomposition. In principle, the matrix \mathbf{D} can be designed using different coding techniques; however, a unique estimation of effect contributions can be obtained only if a factor with α levels is coded by $\alpha - 1$ columns in the design matrix [17]. By ASCA+, the deviation coding was utilized to design the matrix \mathbf{D} , where a factor with α levels is coded by $\alpha - 1$ columns with values of 0 and 1 for the first $\alpha - 1$ levels and with the value of -1 for the last level of this factor [17]. Then the parameter matrix $\boldsymbol{\beta}$ is estimated using the ordinary least square method as $\hat{\boldsymbol{\beta}} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{X}$, and the data matrix \mathbf{X} is approximated as $\hat{\mathbf{X}} = \mathbf{D} \hat{\boldsymbol{\beta}}$. The error of this estimation is determined by the residual matrix $\hat{\mathbf{E}}$ which can be written mathematically as:

$$\hat{\mathbf{E}} = \mathbf{X} - \mathbf{D} \hat{\boldsymbol{\beta}}. \quad (8)$$

Nevertheless, the main advantage of the previous coding is that the sub-design matrix of each effect is orthogonal for a balanced design. For instance, suppose a balanced two-factor crossed design is considered in which six measurements are affected by the factors $A : (a_1, a_2)$ and $B : (b_1, b_2, b_3)$. A response matrix $\mathbf{X} \in \mathbb{R}^{6 \times m}$ of this balanced design can be described with respect to these factor levels as:

$$\mathbf{X} = [X_{1,a_1b_1} \ X_{2,a_1b_2} \ X_{3,a_1b_3} \ X_{4,a_2b_1} \ X_{5,a_2b_2} \ X_{6,a_2b_3}]^T. \quad (9)$$

$X_{i,a_kb_l} \in \mathbb{R}^m$ indicates a measurement of level $a_p : p \in \{1, 2\}$ and level $b_q : q \in \{1, 2, 3\}$, which is oriented in the row $i \in \{1, 2, \dots, 6\}$. Based on the ASCA+ and the deviation coding, the design matrix \mathbf{D} and the sub-design matrix of factor B can be visualized by:

$$\mathbf{D} = \begin{bmatrix} M_0 & A & b_1 & b_2 & A : b_1 & A : b_2 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}; \mathbf{D}_{/B} = \begin{bmatrix} M_0 & A & b_1 & b_2 & A : b_1 & A : b_2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \end{bmatrix}.$$

Then, the effect matrix of factor B , namely \mathbf{Con}_B , is simply estimated by multiplying the matrix $\mathbf{D}_{/B}$ and the parameter matrix $\hat{\beta}$. Likewise, all sub-design matrices and effect matrices can be calculated, and the response matrix \mathbf{X} is subsequently decomposed according to formula (3) while the contribution of each effect into overall variance is estimated using a squared Frobenius norm (see Equation (6)).

4.3. Weighted-Effect ASCA (WE-ASCA)

The deviation coding introduced by ASCA+ provides an orthogonal design matrix only if an experimental design is balanced. In this case, the traditional constraints of analysis of variance models are perfectly satisfied, and the summation of percentage of variance equals to 100%. For unbalanced multifactorial designs, the design matrices introduced by any of ASCA or ASCA+ are non-orthogonal, and the estimated experiment effects are biased; therefore, it is desirable to remove, or at least reduce, these estimation biases. Another coding scheme, which can be considered to design the model matrix of GLMs in unbalanced data, is the weighted-effect (WE) coding. This WE-coding is a type of dummy coding which can be used to facilitate the inclusion of categorical variables in GLMs [19,20]. Thereby, the effect of each level of a categorical variable represents the level deviation from the weighted mean instead of using the grand mean in deviation coding. The WE-coding offers an attractive property related to the constrain in which the sum of all level effects of a categorical variable is equal to zero, which is not fulfilled by other coding schemes in unbalanced designs [19]. Moreover, the results of WE-coding are identical with those obtained by deviation coding if an experiment's design is balanced. Beside this, the estimated effect of a specific variable provided by the WE-coding are unique because a variable of α levels is coded by $\alpha - 1$ columns within the model matrix, and the interaction between two variables of α and β levels is coded by $(\alpha - 1) \times (\beta - 1)$ columns in this model matrix. Using the previous advantages, the WE-coding can be used to improve the performance of ASCA models in unbalanced multifactorial designs. In the following, the implementation of WE-coding in unbalanced multifactorial designs will be described in detail for a two-factor crossed design. However, this coding scheme is still valid for higher multifactorial designs. Let $\mathbf{X} \in \mathbb{R}^{7 \times m}$ be a response matrix collected from an unbalanced two-factor crossed design and presented as:

$$\mathbf{X} = [X_{1,a_1b_1} \ X_{2,a_1b_2} \ X_{3,a_1b_3} \ X_{4,a_2b_1} \ X_{5,a_2b_2} \ X_{6,a_2b_3} \ X_{7,a_2b_3}]^T, \quad (10)$$

where $X_{i,a_p b_q} \in \mathbb{R}^m$ indicates a measurement of level $a_p : p \in \{1, 2\}$ and level $b_q : q \in \{1, 2, 3\}$, which is oriented in rows $i \in \{1, 2, \dots, 7\}$. According to ASCA+, the design matrix \mathbf{D} and the sub-design matrix with respect to factor B can be depicted by:

$$\mathbf{D} = \begin{bmatrix} M_0 & A & b_1 & b_2 & A : b_1 & A : b_2 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}; \mathbf{D}_{/B} = \begin{bmatrix} M_0 & A & b_1 & b_2 & A : b_1 & A : b_2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \end{bmatrix}.$$

We can note that the design matrix \mathbf{D} is non-orthogonal and that the sum of level effects of the factors and their interaction does not equal to zero, which introduces biased estimators of experimental effects. The previous design matrix can be converted into a balanced design matrix if the WE-coding described by the coding matrix in Table 4 is applied. Based on this table, the obtained balanced design matrix \mathbf{BD} and the balanced sub-design matrix $\mathbf{BD}_{/B}$ are determined as the following:

$$\mathbf{BD} = \begin{bmatrix} M_0 & A & b_1 & b_2 & A : b_1 & A : b_2 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -2/3 & -2/3 & -1/1 & -1/1 \\ 1 & -3/4 & 1 & 0 & -1/1 & 0 \\ 1 & -3/4 & 0 & 1 & 0 & -1/1 \\ 1 & -3/4 & -2/3 & -2/3 & 1/2 & 1/2 \\ 1 & -3/4 & -2/3 & -2/3 & 1/2 & 1/2 \end{bmatrix};$$

$$\mathbf{BD}_{/B} = \begin{bmatrix} M_0 & A & b_1 & b_2 & A : b_1 & A : b_2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2/3 & -2/3 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2/3 & -2/3 & 0 & 0 \\ 0 & 0 & -2/3 & -2/3 & 0 & 0 \end{bmatrix}.$$

Here, each effect A, B , and AB is estimated in \mathbf{BD} by $p - 1 = 1, q - 1 = 2$, and $(p - 1) \times (q - 1) = 2$ columns, respectively. Clearly, the considered WE-coding estimates the levels of effects A, B , and AB in a way that the sum is always equal to zero.

Table 4. The coding matrix of a two-factor crossed design based on the weighted-effect coding. In this design, the effects of factors $A : (a_1, a_2), B : (b_1, b_2, b_3)$ and their interaction AB are presented.

Possible Combinations	Factors			Interaction	
	A	b_1	b_2	$A:b_1$	$A:b_2$
$a_1 \& b_1$	1	1	0	1	0
$a_1 \& b_2$	1	0	1	0	1
$a_1 \& b_3$	1	$-n_{b_1}/n_{b_3}$	$-n_{b_2}/n_{b_3}$	$-n_{a_1, b_1}/n_{a_1, b_3}$	$-n_{a_1, b_2}/n_{a_1, b_3}$
$a_2 \& b_1$	$-n_{a_1}/n_{a_2}$	1	0	$-n_{a_1, b_1}/n_{a_2, b_1}$	0
$a_2 \& b_2$	$-n_{a_1}/n_{a_2}$	0	1	0	$-n_{a_1, b_2}/n_{a_2, b_2}$
$a_2 \& b_3$	$-n_{a_1}/n_{a_2}$	$-n_{b_1}/n_{b_3}$	$-n_{b_2}/n_{b_3}$	$n_{a_1, b_1}/n_{a_2, b_3}$	$n_{a_1, b_2}/n_{a_2, b_3}$

In this paper, we update the ASCA+ by replacing the deviation coding of the design matrix in GLMs by the WE-coding. This updated version, namely weighted-effect ASCA

(WE-ASCA), provides a new extension of ASCA in unbalanced multifactorial designs. Thereby, a balanced design matrix \mathbf{BD} is estimated using the WE-coding [19,20], then the parameter matrix β is estimated based on the ordinary least square method:

$$\hat{\beta} = (\mathbf{BD}^T \mathbf{BD})^{-1} \mathbf{BD}^T \mathbf{X}. \quad (11)$$

The response matrix \mathbf{X} can be thereafter approximated as $\hat{\mathbf{X}} = \mathbf{BD} \hat{\beta}$, while the error of this estimation is given by:

$$\hat{\mathbf{E}} = \mathbf{X} - \mathbf{BD} \hat{\beta}. \quad (12)$$

Because the WE-coding estimates each factor of α levels by $\alpha - 1$ columns in the design matrix \mathbf{BD} , the presented WE-ASCA analysis provides a unique solution to solve the statistical models of any multifactorial design. Additionally, it reduces the bias introduced by unbalanced designs. Consequently, the effect matrix of any factor or interaction in multifactorial designs can be uniquely estimated by:

$$\mathbf{Con}_f = \mathbf{BD}_{/f} \hat{\beta}, \quad (13)$$

where $\mathbf{BD}_{/f}$ and \mathbf{Con}_f denotes the balanced sub-design matrix and the effect matrix of factor (or interaction) f , respectively. In case of a two-factor crossed design, the response matrix \mathbf{X} is decomposed linearly into different effects of the factors A and B and their interaction:

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}. \quad (14)$$

The previous estimated effect matrices have the same size of matrix \mathbf{X} . Thus, it is useful to reduce the dimensions of these matrices using PCA models in order to highlight the variations between different effect levels. In the presented two-factor crossed design, the statistical model based on the PCA sub-models can be presented by the decomposition:

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{T}_A \mathbf{P}_A^T + \mathbf{T}_B \mathbf{P}_B^T + \mathbf{T}_{AB} \mathbf{P}_{AB}^T + \hat{\mathbf{E}}, \quad (15)$$

where the matrix \mathbf{T}_f represents the score matrix, which highlights the variations between the levels of effect f . The matrix \mathbf{P}_f denotes the loadings matrix of the same effect.

4.4. The Percentage of Variance

One of the main goals of multifactorial design analysis is to study how different factors influence a particular experiment based on estimating their contributions to the overall variance. In balanced designs, a factor contribution is approximated simply using the type I sum squares. This method of sum squares sequentially computes the factor contributions with respect to their order in the designed model [25]. For the unbalanced multifactorial design, the factor levels have different numbers of measurements, which provides overestimation of some factor contributions. It is recommended according to [8,25,26] to calculate these contributions based on the type III sum squares. Thereby, the effect of one factor is evaluated after all other factors have been considered. This type of sum squares offers identical estimations of factor contributions with those obtained by type I sum squares when the considered design is balanced. In Table 5, the type III sum squares of each effect contribution in a two-factor crossed design is presented. Herein, the mean model decomposes the response matrix \mathbf{X} based on the global mean matrix \mathbf{M}_0 and the overall variance, named $\hat{\mathbf{E}}_1$. Then, the response matrix \mathbf{X} is decomposed by a reduced model that does not consider the contribution of a specific effect f . Subsequently, the residual matrix $\hat{\mathbf{E}}_f$ of this reduced model is estimated, and the sum squares of the effect f is calculated by the difference between $\|\hat{\mathbf{E}}_f\|^2$ and $\|\hat{\mathbf{E}}\|^2$. The explained variance by each effect can be

approximated finally as a percentage of the sum squares of that effect to the sum squares of the residual matrix of the mean model, i.e., $SS(\hat{\mathbf{E}}_1)$. Mathematically, the percentage of variance explained by an effect $f \in \{A, B, AB\}$ and the percentage of variance explained by the residual $\hat{\mathbf{E}}$ of a two-factor crossed model can be formulated as:

$$\%Var_f = \frac{\|\hat{\mathbf{E}}_f\|^2 - \|\hat{\mathbf{E}}\|^2}{\|\mathbf{X} - \mathbf{M}_0\|^2} \times 100 \text{ and } \%Var_{\hat{\mathbf{E}}} = \frac{\|\hat{\mathbf{E}}\|^2}{\|\mathbf{X} - \mathbf{M}_0\|^2} \times 100. \quad (16)$$

Table 5. Type III sum squares for a two-factor crossed design.

	Model	Type III Sum Squares
Mean model	$\mathbf{X} = \mathbf{M}_0 + \hat{\mathbf{E}}_1$	$SS(\hat{\mathbf{E}}_1) = \ \mathbf{X} - \mathbf{M}_0\ ^2$
Two-factor cross design	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}$	$SS(\hat{\mathbf{E}}) = \ \hat{\mathbf{E}}\ ^2$
Without the effect of A	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}_A$	$SS(\mathbf{Con}_A) = \ \hat{\mathbf{E}}_A\ ^2 - \ \hat{\mathbf{E}}\ ^2$
Without the effect of B	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_{AB} + \hat{\mathbf{E}}_B$	$SS(\mathbf{Con}_B) = \ \hat{\mathbf{E}}_B\ ^2 - \ \hat{\mathbf{E}}\ ^2$
Without the effect of AB	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \hat{\mathbf{E}}_{AB}$	$(\mathbf{Con}_{AB}) = \ \hat{\mathbf{E}}_{AB}\ ^2 - \ \hat{\mathbf{E}}\ ^2$

In our study, we compare the results of the WE-ASCA with the classical ASCA and its extension ASCA+ based on the summation of percentage of variances in an unbalanced design.

4.5. Permutation Tests

The basic idea of permutation tests is to check whether a specific effect f in ANOVA models contributes significantly to the variation of an experiment or it has a random influence [14,17]. For a two-factor crossed design, the procedure of permutation test starts by calculating the Frobenius sum squares of the first l principal components of the score matrix \mathbf{T}_f for each effect $f \in \{A, B, AB\}$:

$$SS(f) = \sum_{i=1}^n \sum_{j=1}^l (\mathbf{T}_f)_{ij}^2 \quad (17)$$

where n denotes the number of measurements in a response matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$. Then, we generate N random permutations of the rows of \mathbf{X} , and the Frobenius norm $SS_r(f)$ of each effect is computed for each permutation $r = 1, 2, \dots, N$ with respect to the considered ASCA extension, i.e., ASCA+ and WE-ASCA. The last step of this test is to calculate the p-value $p(f)$ as:

$$p(f) = \frac{\#\{SS_r(f) \geq SS(f)\}}{N}. \quad (18)$$

This p-value determines whether an effect, f , explains a random variation or shows a significant contribution within a considered experiment.

4.6. Data and Software

All computational parts were carried out based on in-house written functions in R version 3.4.2. The utilized Raman spectral data and R functions are freely available via Zenodo through the following links:

- Raman spectra of colon cancer in a mice model: <https://zenodo.org/deposit/3975464>

- Weighted-effect ASCA (WE-ASCA) codes: <https://zenodo.org/deposit/3975471>

5. Conclusions

The presented WE-ASCA provides an updated version of the ASCA and ASCA+ that suits the analysis of variance in unbalanced multifactorial designs. WE-ASCA proved its potential in understanding and analyzing the influence of experimental factors in a complex multifactorial design. Furthermore, the WE-ASCA was presented as a powerful preprocessing tool that can improve the classification performance and increase the classification reproducibility. The current implementations of WE-ASCA were checked only for Raman spectra and for tissue classification tasks; however, the application field is not limited to these previous applications. It can be extended to cover the analysis of variance of any type of multivariate data and any statistical modeling task.

Author Contributions: Conceptualization, T.B., N.A., and J.J.; methodology, T.B., G.H.T., A.v.d.D., J.J., and N.A.; software, N.A. and T.B.; validation, N.A. and A.v.d.D.; formal analysis, G.H.T., T.B., and N.A.; writing—original draft preparation, N.A. and T.B.; writing—review and editing, T.B., J.J., A.v.d.D., N.A., and G.H.T.; visualization, N.A.; supervision, J.J. and T.B.; project administration, J.J. and T.B.; funding acquisition, T.B. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: The funding from the DFG for the project BO4700/4-1 is highly appreciated. Financial support by the BMBF for the project Uro-MDD (FKZ 03ZZ0444) is highly acknowledged. This research received additionally funding from the Netherlands Organization for Scientific Research (NWO) in the framework of the Programmatic Technology Area PTA-COAST3 of the Fund New Chemical Innovations. This publication reflects only the authors' views and NWO is not liable for any use that may be made of the information contained herein. We acknowledge support by the German Research Foundation and the Open Access Publication Fund of the Thueringer Universitaets- und Landesbibliothek Jena Projekt-Nr. 433052568.

Data Availability Statement: The data of Vogler et al. [21] and scripts are available in the following repositories: Raman spectra of colon cancer in a mice model: <https://zenodo.org/deposit/3975464>; Weighted-effect ASCA (WE-ASCA) codes: <https://zenodo.org/deposit/3975471>.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Sample Availability: Not available.

References

1. Sampford, R.M.; Cochran, W.G. Sampling Techniques. *Biometrics* **1978**, *34*, 332–333. [CrossRef]
2. Smith, T.M.F.; Sugden, R.A. Sampling and Assignment Mechanisms in Experiments, Surveys and Observational Studies, Correspondent Paper. *Int. Stat. Rev.* **1988**, *56*, 165–180. [CrossRef]
3. Steinberg, D.M.; Hunter, W.G. Experimental Design: Review and Comment. *Technometrics* **1984**, *26*, 71–97. [CrossRef]
4. Winer, B.J. *Statistical Principles of Experimental Design*; McGraw-Hill: New York, NY, USA, 1962; Volume 3, pp. 381–385.
5. Mead, R.; Curnow, R.N.; Hasted, A.M. *Statistical Methods in Agriculture and Experimental Biology*, 3rd ed.; Chapman and Hall, CRC Press: Boca Raton, FL, USA, 2017; pp. 1–472.
6. Durakovic, B. Design of experiments application, concepts, examples: State of the art. *Period. Eng. Nat. Sci.* **2017**, *5*, 421–439. [CrossRef]
7. Støhle, L.; Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259–272.
8. Christensen, R. Analysis of Variance and Generalized Linear Models. In *International Encyclopedia of the Social & Behavioral Sciences*; Smelser, N.J., Baltes, P.B., Eds.; Pergamon Press: Oxford, UK, 2001; pp. 473–480.
9. Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*; Academic Press: London, UK, 1979; Volume 37.
10. Martens, H.M. Multivariate Analysis of Quality. An Introduction. *Meas. Sci. Technol.* **2001**, *12*, 1746. [CrossRef]
11. Bratchell, N. Multivariate response surface modelling by principal components analysis. *J. Chemom.* **1989**, *3*, 579–588. [CrossRef]
12. Jansen, J.J.; Hoefsloot, H.C.J.; van der Greef, J.; Timmerman, M.E.; Westerhuis, J.; Smilde, A.K. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chemom.* **2005**, *19*, 469–481. [CrossRef]
13. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.-J.A.N.; van der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef] [PubMed]
14. Anderson, M.; ter Braak, C. Permutation tests for multi-factorial analysis of variance. *J. Stat. Comp. Simul.* **2003**, *73*, 85–113. [CrossRef]

15. Bertinetto, C.; Engel, J.; Jansen, J.J. ANOVA simultaneous component analysis: A tutorial review. *Anal. Chim. Acta* **2020**, *6*, 100061. [[CrossRef](#)]
16. Timmerman, M.E.; Hoefsloot, H.C.J.; Smilde, A.K.; Ceulemans, E. Scaling in ANOVA-simultaneous component analysis. *Metabolomics* **2015**, *11*, 1265–1276. [[CrossRef](#)] [[PubMed](#)]
17. Thiel, M.; Féraud, B.; Govaerts, B. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J. Chemom.* **2017**, *31*, e2895. [[CrossRef](#)]
18. Madsen, H.; Thyregod, P. *Introduction to General and Generalized Linear Models*; CRC Press: Boca Raton, FL, USA, 2010.
19. Nieuwenhuis, R.; Grotenhuis, M.; Pelzer, B. Weighted Effect Coding for Observational Data with wec. *R. J.* **2017**, *9*, 477–485. [[CrossRef](#)]
20. te Grotenhuis, M.; Pelzer, B.; Eisinga, R.; Nieuwenhuis, R.; Schimdt-Catran, A.; König, R. A novel method for modelling interaction between categorical variables. *Int. J. Public Health* **2017**, *62*, 427–431. [[CrossRef](#)] [[PubMed](#)]
21. Vogler, N.; Bocklitz, T.; Salah, F.S.; Schmidt, C.; Bräuer, R.; Cui, T.; Mireskandari, M.; Greten, F.R.; Schmidt, M.; Stallmach, A.; et al. Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics. *J. Biophotonics* **2016**, *9*, 533–541. [[CrossRef](#)] [[PubMed](#)]
22. Zwanenburg, G.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Jansen, J.J.; Smilde, A.K. ANOVA–principal component analysis and ANOVA–simultaneous component analysis: A comparison. *J. Chemom.* **2011**, *25*, 561–567. [[CrossRef](#)]
23. Smilde, A.K.; Hoefsloot, H.C.J.; Westerhuis, J.A. The geometry of ASCA. *J. Chemom.* **2008**, *22*, 464–471. [[CrossRef](#)]
24. Neter, J.; Wasserman, W.; Kutner, M. *Applied Linear Statistical Models*, 4th ed.; Irwin: Chicago, IL, USA, 1996.
25. Shaw, R.G.; Mitchell-Olds, T. Anova for Unbalanced Data: An Overview. *Ecology* **1993**, *74*, 1638–1645. [[CrossRef](#)]
26. Langsrud, Ø. ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Stat. Comput.* **2003**, *13*, 163–167. [[CrossRef](#)]