Analyzing Designed Experiments with Multiple Responses

Øyvind Langsrud*, Kjetil Jørgensen**,
Ragni Ofstad* & Tormod Næs*

**MATFORSK, Osloveien 1, N-1430 Ås, NORWAY*

***TINE BA, Center for R&D, PO Box 50, N-4358 Kleppe, NORWAY*

To appear in
Journal of Applied Statistics

ABSTRACT

This paper is an overview of a unified framework for analyzing designed experiments with univariate or multivariate responses. Both categorical and continuous design variables are considered. To handle unbalanced data, we introduce the so-called Type II* sums of squares. This means that the results are independent of the scale chosen for continuous design variables. Furthermore, it does not matter whether two-level variables are coded as categorical or continuous. Overall testing of all responses is done by 50-50 MANOVA, which handles several highly correlated responses. Univariate p-values for each response are adjusted by using rotation testing. To illustrate multivariate effects, mean values and mean predictions are illustrated in a principal component score plot or directly as curves. For the unbalanced cases, we introduce a new variant of adjusted means, which are independent to the coding of two-level variables. The methodology is exemplified by case studies from cheese and fish pudding production.

KEY WORDS : 50-50 MANOVA; General linear model; Least-squares means; Multiple testing; Principal component; Rotation test; Unbalanced factorial design.

1 INTRODUCTION

Experimental designs are useful in a large number of industrial and scientific situations and applications. Traditionally, each response variable has been analyzed separately (Box *et al.* (1978) and Montgomery (1991)), but with the introduction of modern measurement instruments that typically produce a large number of highly correlated variables and with the increased complexity of scientific problems, simultaneous analysis of several response variables has become more in focus (e.g. Smilde *et al.* (2005), Nair *et al.* (2002), Langsrud (2001, 2002), Ellekjær *et al.* (1996)). In this paper we will consider multivariate extensions of univariate ANOVA and regression analysis (general linear modeling). The univariate challenges are still present, but for the multivariate case, there are a number of extra aspects that have to be taken into account. This paper is an overview of a unified framework for analyzing designed experiments that can be used both for univariate and multivariate responses. Our approach has been implemented as a standalone windows program and as functions in Matlab and R/Splus (www.matforsk.no/ola/program.htm).

The design variables are allowed to be categorical (factorial designs) as well as continuous (response surface designs). The design does not need to be perfectly balanced and the results will be independent of the scale chosen if continuous design variables are used (Celsius and Farenheit give same results). The starting point for our approach is the classical balanced factorial design with one single response. The effects are evaluated by standard analysis of variance (ANOVA) and they can be illustrated by presenting ordinary mean values accompanied with standard errors. To generalize this analysis to more complex situations, a lot of important choices have to be made. Below we describe some of our preferred choices and discuss them briefly.

The ANOVA method presented for unbalanced designs (named as Type II*) differs from the standard one (Type III) used in most major statistical programs. One of the reasons for this choice is to ensure that the results are independent to scale changes and to the coding of two-level variables (categorical or continuous). To describe various effects we focus on (adjusted) mean values and mean predictions rather that model parameters. A reason for this choice is that such results can easily be communicated to practitioners. Another reason is that it allows the results from multivariate analyses to be presented in a similar way as ordinary multivariate observations. Multivariate means are illustrated as curves or as points in a principal component plot.

A relatively new multivariate ANOVA (MANOVA) method, named as 50-50 MANOVA, is used to perform overall testing of all responses. Classical MANOVA has been modified so that collinear and several highly correlated responses are handled in a satisfactory way (Langsrud, 2002). When analyzing multiple responses, we will also consider a measure of explained variance associated with each model component. This is based on univariate sums of squares summed over all responses. In the single response case this measure is just a scaled version of the ordinary sums of squares.

In many multiresponse situations one is still interested in testing all the responses by individual significance tests. I these cases p-values from ordinary F-tests are not appropriate - since a lot of of type I errors ("incorrect significance") are expected. Therefore the p-values need to be adjusted. In the present paper the p-values are adjusted by using rotation testing, which makes use of the dependence among the responses. Both family-wise error rates and false discovery rates are considered.

Below we start by describing two case studies. The following presentation of the single response analysis and thereafter the more complex multiresponse analysis will be related to these case studies.

2 CASE STUDIES

2.1 Cheese data, partly replicated fractional factorial

This data set is from a pilot plant production of cheese. The main purposes of the experiments were

- To build an empirical model (that later can be used for process adjustments) that relates the design variables with important quality parameters of the cheese, and
- To find if the process or raw material factors that were a part of the design can be measured by spectroscopy in the finished product. If this is the case this indicates that the spectroscopic measurements later can be used in process monitoring and control.

The data are put together from three experimental design studies that were conducted with somewhat different focuses. The data are therefore both unbalanced, taken over different time periods and has hundreds of highly correlated response variables. All this require special considerations in the analysis.

More specifically, the first of the three experiments is a $2^{(6-1)}$ fractional factorial design with six factors. Only four of these six design variables were varied in the later experiments, so for this paper we used only the eight design points with the same setting of the other two variables as in the later designs. This means that the observations used in our example is a $2^{(4-1)}$ design. The second experiment is the same $2^{(4-1)}$ design, now with four center points. This was performed to test the linearity of some of the effects from the first experiment. The third experiment is a verification experiment where two of the levels of design variable "protein" from the other designs and two new levels of "temperature" were used.

The three experiments were performed in separate time periods. Hence experiment number is regarded as a block variable in the analyses. Some of the observations are missing. Altogether our case study is a partly replicated $2^{(4-1)}$ design with four center points and with six additional points. The design variables are (abbreviations in parentheses): *Protein in cheese milk* (P), *Starter culture added* (S), *Renneting time* (R) and *Heating temperature* (T). The levels are given in Table 1. The response is the percentage of dry matter in the cheese (DM) measured by a routine gravimetric laboratory method.

In this case we have spectroscopic measurements as multiple responses. The cheeses were grined and dissolved according to a procedure developed by LosAB (LosAB, Uppsala,

Table 1: *Cheese: The design variables and the response DM. This is a partly replicated fractional design (nr 1-13) with four center points (nr 14-18) and with six additional points (nr 19-24).*

nr	Block	P	S	R	T	DM
1	2	3.150	1.70	0.0	36.50	54.9
2	1	3.500	2.20	0.0	36.50	55.8
3	2	3.500	2.20	0.0	36.50	57.0
4	2	3.500	1.70	7.0	36.50	56.3
5	1	3.150	2.20	7.0	36.50	54.4
6	1	3.500	1.70	0.0	39.00	57.5
7	2	3.500	1.70	0.0	39.00	56.8
8	1	3.150	2.20	0.0	39.00	57.5
9	2	3.150	2.20	0.0	39.00	57.3
10	1	3.150	1.70	7.0	39.00	56.9
11	2	3.150	1.70	7.0	39.00	57.2
12	1	3.500	2.20	7.0	39.00	58.3
13	2	3.500	2.20	7.0	39.00	57.5
14	2	3.325	1.95	3.5	37.75	57.0
15	2	3.325	1.95	3.5	37.75	56.1
16	2	3.325	1.95	3.5	37.75	56.5
17	2	3.325	1.95	3.5	37.75	57.0
18	3	3.500	1.95	0.0	37.30	56.9
19	3	3.500	1.95	0.0	37.30	56.9
20	3	3.500	1.95	0.0	37.30	56.3
21	3	3.500	1.95	0.0	37.30	55.3
22	3	3.150	1.95	0.0	38.20	56.3
23	3	3.150	1.95	0.0	38.20	55.4
24	3	3.150	1.95	0.0	38.20	56.3

Sweden) and FT-IR spectra were obtained on a Milkoscan FT 120 (Foss AS, Hillerød, Denmark). We removed areas in the spectra that only contain noise form the high amount of water. The resultant spectra contain 477 variables in the range 964-2970 cm^{-1} and have a resolution of 3.858 cm^{-1}. The raw transmittance values were transformed to absorbance by the transformation $\widetilde{A} = \log(1/\widetilde{T})$, where $\widetilde{A}$ is the absorbance and $\widetilde{T}$ is the transmittance measurements. No other preprocessing of the spectra was used. The resulting spectra are shown in Figure 1. To see differences, Figure 2 presents spectra where the mean spectrum has been subtracted.

2.2 Fish pudding data, 6×3 design

Fish pudding is a traditional product in Norway. The common ingredients used are fish mince, salt, starch, skim milk powder, oil and spices which are finely comminuted before gel-set by heating. The texture (hardness) and the sensory properties of the end product are dependent on amount and quality of fish raw material as well as the additives used in the recipe. The two protein sources in this product are fish protein and skim milk protein (SMP). Fish is a high cost ingredient and the amount used has a large impact on the product cost. Therefore we want to study how the product quality is influenced by the amount of fish protein and SMP.

Our example is an experiment designed by varying Fish% and Cost (total cost of all ingredients) according to a 6×3 factorial design. The Fish% was varied at six levels (35%, 37.5%, 40%, 42.5%, 45% and 47.5%). The specified cost levels (coded as 1, 2 and 3) where achieved by adjusting the amount of SMP. Note that before this experiment was run there has been conducted a series of several fish pudding designs. Earlier SMP was used directly as a design factor, but we found out that the most interesting experimental regions were obtained by varying cost instead.

The 18 recipes were processed at two days, which were considered as blocks in this experiment. In this non-replicated situation orthogonal blocking was impossible, but the blocks were manually designed to be nearly orthogonal. Within each day, the order of the recipes was randomized. The design is presented in Table 2 together with Hardness, which is measured instrumentally by a Texture Analyser (SMS Texture Analyser, TA-XT2). In addition, sensory analysis was performed according to a descriptive sensory strategy by a trained sensory panel. A continuous, non-structured scale (1.0-9.0) was used for evaluation. Results from the 24 properties evaluated are given in Table 3.

3 SINGLE RESPONSE ANALYSIS

This section describes the univariate special case of our framework. Various types of analyses will be exemplified using our two case studies. For demonstration purposes variables that are naturally continuous will in some cases be regarded as categorical. Below we start by analyzing a balanced factorial design. Although this analysis is quite standard, it is a useful basis for discussing the more advanced situations.

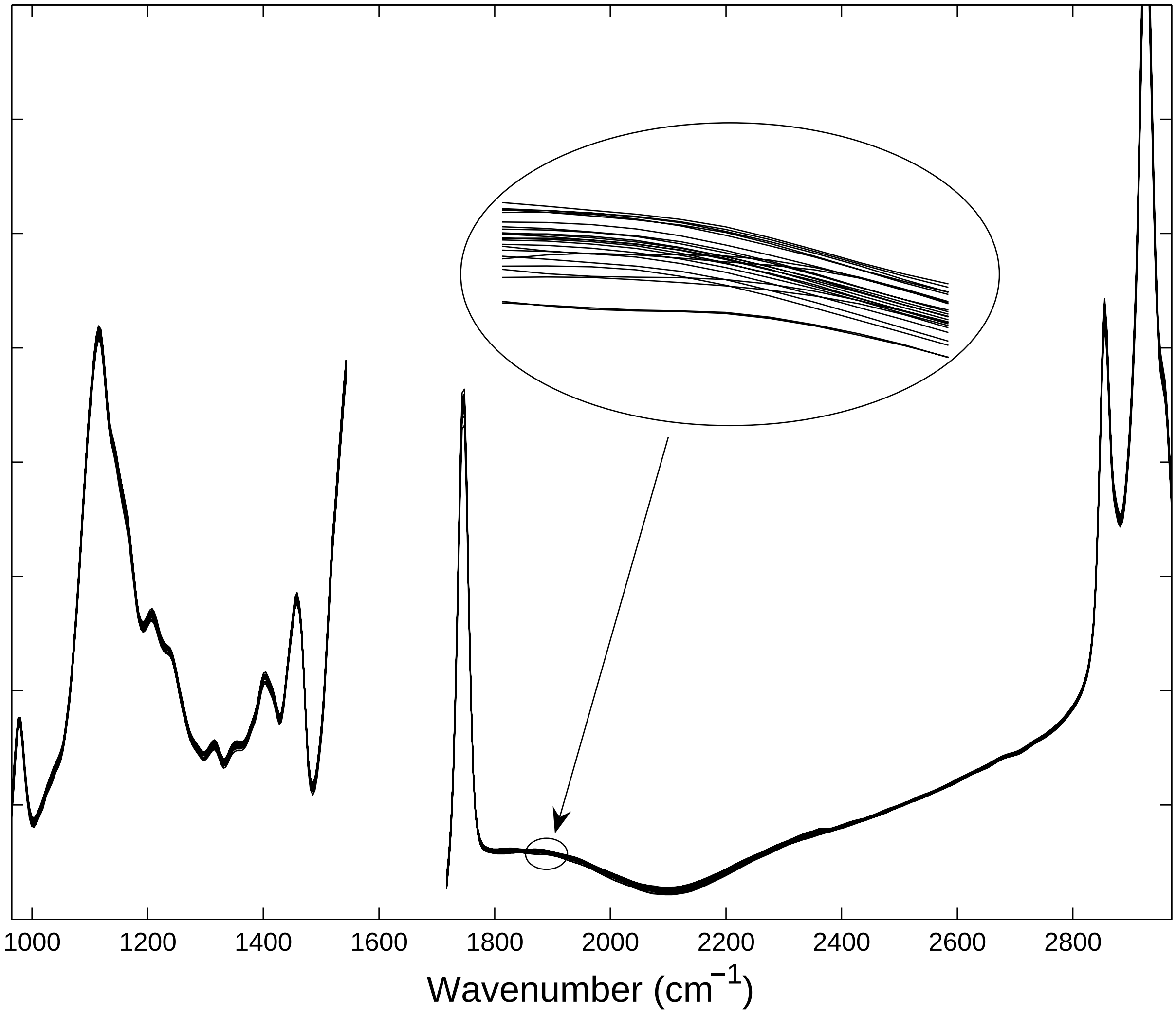

Figure 1: *Cheese: The FT-IR spectra.*

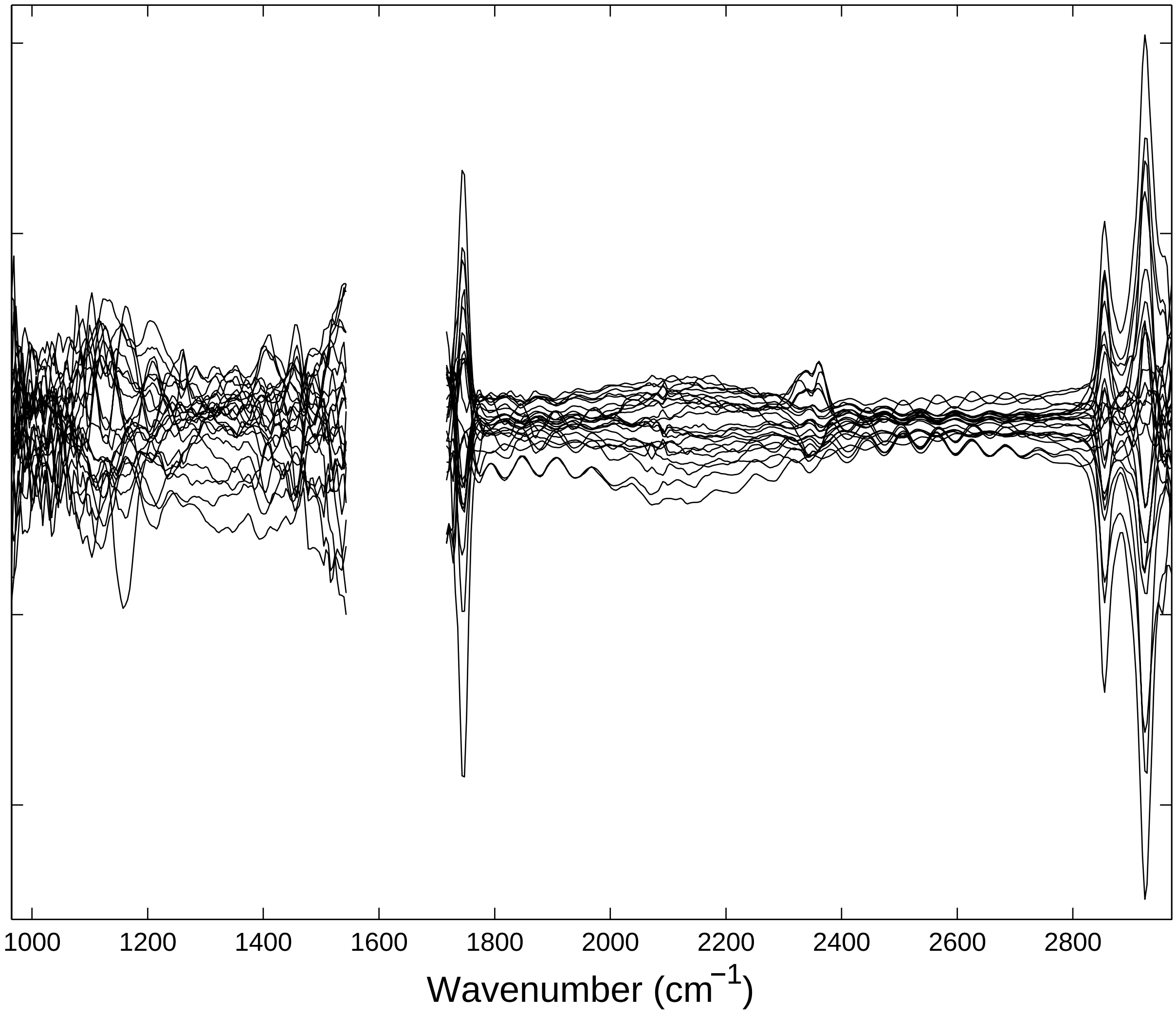

Figure 2: *Cheese: The mean centered FT-IR spectra.*

Table 2: *Fish pudding: The design together with the instrumental hardness response.*

Day	Fish%	Cost	Hardness
2	35.0	1	3.78048
1	35.0	2	3.84826
2	35.0	3	3.94864
1	37.5	1	3.72921
2	37.5	2	3.88381
1	37.5	3	3.94078
2	40.0	1	3.75849
1	40.0	2	3.79866
2	40.0	3	3.95267
1	42.5	1	3.70348
2	42.5	2	3.66002
1	42.5	3	3.83951
2	45.0	1	3.73655
1	45.0	2	3.83846
2	45.0	3	3.84729
1	47.5	1	3.71353
2	47.5	2	3.69649
1	47.5	3	3.79089

Table 3: *Fish pudding: Responses from sensory evaluation (transposed table).*

OdourIntensity	6.12	6.06	6.34	6.13	6.31	6.18	6.10	6.05	6.19	6.14	6.31	6.29	6.14	6.02	6.14	6.49	6.33	6.09
FishOdour	3.44	3.21	2.73	4.10	3.48	3.55	4.60	4.25	3.30	5.96	6.43	4.35	4.38	5.62	4.73	6.64	6.43	5.40
AcidityOdour	3.67	3.47	3.16	4.26	3.94	3.74	4.44	4.28	3.81	5.00	5.19	4.10	4.37	5.03	4.47	4.97	5.07	4.89
SpicyOdour	5.32	5.32	5.68	5.34	5.60	5.47	4.95	5.10	5.61	4.01	4.14	4.84	5.06	4.00	4.73	3.56	3.92	4.42
MilkOdour	3.34	3.26	3.50	2.93	3.42	3.20	2.82	3.06	3.46	2.55	2.07	3.12	2.80	2.54	2.79	2.03	2.36	2.50
BurntOdor	2.58	3.06	3.54	1.95	2.90	2.52	1.88	2.08	2.74	1.24	1.14	2.16	1.78	1.16	1.55	1.07	1.28	1.44
Whiteness	6.91	6.48	6.48	6.95	6.76	6.77	7.30	7.05	6.59	7.80	7.84	7.38	7.20	7.60	7.25	7.94	7.95	7.68
ColourToning	1.78	2.10	2.30	1.55	2.02	1.84	1.41	1.66	1.86	1.20	1.14	1.45	1.40	1.21	1.39	1.09	1.12	1.19
ColourStrength	3.49	4.18	4.74	3.17	3.52	3.86	2.31	2.89	3.81	1.77	1.47	2.51	2.63	1.88	2.30	1.20	1.38	1.80
TasteIntensity	6.55	6.49	6.78	6.27	6.34	6.65	6.38	6.27	6.53	6.43	6.47	6.48	6.28	6.31	6.23	6.38	6.34	6.56
FishTaste	3.33	3.04	2.62	3.80	3.03	3.12	4.91	4.05	3.16	6.37	6.84	3.91	4.76	5.85	4.86	6.77	6.73	5.70
AcidityTaste	3.40	3.13	2.73	3.83	3.02	3.03	4.55	3.86	3.39	4.97	5.30	3.69	4.48	5.04	4.55	4.76	5.00	4.76
Saltiness	4.51	4.21	4.13	4.85	4.22	4.32	4.97	5.04	4.31	5.32	5.53	4.98	5.07	5.17	5.07	5.50	5.56	5.30
Sweetness	4.48	5.33	5.89	3.62	4.89	5.44	2.47	3.37	5.01	1.73	1.44	3.17	2.58	1.80	2.44	1.42	1.62	1.69
Bitterness	4.48	4.64	4.72	4.68	4.78	4.60	4.59	4.70	4.71	3.96	4.36	4.65	4.35	4.47	4.60	4.02	3.98	4.36
MetallicTaste	3.21	3.20	2.91	3.18	3.28	3.16	3.41	3.42	3.15	3.47	3.70	3.41	3.51	3.40	3.34	3.74	3.68	3.50
SpicyTaste	5.52	5.48	5.66	5.21	5.57	5.63	5.06	5.20	5.63	4.01	3.78	5.03	4.90	4.11	4.83	3.42	3.87	4.13
MilkTaste	3.33	3.46	3.62	2.99	3.49	3.51	2.67	2.87	3.53	2.29	1.90	3.01	2.56	2.47	2.77	1.75	2.03	2.40
BurntTaste	2.69	3.22	3.50	2.10	2.96	2.85	1.82	2.21	3.02	1.20	1.04	2.16	1.60	1.08	1.74	1.08	1.13	1.45
Hardness	3.44	3.84	4.11	3.25	3.78	3.78	2.98	3.18	3.88	2.61	2.50	3.48	3.15	3.00	3.20	2.46	2.68	3.18
Juiciness	4.13	3.70	3.02	4.74	3.92	3.63	5.75	4.83	3.78	6.10	6.74	4.21	4.96	5.52	5.14	7.12	6.21	5.55
Graininess	6.34	7.17	7.40	5.85	6.48	7.13	4.72	5.46	6.56	3.99	3.78	5.90	5.28	4.78	5.35	3.30	4.06	4.92
Stickiness	3.45	3.66	3.70	3.30	3.54	3.65	2.71	3.40	3.24	2.57	2.21	3.60	3.10	2.79	3.00	1.95	2.53	2.91
Elasticity	1.84	1.53	1.44	2.17	1.73	1.49	2.71	2.04	1.69	2.92	3.46	1.85	2.17	2.70	2.35	3.76	3.09	2.90

Table 4: *Fish pudding: ANOVA table for the categorical variable model.*

Source	DF	SS	exVarSS	F	p-Value
Fish%	5	0.048455	0.3513	4.86	0.016262
Cost	2	0.069544	0.5042	17.45	0.000548
Error	10	0.019928	0.1445		

Table 5: *Fish pudding: Mean values with standard deviations according to the categorical variable model.*

Fish%	Mean	Std
35.0	3.859	0.026
37.5	3.851	0.026
40.0	3.837	0.026
42.5	3.734	0.026
45.0	3.807	0.026
47.5	3.734	0.026
Cost		
1	3.737	0.018
2	3.788	0.018
3	3.887	0.018

3.1 Balanced factorial design

Table 4 shows the analysis of the fish data (Table 2) with hardness as response and with Fish% and Cost as the two categorical design variables. The p-values are found from F-tests, which are based on the ordinary sums of squares (SS). The result is that both factors are significant.

When significance is found one wants to get more knowledge about the actual effects. Interesting information can already be seen from SS's. The relative size of these numbers reflects the relative importance of the factors. In order to make the SS's more readable we have introduced another column where the SS's have been divided by the total SS. The resulting numbers can be interpreted as explained variances and are therefore denoted as *exVarSS*. Note that exVarSS calculated for the whole model is identical to the ordinary R^2 measure. This means that our exVarSS's can be viewed as a decomposition of R^2 — similar to how the SS's can be viewed as a decomposition of the model SS.

To get even more specific knowledge about the effects, one could look at the estimated parameters of the underlying model or mean values derived from this model. In this balanced case we simply present the mean values (with standard deviations) within the levels of the category variables (Table 5). It is a clear tendency that Hardness increases as Fish% decreases and as Cost increases.

3.2 Unbalanced factorial design

We consider 13 observations of the cheese design (Table 1). The selected observations are the partly replicated design from blocks 1 and 2 (without center points). Preliminary analysis showed that factor T is undoubtedly most important. To illustrate the analysis of an unbalanced design we will therefore use a model with all main factors and all two-factor interactions involving T.

There are different ways to generalize the balanced ANOVA to unbalanced designs. The three main alternatives use sum of squares of Type I, II or III (SAS notation). The Type I analysis corresponds to adding each effect sequentially to the model and it depends on how the model terms are ordered. Different orders may give quite different results. This is a very useful method, but the results must be interpreted with care. As a standard method for reporting experimental results we will prefer a method that does not depend on the ordering of the model terms. Today, Type III is the usual method. However, this method is strongly criticized by Nelder (1977, 1994) and Nelder and Lane (1995). Langsrud (2003) concludes that the Type II method is preferable. Some important points are:

- For testing main effects, Type III assumes models which often are unrealistic — with interactions, but without both corresponding main effects. Type II takes the hierarchy of model terms into account.
- Type II is most powerful when the interactions are negligible. When interactions are present, Type II has still, on average, more power than Type III. In fact, Type III can lead to paradoxes where *"more information is worse than less"* (Senn, 1998).
- The historical reason why Type III has "won" the discussion is that Type III can be viewed as an optimal method for inferences according to a specific parametric formulation of a chosen model. However, *inference about parameters* is seldom the main objective. The aim of the data analysis is always *to answer questions.* For this purpose we need *to choose among different models.* Parameters are necessary for describing and using models. From such a broader viewpoint Type II is a more natural choice — since it is focused on choosing among the models.
- The Type III results depend on constraints (e.g. $\sum \alpha_i = 0$) used to parameterize the model. According to Nelder (1994) *"such constraints are not an intrinsic part of the model".*

Though, in cases with moderate unbalance (e.g. a balanced design with a few missing observations), the difference between Type II and Type III is of minor importance. However, when choosing a general method, we will also have in mind more extreme types of unbalance. Another reason for choosing Type II is the generalization to models with continuous variables, which are discussed in the section below.

In the above section we computed explained variances based on the SS's. Including the error term these add exactly up to one. Such explained variances based on Type I SS will still add up to one, but those according to Type II and Type III will not. Using Type II

Table 6: *Cheese: ANOVA table for the categorical variable model (unbalanced design).*

Source	DF	SS	exVarSS	F	p-Value
Block	1	0.00400	0.0003	0.01	0.918431
P	1	2.00894	0.1364	5.97	0.070950
S	1	0.27285	0.0185	0.81	0.418790
R	1	0.00099	0.0001	0.00	0.959302
T	1	9.94981	0.6758	29.57	0.005552
P*T	1	1.42007	0.0965	4.22	0.109157
S*T	1	0.38025	0.0258	1.13	0.347689
R*T	1	0.18579	0.0126	0.55	0.498726
Error	4	1.34600	0.0914		

SS we still think it is appropriate to report explained variances. Even if their sum is not one, they are useful measures of the importance of the different model components. In some sense, the SS and the corresponding p-value are two aspects of the same information and reporting the SS along with the p-value is very common. Our explained variances here are just scaled versions of the SS's. Type II p-values and the corresponding explained variances for our cheese example are presented in Table 6. Only T is significant, but P has a relatively low p-value.

When the design is unbalanced, illustrating effects by standard mean values is not longer useful. A common alternative is to compute least-squares means. These are estimates of mean values in a fully balanced design. A problem with this approach is that extremely unbalanced designs are not handled satisfactory. Consider, for example, a design where a few special variants (special level of factor A) are added to a large balanced design. The results of these special variants can be illustrated by adjusted mean values according to factor A. But we do not want a few special variants of factor A to have large influence on the adjusted means for factor B — which is the case for ordinary least-squares means.

We will therefore consider a slightly different approach where the means (of factor B) will be calculated according to weights (of factor A) that are proportional to the occurrence in the design. Such a method is already available — SAS's least-squares means with the OM-option (observed margins weighting). A drawback is that means for main factors cannot be estimated when interactions involving that factor are included in the model. Therefore we have, in Appendix B, made an extension of SAS's OM-option that defines adjusted means also in these cases.

Another reason for not using the traditional least-squares means is that these can not be directly generalized to continuous variable models. Two-level factors are interesting special cases. Consider a two-way design where factor A has two levels. When calculating the default least-squares means in SAS, the results for factor B depend on whether factor A is specified as continuous or categorical. Below we generalize our adjusted means to models with continuous design variables. Using this method, the results do not depend on how two-level factors are defined.

Our adjusted means are, for all the main factors, presented in Table 7. To illustrate

Table 7: *Cheese: Adjusted mean values with standard deviations according to the categorical variable model (unbalanced design).*

```
Block           Mean      Std
1              56.744    0.255
2              56.704    0.234
   P
3.15           56.277    0.239
3.50           57.106    0.221
   S
1.70           56.584    0.245
2.20           56.842    0.226
   R
0.00           56.719    0.221
7.00           56.728    0.239
   T
36.5           55.616    0.263
39.0           57.415    0.207
   P * T
3.15  36.5     54.641    0.415
3.15  39.0     57.233    0.291
3.50  36.5     56.351    0.348
3.50  39.0     57.533    0.291
```

two interacting factors, we can present adjusted mean values for all level combinations of these factors. Here we have chosen to present the adjusted mean values according to P and T.

3.3 Models with continuous variables

When all experimental factors can be treated as continuous variables, it is straightforward to formulate a polynomial model as an ordinary multiple regression model. Accordingly, it is straightforward to perform significance testing of the parameters (= Type III testing). However, such an analysis can give misleading results. An important problem is that the analysis is not invariant to scale changes (e.g. Celsius and Farenheit give different results). Especially, variables with mean values relatively far from zero lead to problematic dependence among the regressors. One way to reduce this problem is to center all the design variables before the analysis. But centering of design variables can not solve such dependence problems in general. E.g. in response surface analysis, A and AB^2 can be quite dependent — even if the design is balanced.

We prefer another approach — the Type II analysis. In the ordinary Type III analysis all model terms are adjusted for all other terms. In our Type II analysis each term is adjusted for all other terms except terms that “contain” the effect being tested. For example in a three-factor experiment (A, B and C), the term A is not adjusted for the interactions AB, AC and ABC. And the two-factor interactions are not adjusted for ABC.

Table 8: *Cheese: ANOVA table for the continuous variable model.*

```
Source    DF    exVarSS       F   p-Value
Block      2     0.0448    2.66  0.298150
P          1     0.1638    9.71  0.008170
S          1     0.0144    0.85  0.372108
R          1     0.0001    0.00  0.956868
T          1     0.5246   31.13  0.000089
P*T        1     0.0748    4.44  0.055203
S*T        1     0.0197    1.17  0.299426
R*T        1     0.0095    0.56  0.466837
T*T        1     0.0066    0.39  0.542859
Error     13     0.2191   13.00
```

This is equivalent to the Type II rule for categorical models. Furthermore, we now define A to be contained in A^2. This means that A^2 should be adjusted for A, but A should not be adjusted for A^2. This general rule is appropriate for models with both categorical and continuous variables. The Type II analysis above (Table 6) is a special case. However, this generalized Type II analysis can not be named as Type II. This name was invented by SAS and their general Type II method apply the Type II rule only for categorical variables. Type II in SAS applied to a response surface model produces exactly the same output as the ordinary Type III method. Instead of Type II, we name our general method as Type II*. This method has several nice properties:

- Invariant to ordering of the model terms.
- Invariant to scale changes (centering has no effect).
- Invariant to how the overparameterization problem of categorical variable models is solved (how constraints are defined).
- Whether two-level factors are defined to be continuos or categorical does not influence the results.
- Analysis of a polynomial model with a single experimental variable $(x, x^2, x^3, \ldots)$ produce results equivalent to the results using an orthogonal polynomial.

The more standard Type III method shares only the first of these properties. Further remarks on Type II* testing including relation to other literature is given in Appendix A.

Recalling the cheese example above (Table 6) we now define the four design factors as continuous. To illustrate this analysis we will again present the results for one model. Since we know that factor T is most important, we extend the model in Table 6 with the quadratic term, T^2. This time the analysis uses all the 24 observations of the cheese design (Table 1). Explained variances and p-values are presented in Table 8. Now both P and T are clearly significant and P*T is nearly significant at the 5% level.

With continuous variables, the effects can not be illustrated by (adjusted) mean values as for categorical variables. But we can easily calculate predicted values at certain levels of the design variables. Now, to illustrate the effect of a single factor, we will calculate means of such predictions. For factor P we present mean predictions at the minimal and the maximal value of P. When P=3.15 and when P=3.50, the mean predictions are 56.095 and 56.951 respectively. Looking at these numbers it is easy to see how P affects the response.

If the other terms in the model are only continuos main effects, our mean predictions could be easily obtained by setting those factors to their mean value and calculate the predictions directly. However, higher order terms makes it more complicated. The mean predictions are therefore defined similar to the adjusted means above.

Similar to how P was illustrated, we can illustrate the effect of T. The quadratic term is, however, included in the model. When this term is important, we suggest that an extra mean prediction is presented — at middle between the to extremes. Although, T^2 is far from being significant, we have presented such a prediction in Table 9. We can see that the prediction at the middle (56.593) is not very far from the mean of the predictions at the two extremes (55.487 and 57.245). This is as expected when T^2 is not important.

Since the model involves P*T it is not straightforward how to calculate mean predictions for P and T. The problem is similar to the problem of calculating adjusted means. Appendix B presents a general method that covers both categorical and continuous variables. With P*T in the model it could be useful to present mean predictions at level combinations of P and T. In Table 9 we have chosen to combine three levels of T (since T^2 is in the model) with two levels of P. Looking at these numbers we see that the effect of T is slightly larger when P=3.15 compared to P=3.50.

To analyze the fish data we use a full second order model in Fish% and Cost. In addition Day is included as a block effect. Results are shown in Table 10 and Table 11. Again, since the model contain quadratic terms, we show mean predictions at three levels of the continuous factors. The conclusion is, however, that only the two linear main effects are important.

4 MULTIPLE RESPONSES

This section describes the multivariate generalization of the above univariate methodology. We describe two significance testing approaches — a single p-value for all responses (50-50 MANOVA) and adjusted single response p-values for each response (rotation testing). Furthermore we describe two ways of illustrating multivariate means or mean predictions (curve plotting and principal component plotting).

4.1 50-50 MANOVA

With multiple responses it can be useful to perform a principal component analysis (PCA) (Martens and Næs, 1989). The data is then decomposed as vectors of scores and loadings. The scores are the transformed response values according to the principal components.

Table 9: *Cheese: Mean predictions with standard deviations according the continuous variable model.*

Block		Mean	Std
1		56.759	0.313
2		56.726	0.179
3		56.124	0.302
P			
3.15		56.095	0.185
3.50		56.951	0.169
S			
1.70		56.423	0.209
2.20		56.683	0.197
R			
0.00		56.560	0.158
7.00		56.555	0.244
T			
36.50		55.487	0.269
37.75		56.593	0.240
39.00		57.245	0.235
P * T			
3.150	36.50	54.494	0.421
3.150	37.75	56.005	0.279
3.150	39.00	57.062	0.284
3.500	36.50	56.173	0.307
3.500	37.75	56.985	0.273
3.500	39.00	57.344	0.315

Table 10: *Fish pudding: ANOVA table for the continuous variable model.*

Source	DF	exVarSS	F	p-Value
Day	1	0.0001	0.00	0.958269
Fish%	1	0.2290	10.91	0.007031
Cost	1	0.4873	23.22	0.000537
Fish%*Cost	1	0.0344	1.64	0.226919
Fish%*Fish%	1	0.0001	0.01	0.941576
Cost*Cost	1	0.0170	0.81	0.387976
Error	11	0.2308		

Table 11: *Fish pudding: Mean predictions with standard deviations according to the continuous variable model.*

Day	Mean	Std
1	3.804	0.018
2	3.803	0.018
Fish%		
35.00	3.867	0.028
41.25	3.803	0.020
47.50	3.743	0.028
Cost		
1	3.737	0.022
2	3.788	0.022
3	3.887	0.022

The loadings relate the principal components (PC's) to the original variables. In many situations, a lot of information is contained in the first two components. Therefore plots of scores and loadings for these components can be very useful (see the section on principal component plotting below).

To analyze multiresponse data one possibility is to analyze the first few score vectors as univariate responses (Ellekjær *et al.*, 1996). Using this method, the number of score vectors to consider must be chosen. When looking at the p-value for several principal components to pick out the smallest, we are in a situation where lack of error rate control can be a problem. Therefore, to apply this method correctly these p-values should be corrected for multiplicity by e.g. using Bonferroni correction. That is, multiply the smallest p-value by the number of components.

Inspired by this method, Langsrud (2002) developed a generalized multivariate ANOVA method named as 50-50 MANOVA. Instead of testing single PC's by separate tests, one multivariate test for several PC's is performed. This means that 50-50 MANOVA looks into a space of PC's rather than single PC directions. The number of PC's ($= nPC$) for this test is based on an explained variance criterion associated with the PCA decomposition. Since it can be difficult to choose nPC "correctly", a group of nBu buffer components is introduced. The inclusion of buffer components will reduce the power loss caused by choosing too few components for testing.

To describe 50-50 MANOVA in more detail, we consider the special case of a single continuous design factor (x). Note that the p-value of the classical Hotelling T^2 test (Mardia *et al.*, 1979) can be obtained by reversing the regression model. That is, we can do multiple regressing with x as the single response. The p-value is obtained by the standard F-test corresponding to the whole regression model. The 50-50 MANOVA p-value can also be obtained by reversing the regression model. For simplicity, we will assume that the number of responses exceeds the number, n, of observations. In this case, the reversed model is a principal components regression (PCR) model (Martens and Næs, 1989). The total sum of squares of x can then be decomposed according to the PC's as

$SS_1 + SS_2 + \cdots + SS_{n-1}$. The p-value of the 50-50 MANOVA test with k components and d buffer components can now be obtained by doing an F-test using the statistic

$$F = \frac{(SS_1 + SS_2 + \cdots + SS_k)/k}{(SS_{k+d+1} + SS_{k+d+2} + \cdots + SS_{n-1})/(n-1-k-d)} \tag{1}$$

We can see that the effect of k components are combined in the numerator. The buffer components $(SS_{k+1}, \ldots, SS_{k+d})$ are neither in the numerator nor in the denominator. For comparison, recall the method mentioned above that uses a single score vector (PC number k) as univariate response. The p-value of this analysis can also be obtained by using the reversed PCR model. Then, the F statistic has SS_k in the numerator and all the other $n-2$ SS's in the denominator.

For models with several model terms, the PCA decomposition in 50-50 MANOVA is not based on the original response data. The decomposition is instead calculated from residual data obtained by fitting a model with all terms except the one that is tested. Accordingly, a separate PCA is performed for each model term in the ANOVA table. This means that the variation caused by factor A is regarded as irrelevant for testing factor B and vice versa. Some details (not covered in Langsrud (2002)) on the criterion for choosing the number of components and how to calculate the p-value are given in Appendix C.

The 50-50 MANOVA method is a true generalization of the univariate F-test. Similar to how an F-test is based on sums of squares for hypothesis and error, these multivariate tests are based on two (hypothesis and error) matrices of sums of squares and cross-products (Mardia *et al.*, 1979). We will also formulate the other elements of the analysis as a generalization of the above univariate methodology. The testing will follow the Type II* principle. We will compute the explained variance (exVarSS) for each model term as the SS for the term summed over all responses divided by the sum of the total SS's. This measure is independent of the 50-50 MANOVA testing. We will, however, compute additional explained variances associated with the PCA underlying the multivariate testing. We report both the explained variance after nPC components (exVarPC) and the explained variance after nPC+nBu components (exVarBu). These explained variances are calculated from the eigenvalues corresponding to the PCA. It is important to note that exVarPC and exVarBu have no direct relation to exVarSS. The former depends on the correlations between responses and the number of components (nPC and nBU). The latter depends on how the design variable affects the responses.

Table 12 presents the 50-50 MANOVA results of the fish pudding design (Table 2) with the 24 sensory attributes (Table 3) as responses and with the same model as in Table 10. Again the two linear main effects are most important. The interaction and the quadratic Fish% effect are also significant.

The 50-50 MANOVA results of the cheese design with the 477 spectroscopic wavenumbers as responses are presented in Table 13 (same model as in Table 8). Factor T is again most important. Now there is also a very important block effect and we have significant interaction between S and T. In this example we have for all terms that $nPC = 2$. It is not unusual that nPC is not varying much — a large amount of the data source for PCA is common. Note, however, that $nPC = 2$ will be quite common when using the rule in

Table 12: *Fish pudding: 50-50 MANOVA table for the continuous variable model with responses from sensory evaluation.*

Source	DF	exVarSS	nPC	nBu	exVarPC	exVarBU	p-Value
Day	1	0.0040	2	3	0.898	0.969	0.472886
Fish%	1	0.6661	1	3	0.968	0.991	0.000012
Cost	1	0.1221	1	3	0.910	0.972	0.004913
Fish%*Cost	1	0.0097	2	3	0.897	0.969	0.017699
Fish%*Fish%	1	0.0101	2	3	0.890	0.967	0.009342
Cost*Cost	1	0.0171	2	3	0.910	0.970	0.648315
Error	11	0.1582					

Table 13: *Cheese: 50-50 MANOVA table for the continuous variable model with responses from IR spectroscopy.*

Source	DF	exVarSS	nPC	nBu	exVarPC	exVarBU	p-Value
Block	2	0.1927	2	4	0.775	0.972	0.000000
P	1	0.0534	2	4	0.825	0.965	0.064514
S	1	0.0375	2	4	0.803	0.965	0.060738
R	1	0.0354	2	4	0.791	0.965	0.118843
T	1	0.3150	2	4	0.883	0.980	0.000001
P*T	1	0.0239	2	4	0.781	0.963	0.401477
S*T	1	0.0457	2	4	0.748	0.963	0.006130
R*T	1	0.0128	2	4	0.808	0.963	0.700330
T*T	1	0.0091	2	4	0.792	0.957	0.903666
Error	13	0.2741					

Appendix C for selecting nPC. The default rule says that 50% of the variance should be explained, but 90% is required to trust in a single component.

4.2 Curve plotting

In the univariate case we computed means and mean prediction adjusted for the unbalance. With several responses we may compute such means for each response. When these responses are digitizations of continuous curves we can draw mean curves. Such mean curves for our spectroscopic data are presented in Figure 3. Four level combinations of S and T are shown. These curves are based purely on the unviariate methodology (separate calculation for each response).

Even if the multivariate tests gave significance, we can not be sure which part of these curves that are really interpretable. A possible extension of the curve plotting is to add information about the standard deviations. Another possibility is to compute adjusted single response p-values as described below.

4.3 Principal component plotting

As mention above one could analyze principal component scores as univariate responses. When doing such an analysis we can compute means or mean predictions for each principal component. A mean value for the first component and the corresponding mean value for the second component are then co-ordinates for a point in the score plot. That is, means and mean predictions can be illustrated in a score plot. Note that this method is based on PCA of the original data and is therefore different from the PCA's underlying 50-50 MANOVA — unless the model has only a single term.

Figures 4 and 5 show the loading plot and the score plot for the sensory data of the fish pudding design. To illustrate the effect of Fish% we have added means according to three levels of Fish%. The middle point (41.25) is included since we have a significant quadratic effect.

We can see that all the scores lie more or less along a curve in the plot. The three mean points also tend to follow this curve. However, when interpreting this plot one should have in mind that the first component contains most of the information (95.2%) and the linear effect of Fish% is most important. A possible extension of this plot is to add standard deviation information to the mean points.

4.4 Adjusted p-values by rotation testing

Instead of doing multivariate testing, such as 50-50 MANOVA, one could analyze each response by ordinary univariate F-tests. Then, the Type I error rate is controlled by the significance level individually for each response and for each model term. With a large number of responses, we will expect several Type I errors and it is questionable whether significant results can be interpreted as real effects. We can avoid this problem by adjusting the p-values so that the familywise Type I error rate (FWE) is controlled. This means that,

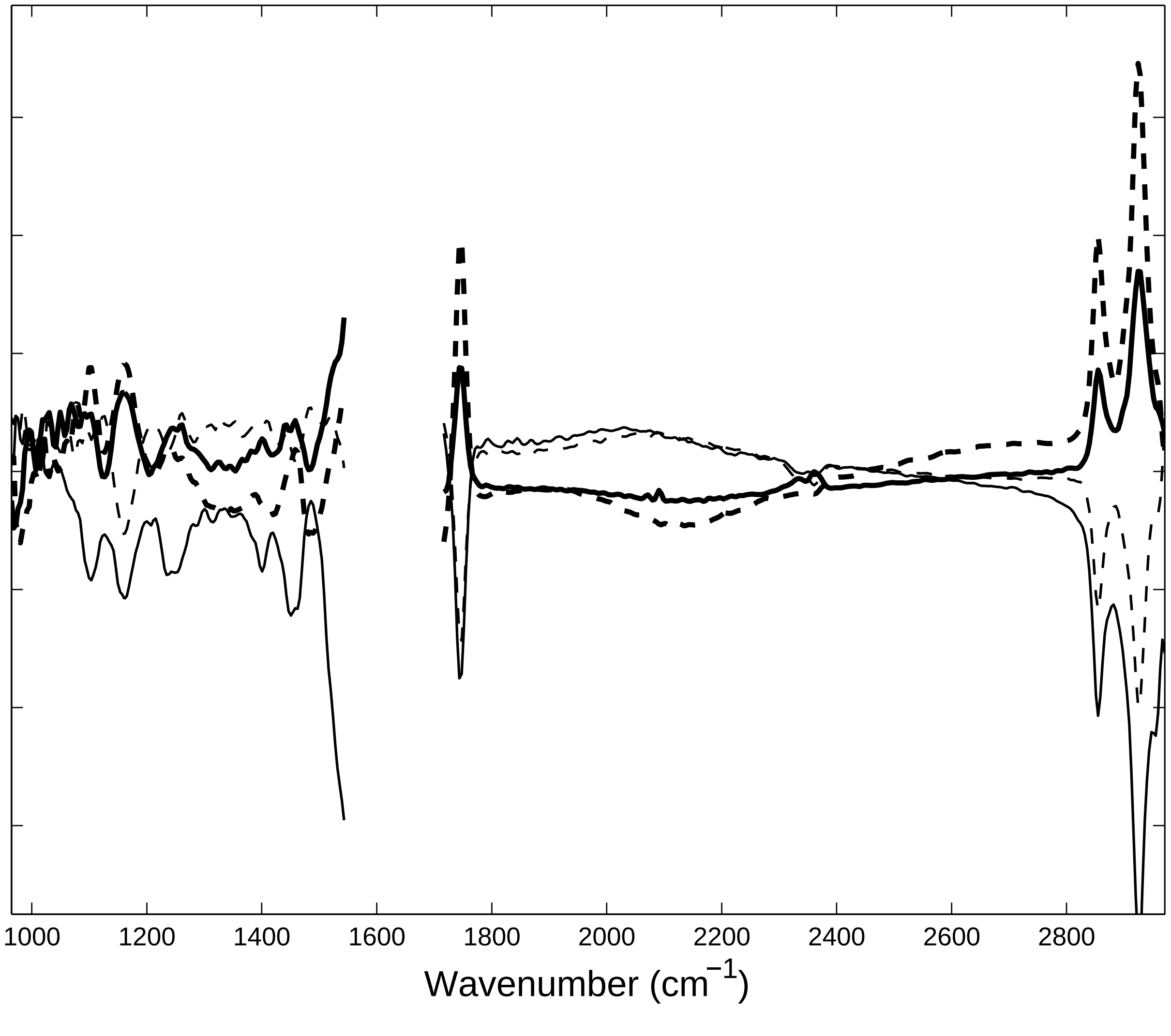

Figure 3: *Cheese: Mean spectra according to level combinations of S and T. The levels of S are 1.7 (solid lines) and 2.2 (dashed lines). The levels of T are 36.5 (thin lines) and 39.0 (thick lines).*

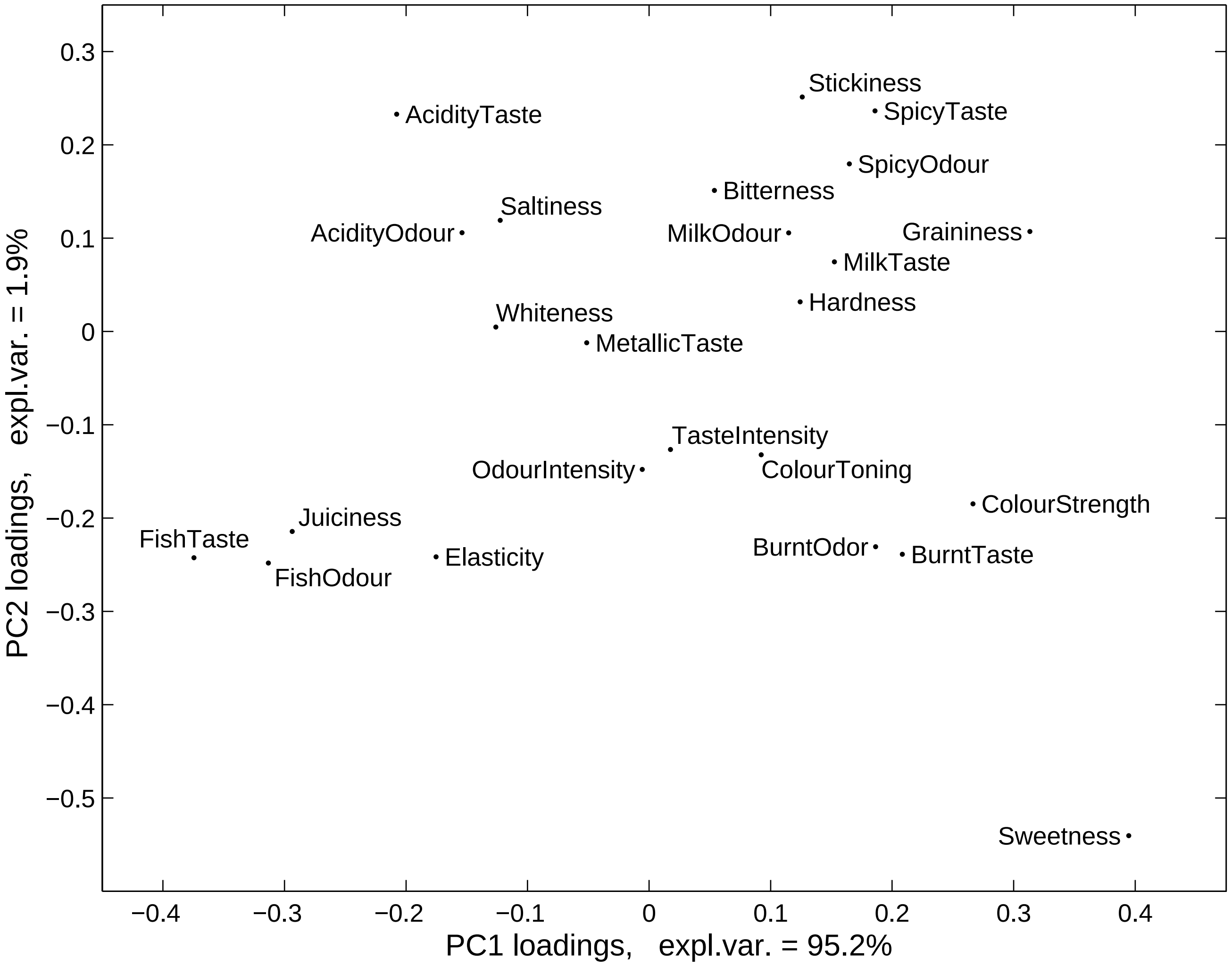

Figure 4: *Fish pudding: Loadings from PCA of the sensory responses.*

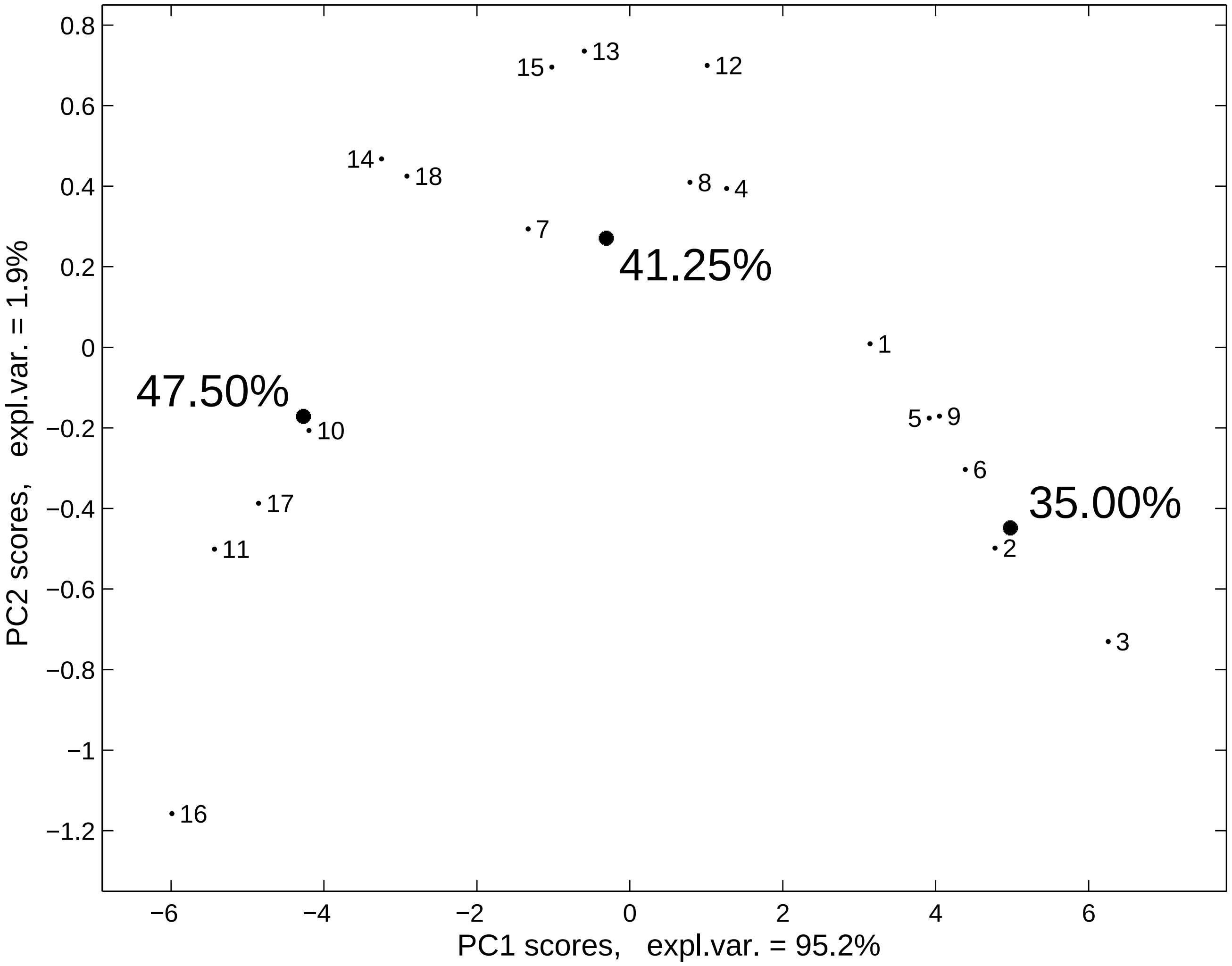

Figure 5: *Fish pudding: Scores from PCA of the sensory responses together with mean scores for three Fish% levels.*

for each model term, the probability of at least one Type I error among all the responses does not exceed the significance level. When adjusting the p-values by using Bonferroni's correction, the p-values are simply multiplied by the number of responses. This classical method is, however, extremely conservative and becomes useless in cases with a large number of highly correlated responses. Recently, Langsrud (2005) has described an exact and non-conservative method based on so-called rotation testing. Adjusted p-values are then calculated by a simulation procedure that makes use of the dependence among the responses.

Note that rotation testing is a framework for doing significance testing using computer simulations and is closely related to permutation testing (Anderson, 2001). Rotation testing relies, however, directly on the multivariate normal distribution. This means that all the classical normal based tests (F-tests, Hotelling T^2-tests and the various MANOVA tests) are special cases of rotation testing — except that simulations are used to calculate the p-values. Rotation testing is useful when the p-value cannot be calculated directly and an important application is the adjustment of univariate F-test p-values in multiresponse experiments.

In cases where the number of responses is very high, the FWE criterion can be viewed as being too strict. Instead of considering the probability of at least one type I error, an alternative is to estimate the false discovery rate (FDR) which is the (expected) proportion of type I errors among all responses reported as significant.

When using adjusted p-values according to FDR, one accepts that 5% of the responses reported as significant at the 5% level are type I errors. As described in Moen *et al.* (2005), rotation testing can also be used to adjust the p-values according to an FDR criterion. Compared to other FDR variants (Storey and Tibshirani, 2003), an advantage of this rotation testing method is that any kind of dependence among the responses is allowed.

According to the above modeling of the spectroscopic data (Table 13), we have calculated adjusted p-values for the model term S*T. Table 14 contains the results for the 20 most significant wavenumbers. Both Bonferroni corrected and the two variants (FWE and FDR) of rotation adjusted (10^6 simulations) p-values are presented. We can clearly see the conservativeness of Bonferroni. Though, this method also gave significant results. At the left part of Figure 3, the curves seem rather chaotic. As seen in Table 14, the S*T interaction is, however, significant for several wavenumbers in this area. Therefore, the chaotic structure is not just noise.

5 CONCLUDING REMARKS

Our approach is invariant to scale changes of the design variables. Scale changes of the response variables will, however, affect the results of PCA and 50-50 MANOVA. When the response variables are measured on different scales or a common scale with widely differing ranges, the data should be standardized before PCA is applied. In other words, PCA is then based on a correlation matrix rather than a covariance matrix. In these cases one

Table 14: *Cheese: P-values for the S*T interaction using the continuous variable model. Ordinary raw (pRaw), Bonferroni adjusted (pBon), familywise adjusted (pAdjFWE) and false discovery rate adjusted (pAdjFDR) p-values are shown for the 20 most significant IR wavenumbers.*

rank	wavenumber	pRaw	pBon	pAdjFWE	pAdjFDR
1	1528	0.000050	0.023850	0.00358	0.002778
2	1543	0.000057	0.027189	0.00393	0.002778
3	1532	0.000057	0.027189	0.00393	0.002778
4	1539	0.000067	0.031959	0.00452	0.002827
5	1535	0.000082	0.039114	0.00533	0.002921
6	1524	0.000087	0.041499	0.00558	0.002921
7	1520	0.000181	0.086337	0.01017	0.005168
8	1443	0.000196	0.093492	0.01078	0.005187
9	1516	0.000268	0.127836	0.01386	0.006207
10	1447	0.000274	0.130698	0.01407	0.006207
11	1254	0.000373	0.177921	0.01814	0.007781
12	1250	0.000402	0.191754	0.01927	0.007946
13	1439	0.000525	0.250425	0.02384	0.009662
14	1246	0.000605	0.288585	0.02680	0.010583
15	1258	0.000648	0.309096	0.02826	0.010852
16	1242	0.000808	0.385416	0.03391	0.012866
17	1512	0.000860	0.410220	0.03550	0.013193
18	1262	0.001140	0.543780	0.04431	0.016571
19	1451	0.001316	0.627732	0.04951	0.018363
20	1265	0.001722	0.821394	0.06078	0.022838

should also use the standardization option when applying 50-50 MANOVA. This means that the PCA's performed within the 50-50 MANOVA method is based on correlation matrices. Note that the adjusted p-values calculated by rotation testing are independent to scale changes.

The cheese example was partly based on fractional factorial designs. In the present paper we have not treated the special problems of such designs. The new problems of multiple responses are, however, independent of the question of confounding, which is related to the model and the design matrix. Another problem of fractional factorial designs is the lack of degrees of freedom (DF) for error. The approach in the present paper assumes that some DF's for error are available. It is possible to perform ordinary ANOVA with a single error DF, but it is well know that such an analysis has very low power. This power problem is similar for 50-50 MANOVA and rotation testing. Note that a method for analyzing fractional designs with multiple responses and with no (or few) error DF's is described in Langsrud (2001). That method relates to 50-50 MANOVA very similar to how the univariate method of Langsrud and Næs (1998) relates to ANOVA.

Split plot designs and designs with random factors is another topic that is not treated above. Often, the univariate solution is to change the error terms of the F-tests. In these cases we can apply 50-50 MANOVA and rotation testing with similar error terms — since both 50-50 MANOVA and rotation testing are generalizations of the univariate F-test. Our current software (available at www.matforsk.no/ola) handles only one error term. As described in Bjerke *et al.* (IN PRESS) a practical solution to this problem, in the case of split plot designs, is to average responses over the whole plot experimental units.

Detecting outliers and discovering the need for transformations are other important issues. One possibility is to plot residuals as curves or to perform PCA of residuals. One could also apply univariate methodology to individual responses or to PCA scores. Further research is, however, needed in this problem area.

Appendix A: Some remarks on Type II* testing

As pointed out by Peixoto (1990) and Nelder (2000), the model for the analysis should be well-formed. This means e.g. that if the model contain A^2BC, the model should also contain A^2B, A^2C, ABC, A^2, AB, AC, BC, A, B and C. If this rule is not followed, the model's fit to the data is scale-dependent and strong assumptions of marginal relations are made (also relevant for categorical variables). Therefore, one should have good reasons for making exceptions to this rule. The statement above saying that Type II* is scale-independent assumes that this rule is followed. Note that Type II* can be said to based on the well-formed rule since the null hypotheses underlying the significance tests can be formulated without breaking this rule (see Langsrud (2003)).

Aiken and West (1991) address the scale invariance problem of ordinary (Type III) testing. To analyze complex models they strongly recommend the well-formed rule and they suggest a hierarchical step-down procedure. Only the scale-independent terms (i.e. highest order terms) are allowed to be tested. The non-significant terms are eliminated and the model is revised. In other words, Aiken and West allow only the Type III tests that are equivalent to our Type II* tests. Since, in addition, our Type II* SS for term A is unchanged when eliminating higher order terms involving A, the procedure of Aiken and West relates very closely to Type II* testing. An important difference is, however, that Aiken and West do not allow a first order term to be tested if there is a significant interaction involving that term. As follows from the well-formed rule, we agree that there is no reason to hypothesize a zero first order term when there is a known interaction. But if the method of Aiken and West is automated, a strictly defined significance level is needed. By doing Type II* testing for all terms we leave to user to interpret the various p-values as degree of significance in light of plausible model alternatives.

Note that our Type II* method are almost identical to the default ANOVA approach (named as Type II) implemented in the “car” library for R and S-plus (Fox, 2002). The difference is that “car” will not recognize linear terms (A) as being contained in quadratic terms (A*A). To perform a separate test of each model term, such quadratic (and cubic) terms must be specified as new variables.

Appendix B: Adjusted means and mean predictions

To compute adjusted means we consider the following algorithm.

- Parameterize the model as a full rank multiple regression model: $y = \boldsymbol{XB}$. This means that each model term use a number of x-variables corresponding to its degrees of freedom (constant term is represented by the first variable). Estimate the regression parameters as usual. Predictions can now be made by $\widehat{y_{\text{new}}} = x_{\text{new}}\widehat{\boldsymbol{B}}$ where x_{new} is a row vector of new x-values.

- A mean prediction for a certain level of a certain variable (or a level combination of several variables) can be obtained as $\widehat{y_{\text{new}}}$ by using a specific x_{new}-vector. The variable(s) involved in this prediction specify(s) x_{new} partially. For example, with reference to Table 7, we now consider the mean prediction at P=3.15 and T=36.5. The elements of x_{new} that corresponds to the main effect of P, the main effect of T and the interaction between them is then specified. The unspecified elements of x_{new} are simply set to the mean value of those x-variables (even for dummy variables).

This algorithm results in the same mean values as obtained by using SAS’s least-squares means with the OM-option. Furthermore, this algorithm is appropriate for illustrating effects of continuous variables.

The difficulty occurs when one wants to illustrate the effect of, say A, in a model containing interactions involving A. Using the above algorithm one would just plug in mean values for x-variables corresponding to these interaction terms. In that case, the resulting means would be an illustration of the parameter estimates corresponding to the main effect of A. This could be a misleading illustration of A’s effect. Different parameterizations of the model give different results. The problem is that parameters corresponding to the main factor and the interaction parameters can be very dependent.

To solve this problem we will require a certain parameterization of the model. We will orthogonalize the x-variables according to the model hierarchy. For example, the x-variables for A*B are made orthogonal to the intercept (vector of ones), the main A variable(s) and the main B variable(s). This way the above dependence problem is eliminated. With this modification we will apply the above algorithm to compute various types of adjusted means. It is interesting to note that, since all the x-variables are made orthogonal to the constant term (except for the constant term itself), the mean value of all variables are zero. That is, the parameters corresponding to the unspecified elements of x_{new} are not involved in the calculations.

Appendix C: 50-50 MANOVA details

The rules for selecting the number of components (k) and the number of buffer components (d) can be written as

- Choose $k = 1$ if $g(k) \geq 0.90$. Otherwise, choose the smallest $k > 1$ so that $g(k) \geq 0.50$.
- Choose $d = \min(\widetilde{d}, r - k)$, where $\widetilde{d} = (r_{\max} - \nu - k - 3)/2$ (truncated).

Here ν is the DF for the term that is tested and we have defined

$$g(k) = \frac{\sum_{i=1}^{k} e_i}{\sum_{i=1}^{k} e_i + c(k) \sum_{i=k+1}^{r} e_i} \tag{2}$$

where the e_i's are eigenvalues corresponding to the PCA underlying the test and where r is the corresponding rank. The number $r_{\max}$ is the maximum possible value of r. That is, $r_{\max}$ equals the value of r in cases where the number of responses exceeds the number of observations. In such cases we will always have that $c(k) = 1$ and the rule above is identical to the rule described in Langsrud (2002). However, it was mentioned in Langsrud (2002) that the rule should be modified when we have few responses. The idea is that 50-50 MANOVA should be made more similar to classical MANOVA in these cases. The factor c that we have introduced here represent this modification and we define $c(k) = \left(\sum_{i=k+1}^{r_{\max}} i^{-1}\right) / \left(\sum_{i=k+1}^{r} i^{-1}\right)$. This rule is an intuitive suggestion and seems to work satisfactory. Optimal rules for selecting k and d can, however, be an interesting topic for future research.

Note the 50-50 MANOVA can be viewed as a method that reduces dimensions in classical MANOVA. The final p-value is always calculated according to an ordinary MANOVA test statistic. There exist several test statistics for this testing problem and the four most popular are reviewed by Olson (1976). These are: *Wilks' Λ*, *Roy's Largest Root*, *Hotelling-Lawley Trace Statistic* and *Pillay-Bartlett Trace Statistic* . When one test statistic is chosen, Wilks' Λ is very common. On the other hand we find Roy's Largest Root intuitively appealing since it focuses on the most important dimension of the “MANOVA space”. However, our choice for a single test statistic is Hotelling-Lawley Trace Statistic. This statistic can be viewed as lying between Wilks' Λ and Roy's Largest Root. Note that the four statistics are equivalent whenever $nPC = 1$ or the DF for the tested term is one.

ACKNOWLEDGEMENTS

The work in this paper is a part of the IBION project (Industrial Biostatistics Network) financed in part by the Research Council of Norway through grant 145456/130. The work was also partially supported by The Norwegian Research Council's grant 132975/112 (Development of Methods for Recipe optimization in the Food Industry). Dr. Frøydis Bjerke at Matforsk is thanked for her helpful comments.

REFERENCES

Aiken, L.S. & West, S.G. (1991), *Multiple regression: Testing and interpreting interactions*, Sage Publications, Newbury Park.

Anderson, M. J., (2001), Permutation Tests for Univariate or Multivariate Analysis of Variance and Regression, *Canadian Journal of Fisheries and Aquatic Sciences*, 58, pp. 626–639.

Bjerke, F., Langsrud, Ø. & Aastveit, A. H. (IN PRESS), Restricted randomisation and multiple responses in industrial experiments, *Quality and Reliability Engineering International.*

Box, G. E. P., Hunter, W. G. & Hunter, J. S. (1978), *Statistics for Experimenters*, John Wiley and Sons, New York.

Ellekjær, M. R., Ilseng, M. A. & Næs, T. (1996), A case study of the use of experimental design and multivariate analysis in product development, *Food Quality and Preferences*, 7, pp. 29–36.

Fox, J. (2002), *An R and S-PLUS Companion to Applied Regression*, Sage Publications, Thousand Oaks.

Langsrud Ø. (2001), Identifying significant effects in fractional factorial multiresponse experiments, *Technometrics*, 43, pp. 415–424.

Langsrud, Ø. (2002), 50-50 Multivariate Analysis of Variance for Collinear Responses, *Journal of The Royal Statistical Society Series D – The Statistician*, 51, pp. 305–317.

Langsrud, Ø. (2003), ANOVA for Unbalanced Data: Use Type II Instead of Type III Sums of Squares, *Statistics and Computing*, 13, pp. 163–167.

Langsrud, Ø. (2005), Rotation Tests, *Statistics and Computing*, 15, pp. 53–60.

Langsrud, Ø. & Næs, T. (1998), A Unified Framework for Significance Testing in Fractional Factorials, *Computational Statistics and Data Analysis*, 28, pp. 413–431.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press Limited, London.

Martens, H. & Næs, T. (1989), *Multivariate Calibration*, John Wiley and Sons, New York.

Moen, B., Oust, A., Langsrud, Ø., Dorrell, N., Gemma, L., Marsden, G.L., Hinds, J., Kohler, A., Wren, B.W. & Rudi, K. (2005), An explorative multifactor approach for investigating global survival mechanisms of Campylobacter jejuni under environmental conditions, *Applied and Environmental Microbiology*, 71, pp. 2086–2094.

Montgomery, D. C. (1991), *Design and Analysis of Experiments, 3rd. ed.*, John Wiley and Sons, New York.

Nair, V. N., Taam, W. & Ye, K. Q. (2002), Analysis of functional responses from robust design studies, *Journal of Quality Technology*, 34, pp. 355-370.

Nelder, J. A. (1977), A Reformulation of Linear Models (with discussion), *Journal of the Royal Statistical Society Series A* , 140, pp. 48–77.

Nelder, J. A. (1994), The statistics of linear models: back to basics, *Statistics and Computing*, 4, pp. 221–234.

Nelder, J. A. (2000), Functional marginality and response-surface fitting, *Journal of Applied Statistics*, 27, pp. 109–112.

Nelder, J. A. & Lane, P. W. (1995), The Computer Analysis of Factorial Experiments: In Memoriam — Frank Yates, *The American Statistician*, 49, pp. 382–385.

Olson, L. (1976), On choosing a test statistic in multivariate analysis of variance, *Psychological Bulletin*, 83, pp. 579–586.

Peixoto, J. L. (1990), A property of well-formulated polynomial regression-models, *The American Statistician*, 44, pp. 26–30.

Senn, S.(1998), Some Controversies in Planning and Analysing Multi-Centre Trials, *Statistics in Medicine*, 17, pp. 1753–1765.

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

Storey, J. D., & Tibshirani., R. (2003), Statistical significance for genomewide studies, *Proceedings of the National Academy of Sciences of the United States of America*, 100, pp. 9440–9445.