Explaining Correlations by Plotting Orthogonal Contrasts

Øyvind Langsrud*
MATFORSK, Norwegian Food Research Institute.
www.matforsk.no/ola/

To appear in
The American Statistician
www.amstat.org/publications/tas/

ABSTRACT

This article describes a new plot that aids understanding the relationship between two response variables in a designed experiment. In addition to plotting the observed values directly, we make a scatter plot of orthogonal contrasts from the general linear model. This plot contains the same correlation information as the ordinary scatter plot. Therefore, one can interpret how the effects of the various design variables contribute to the correlation coefficient. This idea is also useful in more general cases. Any graphic presentation of the original observations can be accompanied by a corresponding plot of orthogonal contrasts that often will clarify the interpretation.

KEY WORDS : Design of experiments, Fractional factorial design, Scatterplot, General linear model, Partial least squares, Principal component analysis.

*Øyvind Langsrud is a research scientist at MATFORSK (Norwegian Food Research Institute), Osloveien 1, N-1430 Ås, NORWAY (E-mail: *oyvind.langsrud@matforsk.no*). The author thanks Per Lea for helpful comments. The author is also grateful to the editors and referees of *The American Statistician* for their valuable comments.

1 INTRODUCTION

This article introduces new plotting methodology that is useful for illustrating relationships between responses in designed experiments. The main idea is to make plots of orthogonal contrasts for pairs of response variables that yield the same correlations as the ordinary plots of observed values. In Section 2 we start the discussion by making scatter plots of estimated effects from a fractional factorial design. A more general situation is treated in Section 3 where orthogonal contrasts are derived from a response surface model. Section 4 applies the new scatter plot to chemometrics regression; the relationship between predicted and measured values is illustrated. In Section 5, the concept of plotting orthogonal contrasts is extended to principal component analysis. Section 6 concludes with some final remarks.

2 PLOTTING EFFECTS IN TWO-LEVEL DESIGNS

We consider a fractional factorial 2^{5-1} design (Box *et al.*, 1978), which has been analyzed in Langsrud (2001). The effect of five different ingredients or processing factors on the sensory quality of baguettes was studied. Our response variables are 16 sensory attributes that were evaluated by a sensory panel. For each sensory response we can calculate 15 estimated effects; five main effects, seven (confounded) two-factor interactions and three higher order interactions. An interesting question is how to identify significant effects, but here we focus on illustrating the correlation between some of the responses.

We will first note that the uncentered correlation between two $n \times 1$ data vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$, can be expressed as

$$\widetilde{r}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T \boldsymbol{a}\ \boldsymbol{b}^T \boldsymbol{b}}} \ . \tag{1}$$

To calculate the ordinary correlation between two response variables, we need to center the data (subtracting the means) before applying the above function.

An ordinary scatter plot of two highly correlated (r=0.9861) responses, Garlic Flavour and Odour Intensity, is presented in Figure 1(a). We see that the main cause of the high correlation is a separation into two clusters. It turns out that this grouping corresponds to one of the design factors, namely *garlic content*.

Figure 1(b) shows a corresponding scatter plot of the estimated effects. In this figure, each of the 15 points corresponds to one of the 15 estimated effects. For instance, the point at the upper-right corner plots the effect of garlic content on the Garlic Flavour against the effect of garlic content on the Odour Intensity. By assuming a regression line through the origin, we can interpret this plot as an another illustration of the correlation between the two responses. That is, the above correlation coefficient (0.9861) can be computed from the estimated effects by applying expression (1) directly without centering. In the section below, we discuss this relationship further.

The two figures represent two different ways of illustrating that garlic content is causing correlation. In Figure 1(a) we can also see that the correlation within each group seems

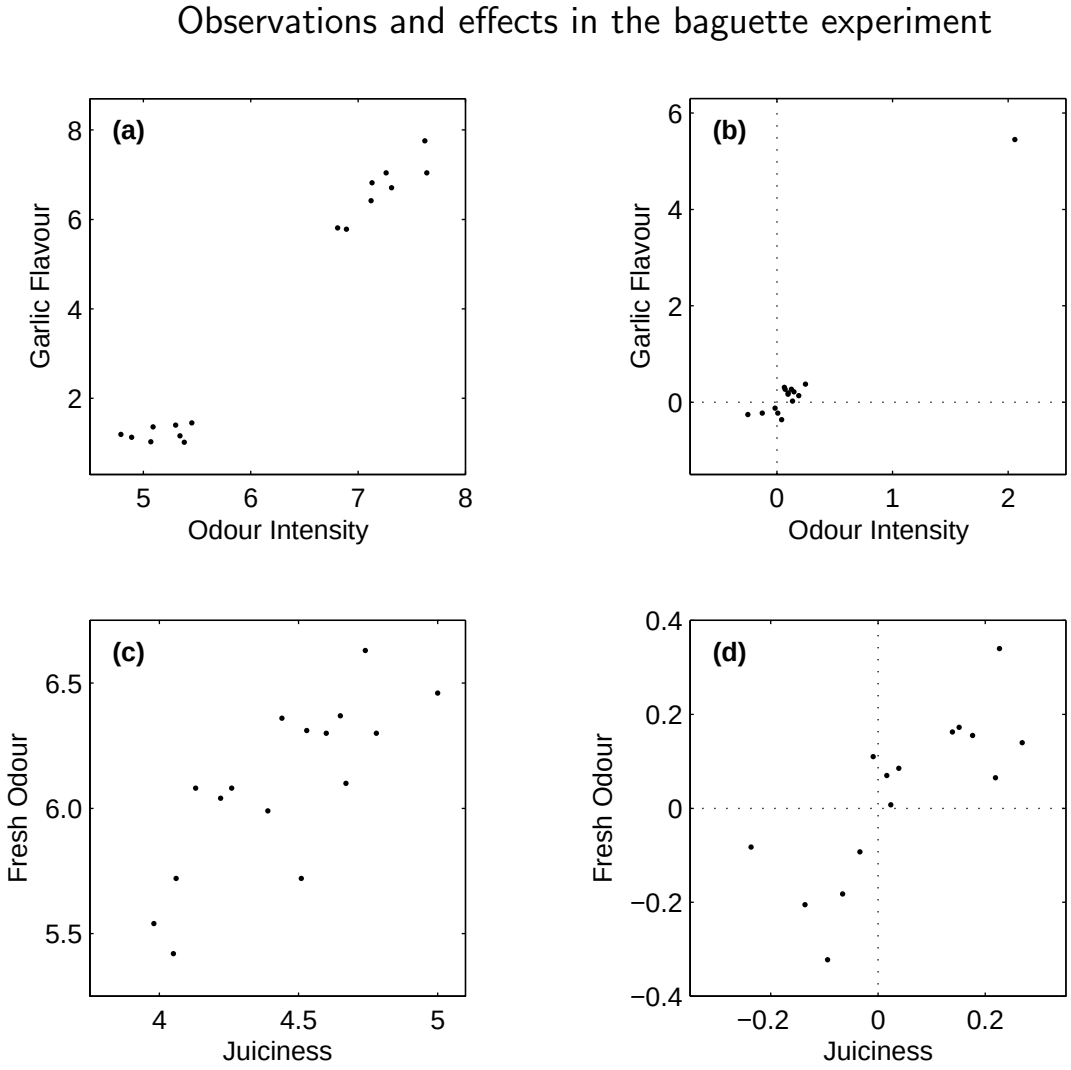

Figure 1: The left panels contain ordinary scatter plots of original observations in the baguette experiment: Garlic Flavour versus Odour Intensity (a) and Fresh Odour versus Juiciness (c). Panels (b) and (d) contain corresponding plots of estimated effects, which are calculated according to the underlying fractional factorial design.

to be different. This type of information is lost in Figure 1(b). However, in many cases, the ordinary scatter plot is less interpretable and Figure 1(b) represents a very useful alternative.

Figure 1(c) shows a scatter plot of Fresh Odour versus Juiciness, which are moderately correlated (r=0.7852). In this case, the corresponding plot (d) of estimated effects looks quite similar. There is no indication that any particular design variables are causing the correlation. Furthermore, by analyzing these responses, by e.g. normal probability plotting, there is no indication of any significant effects. Although, a positive correlation between these two sensory attributes seems reasonable.

3 PLOTTING CONTRASTS FROM RESPONSE SURFACE MODELING

We consider the sausage experiment of Ellekjær *et al.* (1994), where Fat, Salt and Starch were varied according to a $6 \times 3 \times 3$ design. Fat has six levels (8%, 12%, 16%, 20%, 24%, 28%), Salt has three levels (1.3%, 1.6%, 1.9%) and Starch has three levels (1.5%, 4.5%, 7.5%). The 54 sausages were measured by both sensory analysis and by near infrared (NIR) spectroscopy.

A suitable analysis is based on a model with the terms: Fat, Salt, Starch, Fat*Salt, Fat*Starch, Salt*Starch, Fat^2, Salt^2, Starch^2. Dependence between the regressors is avoided by using an orthogonal polynomial specification, which is equivalent to an analysis based on sequential sums of squares (Box and Draper, 1987). However, in order to compute an entire set of $n-1$ orthogonal contrasts, we fit a saturated model, which includes all the 53 terms up to Fat^5*Salt^2*Starch^2. Including the intercept term, this model specification defines a 54×54 matrix, $\boldsymbol{X}$, of (hierarchically ordered) regressors. The orthogonal polynomial specification is obtained by performing a Gram-Schmidt orthogonalisation of $\boldsymbol{X}$ (Strang, 1988). The resulting orthogonal matrix, denoted as $\boldsymbol{M}$, can be used to compute contrasts.

In general, to obtain an entire set of orthogonal contrasts, we transform the data through

$$\boldsymbol{Z} = \boldsymbol{M}^T \boldsymbol{Y}, \tag{2}$$

where the $n \times q$ matrix $\boldsymbol{Y}$ contains n observed values for each of q response variables and where $\boldsymbol{M}$ is an orthogonal $n \times n$ matrix derived from the general linear model. The first column of $\boldsymbol{M}$ represents the intercept and accordingly the first row of $\boldsymbol{Z}$ contains scaled mean values. The other $n-1$ rows of $\boldsymbol{Z}$ are our observed contrasts.

We consider the first NIR measurement (1100 nm) and the first sensory attribute (Whiteness) whose correlation is -0.7318. The ordinary scatter plot and the corresponding plot of contrasts are presented in Figure 2 (a) and (b) respectively. As above, we can apply expression (1) to compute the correlation coefficient from the contrast values. It is easy to show that the uncentered correlation (1) is invariant under orthogonal transformations. That is, $\tilde{r}(\boldsymbol{a}, \boldsymbol{b}) = \tilde{r}(\boldsymbol{Ra}, \boldsymbol{Rb})$, where $\boldsymbol{R}$ is an arbitrary orthogonal matrix. However, to

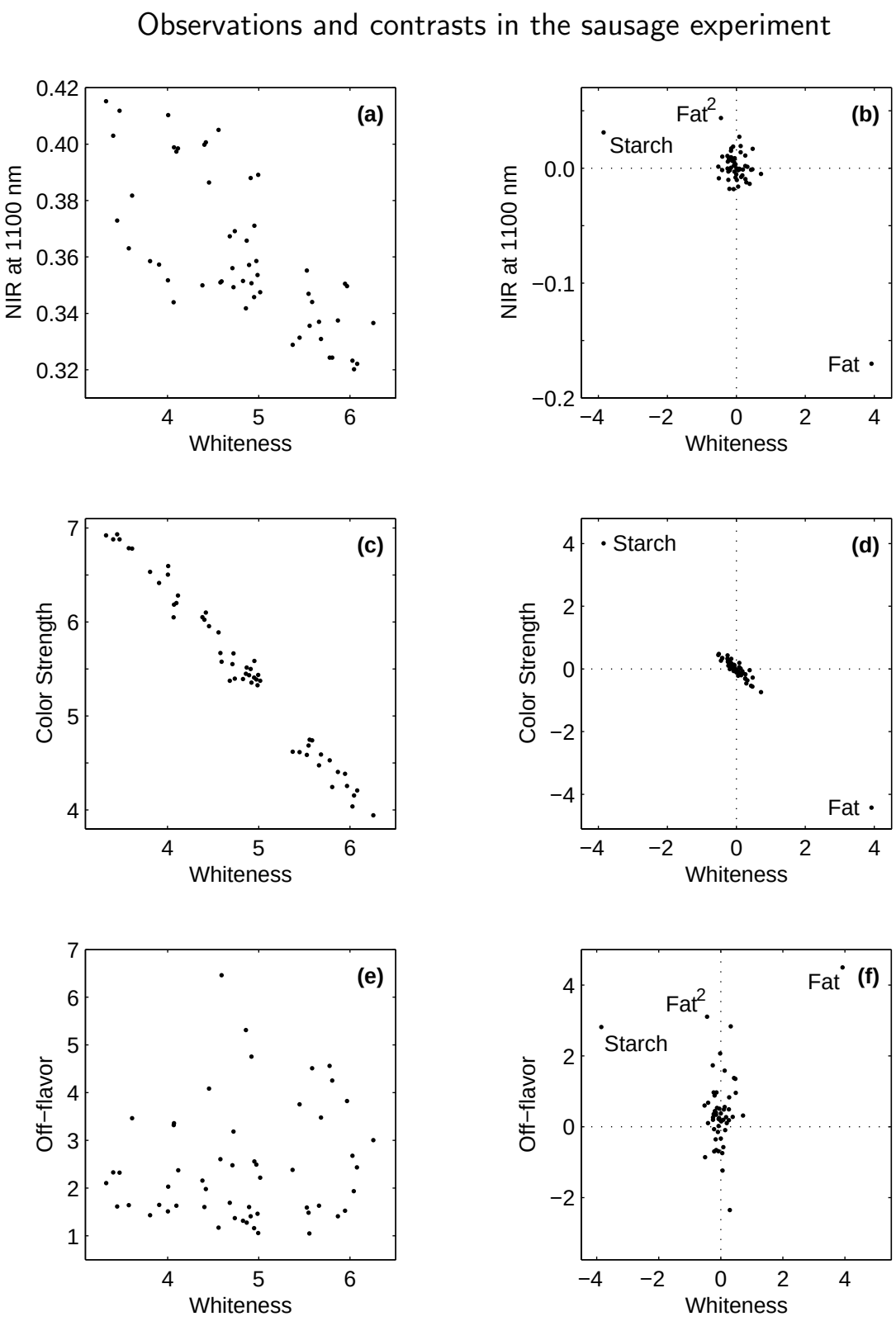

Figure 2: Panels (a), (c) and (e) contain ordinary scatter plots of original observations in the sausage experiment, where NIR at 1100 nm, Color Strength, and Off-flavor are plotted against Whiteness. Panels (b), (d) and (f) contain corresponding plots of contrasts, which are calculated according to response surface modeling.

compute contrasts from the original data we require an orthogonal matrix, $\boldsymbol{M}$, where the first column represents the intercept. Then, omitting the first row of $\boldsymbol{Z} = \boldsymbol{M}^T\boldsymbol{Y}$ before applying expression (1) corresponds to centering the data. Note that if we had centered $\boldsymbol{Y}$ before performing the transformation (2), then all entries in the first row of $\boldsymbol{Z}$ would be zero. Also note that the ordinary standard deviation of each response can be calculated directly from the contrasts by using a formula where subtracting the mean is omitted.

Figure 2(b) shows that the Fat term (first order) is the main cause of the correlation. Also the Starch term contributes importantly to the variation in the data, but that point does not support a line from the Fat point through the origin. In fact, if we exclude Starch from Figure 2(b), the correlation becomes -0.8454.

Figure 2 (c) and (d) illustrate two highly correlated (r = -0.9901) sensory responses, Color Strength and Whiteness. The contrasts for the two important regression terms (Starch and Fat) lie on a common line through the origin. The direction of this line is also supported by the other contrasts and we cannot conclude that the correlation is caused by any particular design variable. However, most of the variation in the data is explained by Starch and Fat.

Both responses in Figure 2 (e) and (f) are affected by Starch and Fat, but these terms have opposite influence on the correlation. The result is almost no correlation (r=0.1361).

4 PLS PREDICTIONS VERSUS MEASURED VALUES

With reference to the above sausage experiment, the sensory attribute, Meat Flavor, has been modeled as function of the 351 NIR wavelengths by using partial least squares (PLS) regression (Martens and Næs, 1989). Note that this is done without using any design information. By using leave-one-out cross validation, six PLS components are found to be optimal.

A common way of illustrating the accuracy of such predictions is to plot predicted values according to the cross validation against the measured values. Such a scatter plot is presented in Figure 3 (r = 0.7130) together with a corresponding plot of contrasts. The latter plot is made using exactly the same procedure as in Section 3. The Starch term is the main cause of the correlation. If we exclude the Starch point from Figure 3(b), the correlation decreases to 0.2612.

An interpretation of this result is that the Meat Flavor predictor is in reality an indirect predictor of Starch. Meat Flavor is importantly affected by Starch. To predict Meat Flavor, PLS uses Starch information that is contained in the NIR spectra. There is almost no indication of the model's ability to predict Meat Flavor beyond the relationship caused by the Starch effect.

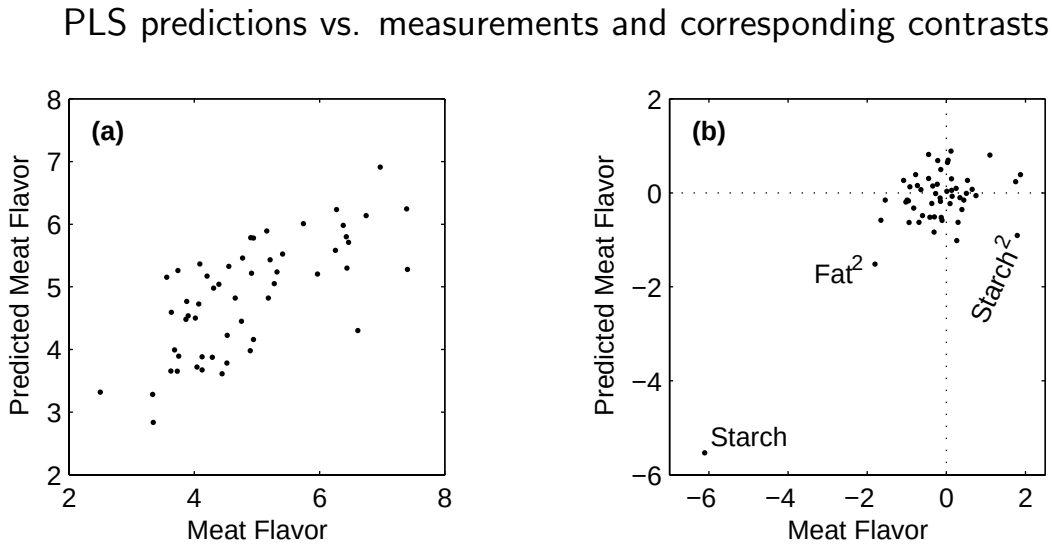

Figure 3: Meat Flavor predicted from NIR spectra versus the original Meat Flavor observations (a) together with the corresponding plot of contrasts (b).

5 PCA BASED ON CONTRASTS

Principal component analysis (PCA) can be based on the singular value decomposition (SVD) (Strang, 1988),

$$\boldsymbol{Y}_c = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T \,, \tag{3}$$

where $\boldsymbol{Y}_c$ is the centered data matrix. We assume that the diagonal matrix, $\boldsymbol{\Lambda}$, is square with only nonzero diagonal entries (the "economy size" decomposition). The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal. In the context of PCA, $\boldsymbol{V}$ is the matrix of loadings and $\boldsymbol{U}\boldsymbol{\Lambda}$ contains the scores.

The loadings are unchanged if we perform SVD of the matrix of observed contrast values. This follows from the fact that the SVD of $\boldsymbol{M}^T\boldsymbol{Y}_c$ can be written as

$$\boldsymbol{M}^T\boldsymbol{Y}_c = (\boldsymbol{M}^T\boldsymbol{U})\boldsymbol{\Lambda}\boldsymbol{V}^T \,. \tag{4}$$

Since $\boldsymbol{Y}_c$ is centered and since the first column of $\boldsymbol{M}$ represents the intercept, the entries in the first row of $\boldsymbol{M}^T\boldsymbol{Y}_c$ and $\boldsymbol{M}^T\boldsymbol{U}\boldsymbol{\Lambda}$ are zero. The other rows of $\boldsymbol{M}^T\boldsymbol{U}\boldsymbol{\Lambda}$ contain the scores according to the contrasts.

Recalling the sausage experiment, we will consider the PCA of the NIR spectra (351 variables). The score plot is presented in Figure 4 together with the corresponding plot of contrast scores. When interpreting these plots, it is important to have in mind that the axes are not scaled according to their importance.

From Figure 4(b) we can see that the linear effect of Fat is dominating the first component. Variation caused by the design is also very important for the second component.

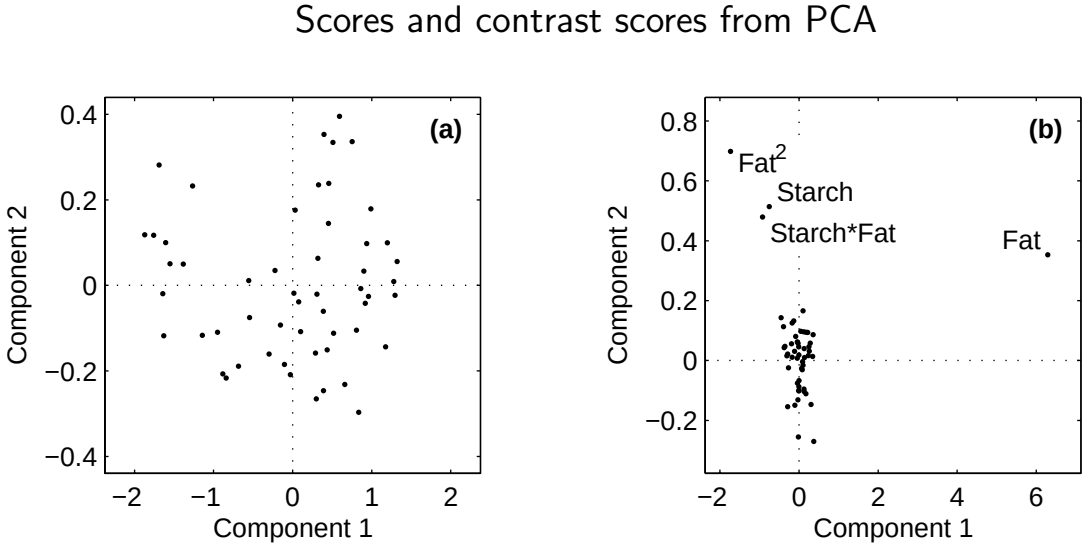

Figure 4: Scores from PCA of NIR (a) together with the corresponding plot of contrast scores (b). The first and the second component explain, respectively, 95.9% and 3.3% of the variance.

Both Fat^2, Starch and Starch*Fat are influencing this direction. If we had colored the points in Figure 4(a) according to the Fat content we would also see from that figure how the first component is dominated by Fat. However, this information is seen more directly in Figure 4(b).

The two plots represent two ways of illustrating the same data variation. Figure 4(a) shows how the single observations contribute to the variation and Figure 4(b) illustrates in a direct way whether there are systematic structures according to the response surface model.

6 CONCLUDING REMARKS

The above analysis of the sausage experiment is just one example of how to construct an entire set of orthogonal contrasts according to a general linear model. To distinguish from "a complete set of contrasts" (decomposition of a single model term) we have introduced the expression "an entire set of contrasts" when referring to a decomposition into $n-1$ contrasts. This means that we have one contrast associated with each degree of freedom (DF) in the ANOVA table. Thus, model terms with several DFs have to be decomposed as a complete set of contrasts. Furthermore, contrasts associated with the error term are also constructed. Above, the error term was decomposed by saturating the model. In other cases we may choose any arbitrary decomposition.

In unbalanced cases the problem of defining contrasts corresponds to the problem of choosing sums of squares. Since we require orthogonal contrasts, a decomposition according to sequential sums of squares is the only possible choice. One could, however, imagine "approximate" plots where the contrasts are defined according to other types of sums of squares. This problem is, however, beyond the scope of the present paper.

In this paper we have demonstrated several examples of orthogonal contrast plotting. In general, when original observations are plotted in some way, we can always make a corresponding plot of orthogonal contrasts. In such a plot, information about the single observations is lost. Instead, information about the contribution of single model terms is obtained. The contrast plot contains exactly the same information about variances, correlations and linear multivariate relations as the ordinary plot. We believe that plotting orthogonal contrasts is a useful supplement to the established tools for analyzing designed experiments.

REFERENCES

Box, G. E. P. and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, New York: John Wiley and Sons.

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

Ellekjær, M. R., Isaksson, T. and Solheim, R. (1994), "Assessment of Sensory Quality of Meat Sausages Using Near Infrared Spectroscopy", *Journal of Food Science*, **59**, 456–464.

Langsrud, Ø. (2001), "Identifying Significant Effects in Fractional Factorial Multiresponse Experiments", *Technometrics*, **43**, 415–424.

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

Strang, G. (1988), *Linear Algebra and its Applications, 3rd ed.*, San Diego: Harcourt Brace Javanovich.