<% response.buffer=true%>

GEMANOVA for complicated experimental design

Based on Bro, R, Ph.D. thesis, 1998

Background

The use of multiplicative models for analysis of variance (ANOVA) is little known. However, as the responses of a factorial design with more than one factor give rise to data that have the structure of an array it is only natural to suppose that a decomposition model may in some cases be more appropriate for modeling the data than a traditional ANOVA model. Already Fisher [1923] proposed the use of a PCA-like method for obtaining more interpretable models of two-factor experiments, and several other authors have later proposed similar approaches [Gollob 1968, Hegemann & Johnson 1976, Mandel 1969, Mandel 1971]. Kettenring [1983] has suggested the use of three-way PARAFAC for decomposing three-factor experiments. Here, a general multiplicative model is suggested partly in line with the above references, extending them to a higher level of generality. The model described was first described in Bro [1997] but also Heiser & Kroonenberg [1997] have described a similar method.

It is often stated that higher-order interactions are hard to interpret in ANOVA, and the reason is simple. Given a set of experimental data where two factors are varied on I and J levels respectively, the responses of one variable can be represented by an I x J matrix, the ijth element, y_{ij}, representing the response when the first factor is on the ith level and the second factor is on the jth level. If no interaction is present the standard ANOVA model for qualitative factors [Montgomery 1991] is

$$y_{ij} = \mu + a_i + b_j + e_{ij}. \qquad (1)$$

Here μ is the grand level, a_i is the effect of the first factor at level i, b_j the effect of the second factor at level j and e_{ij} the residual. This model of the data is a simplification. Instead of IJ elements in the original data array, only $1+I+J$ terms of which $I+J-1$ are independent has to be understood and interpreted. If interaction between the two factors is present the model is

$$y_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}. \qquad (2)$$

This model consists of $1+I+J+IJ$ parameters, of which IJ are independent. Clearly, no reduction in the complexity of the representation has been achieved, and therefore the interaction is hard to interpret. For third- and higher-order interactions the problem is even worse.

Another model is normally used for experiments with quantitative or continuous factors. For quantitative factors a linear effect of the factor settings over all levels is estimated. Mathematically the corresponding model underlying this approach is

$$y_{ij} = b_0 + b_1 x_{1i} + b_2 x_{2j} + b_{12} x_{1i} x_{2j} + e_{ij}. \qquad (3)$$

Here x_{1i} refer to the value of the first factor at the ith level etc. These values are fixed as they are determined by the experimental design, hence only the parameters b have to be estimated. For the first main effect only the scalar b_1 must be estimated and so on.

For models involving both quantitative and qualitative factors the two models are easily merged. The qualitative ANOVA model is quite flexible, and thus quite prone to overfit, especially for interactions. The quantitative model on the other hand is very restricted in its model formulation. The drawback of both models is that, even though the models can theoretically handle interactions of any order and complexity, data that are mainly generated by interactions can be difficult to model adequately. The model proposed here can be seen as a complement of intermediate complexity.

The GEneral Multiplicative ANOVA (GEMANOVA) model is defined for a two-factor experiment as

$$y_{ij} = \mu + a_{i1} + b_{j1} + \sum_{f=1}^{F} c_{if} d_{jf} + e_{ij} . \quad (4)$$

For a three-factor experiment a full model would contain main effects, two-way interactions as well as three-way interactions. As for the standard ANOVA model only significant terms should be included in the model. It may be noted that the model contains the qualitative ANOVA model as a limiting case. If for example the following ANOVA model is found for a two-factor experiment

$$y_{ij} = (ab)_{ij} \quad (5)$$

the exact same model can be fitted by a full rank bilinear model as

$$y_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} \quad (6)$$

where F equals the rank of the matrix **Y**, with typical elements y_{ij}. The GEMANOVA model also contains more specialized models as the shifted multiplicative model (Cornelius et al. 1992, van Eeuwijk 1996) and thus theoretically gives a unifying algorithm for fitting these models. An important distinction between the GEMANOVA model and other suggested multiplicative models is that the GEMANOVA model is a least squares model. This cannot in general be guaranteed for earlier proposed multiplicative ANOVA models as evidenced in Kruskal (1977b). Any GEMANOVA model can be fitted with a PARAFAC algorithm with certain elements fixed to ones. This points to the other significant advantages of the GEMANOVA model. It generalizes to any number of factors, and due to its close relation to the PARAFAC model many GEMANOVA models are identified by the structural model alone.

Consider a three-factor experiment with response y_{ijk} for factor one at level x_i, factor two at level x_j, and factor three at level x_k. A main effect for factor one will be modeled as

a_i	ANOVA	qualitative
bx_i	ANOVA	quantitative
a_i	GEMANOVA	

A three-way interaction will be modeled as

$(abc)_{ijk}$	ANOVA	qualitative
$bx_i x_j x_k$	ANOVA	quantitative
$a_i b_j c_k$	GEMANOVA	one-component
$\sum_{f=1}^{F} a_{if} b_{jf} c_{kf}$	GEMANOVA	F-component

The GEMANOVA model obviously has stronger similarity to the qualitative ANOVA model than the quantitative model being parametrically equivalent for main effects. However, for interactions, the GEMANOVA model fills a gap between the very flexible qualitative model B $(abc)_{ijk}$ using a total of IJK parameters B and the very restricted quantitative model B $bx_i x_j x_k$ using one parameter. A one-component three-way GEMANOVA effect is $a_i b_j c_k$ using $I+J+K$ parameters. When used to model experiments with qualitative factors the possible advantage is similar to the gain in insight and robustness obtained by using PCA for exploring two-way data. The qualitative ANOVA interaction terms can often be expected to overfit as no structure is imposed at all between levels and factors. For quantitative experiments the GEMANOVA model may also be advantageous especially if many experiments are performed. The GEMANOVA model is intrinsically more flexible than the quantitative ANOVA model, i.e., compare the main effect for both models. If the variation in response is primarily caused by simple main effects of quantitative factors the quantitative ANOVA model is most likely to provide adequate answers, but if several interactions or cross-products are present, it is quite possible that they can be better modeled by a multiplicative model.

How then, to choose between the different possible modeling strategies? The most sensible way to start is to use the standard ANOVA model. *If* there are indications that the model is mainly governed by

interactions or complicated cross-products or *a priori* knowledge suggests a multiplicative model would be more appropriate, it is possible that the GEMANOVA model can be a fruitful alternative.

Several possibilities exist for validating a GEMANOVA model, i.e., determining the complexity of the model. F-like tests, resampling techniques like bootstrapping and cross-validation, or technological and methodological insight may be used. The details of these approaches will not be touched upon here. The only important point that should be made, is that degrees of freedom are not available for estimating variances in PARAFAC/GEMANOVA models. To circumvent this problem approximate estimated degrees of freedom can be obtained from Monte Carlo studies as described for a similar problem by Mandel (1969 & 1971).

ASPECTS OF GEMANOVA

Fractional factorial designs. Fractional designs correspond to full factorial designs with missing elements and are easily handled as such.

Multiple responses. Throughout it has been assumed that there is only one response variable, but the model also holds for multiple responses. If the responses are not correlated it is of course possible to analyze the responses separately, but it is likely that the model can be stabilized by decomposing several responses simultaneously. This can be accomplished by introducing an extra mode called the response mode with its *m*th level being the *m*th response. Appropriate scaling may be necessary for handling differences in scale of different responses.

Heteroscedasticity. If data are heteroscedastic this may be dealt with by using scaling or a weighted loss function.

Application of GEMANOVA to enzymatic activity

Problem

A major contributor to undesirable browning effects in fruit and vegetables is the enzymatic browning caused by PPO, polyphenol oxidase (Martinez & Whitaker 1995). Enzymatic browning can henceforth be expressed by the dioxygen consumption of PPO, as this is directly related to the activity of PPO. To avoid enzymatic browning of fruits and vegetables it is important to store them under conditions that suppress the enzymatic activity.

The activity of PPO was investigated as a function of different levels of O_2, CO_2, temperature, pH, and substrate according to a factorial design. A traditional ANOVA model of the data did not shed much light on the relation between these factors and the activity, while a multiplicative model based on GEMANOVA/PARAFAC was easy to interpret. The problem presented here is part of an investigation conducted by Hanne Heimdal and described in detail in [Bro & Heimdal 1996, Heimdal et al. 1997].

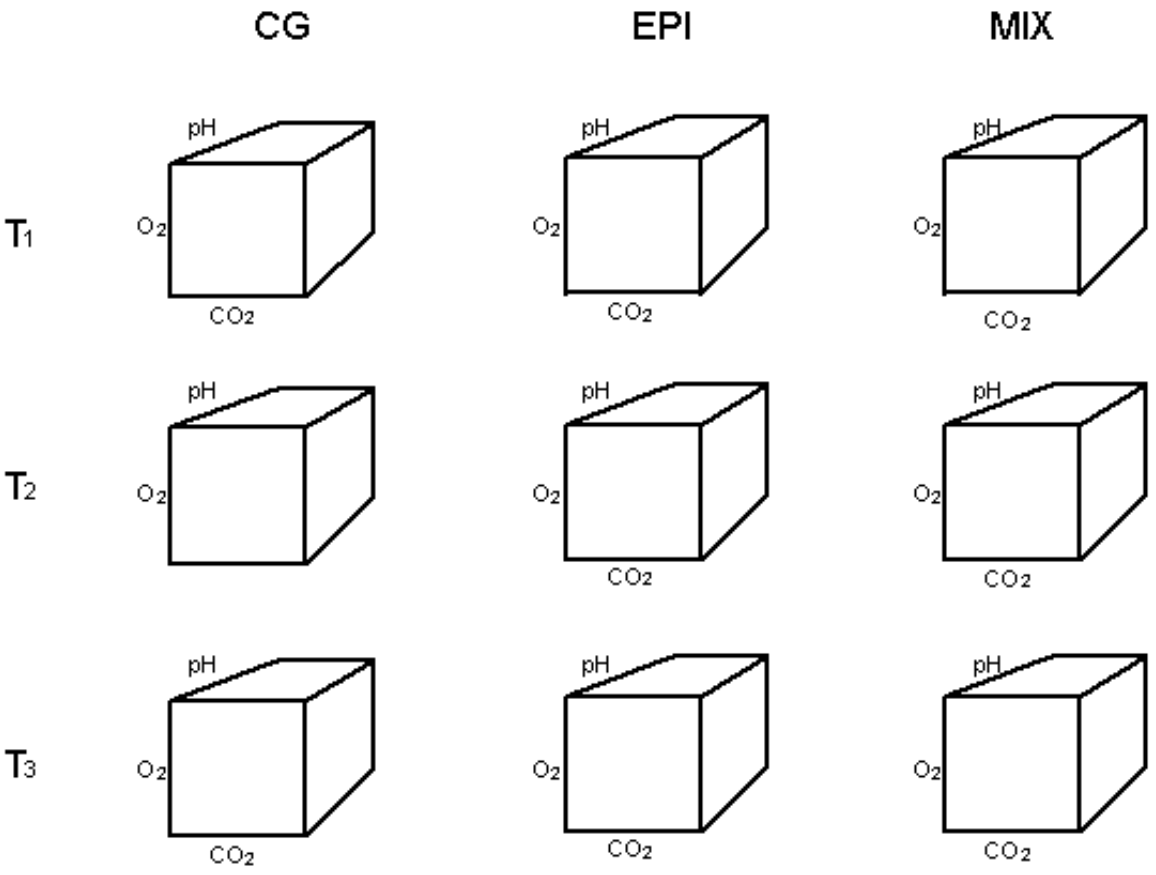
Data

The PPO used was obtained from fresh iceberg lettuce and extracted and purified according to Heimdal et al. (1997). PPO activity was measured in nanomoles of O_2 consumed per second by a polarographic polyphenol oxidase assay. For five O_2 levels, three CO_2 levels, three substrate types, three pH values and three different temperatures B all varied independently B the activity of PPO was determined in replicate. The substrates used were 0.01 M chlorogenic acid (CG), 0.01 M epicatechin (EPI) and an equimolar mixture of both (MIX), where MIX = 0.005M CG + 0.005M EPI, hence the substrate concentration is also 0.01 M.

Table 1. Experimental design for the experiment.

Factor	Code	Levels
O_2 /%	*O*	0, 5, 10, 20, 80
CO_2 /%	*C*	0, 10, 20
Substrate	*S*	CG, EPI, MIX
pH	*P*	3.0, 4.5, 6.0
Temperature / °C	*T*	5, 20, 30

Building a calibration model to predict the activity from the experimental conditions can give important information on how the PPO activity - and therefore the color formation - is influenced by the different factors. The different levels of the factors are shown in Table 9. The number of samples in the replicated full factorial design is 5 x 3 x 3 x 3 x 3 x 2 = 810. The data constitute a full five-factor factorial design, but in the PARAFAC/GEMANOVA model, the data are interpreted as a multi-way array of activities, specifically a five-way array. The five different modes are: O_2 (dimension five), CO_2 (dimension three), pH (dimension three), temperature (dimension three), and substrate type (dimension three). The *ijklm*th element of the five-way array contains the activity at the *i*th O_2 level, the *j*th CO_2 level, the *k*th level of pH, the *l*th level of temperature, for the *m*th substrate type. The five-way array is depicted in below.



A graphical representation of the five-way array of enzymatic activities.

RESULTS

To choose the model, i.e., the number of components, GEMANOVA models (page 192) were fitted using half the data as calibration set and half as a test set. The predictions of activity from each model given by the loadings **O**, **C**, **S**, **P** and **T** were compared to the test set activities. The number of components, *F*, was chosen to minimize the predicted residual sum of squares, PRESS, calculated as

$$\mathrm{PRESS} = \left[\left(\sum_{f=1}^{F} o_{if} c_{jf} s_{kf} p_{lf} t_{mf} \right) - act_{ijklm} \right]^2 \qquad (7)$$

act_{ijklm} being the *ijklm*th element/activity of the replicate set not used to build the model. One component with no fixed elements gave the lowest prediction error, which furthermore was in the neighborhood of the intrinsic error of the reference value. The resulting GEMANOVA model is thus very simple, namely a one-component PARAFAC model, which is equivalent to a five-way multiplicative ANOVA model

$$act_{ijklm} = o_i c_j s_k p_l t_m + e_{ijklm} \qquad (8)$$

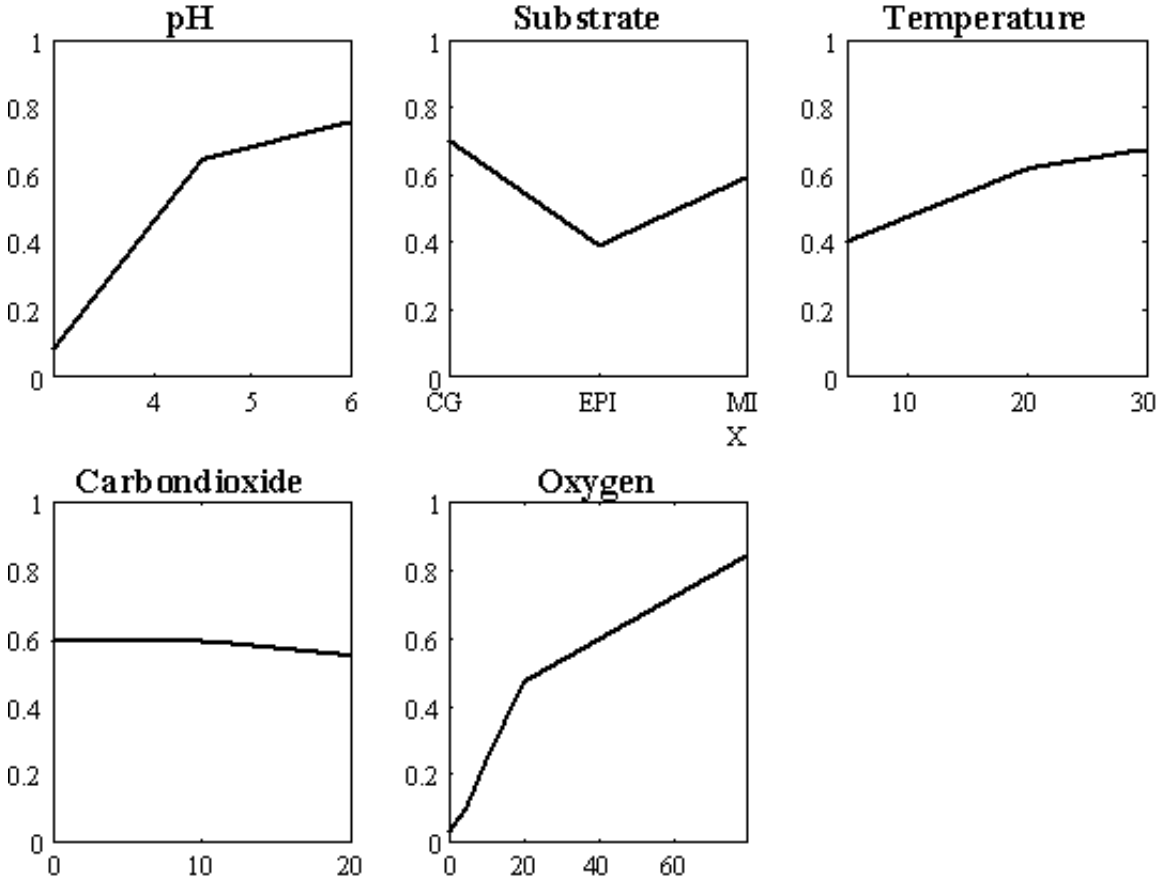
as compared for example to a traditional ANOVA model which after pruning of insignificant terms reads

$$\log(act_{ijklm}) =$$

$$b_0 + b_1 x^{o,i} + b_2 x^{s,k} + b_3 x^{p,l} + b_4 x^{t,m} + b_5 (x^{o,i})^2 + b_6 (x^{p,l})^2 + b_7 x^{o,i} x^{p,l} + b_8 x^{s,k} x^{t,m} + e_{ijklm} \qquad (9)$$

where $x^{o,i}$ means the *i*th setting of oxygen (scaled appropriately). The PARAFAC model is given by the five loading vectors depicted in Figure 37. The loadings are interpreted in the following way. For given settings of the factors simply read the corresponding five loading elements on the respective plots and multiply these five numbers. The product is the estimated PPO activity.

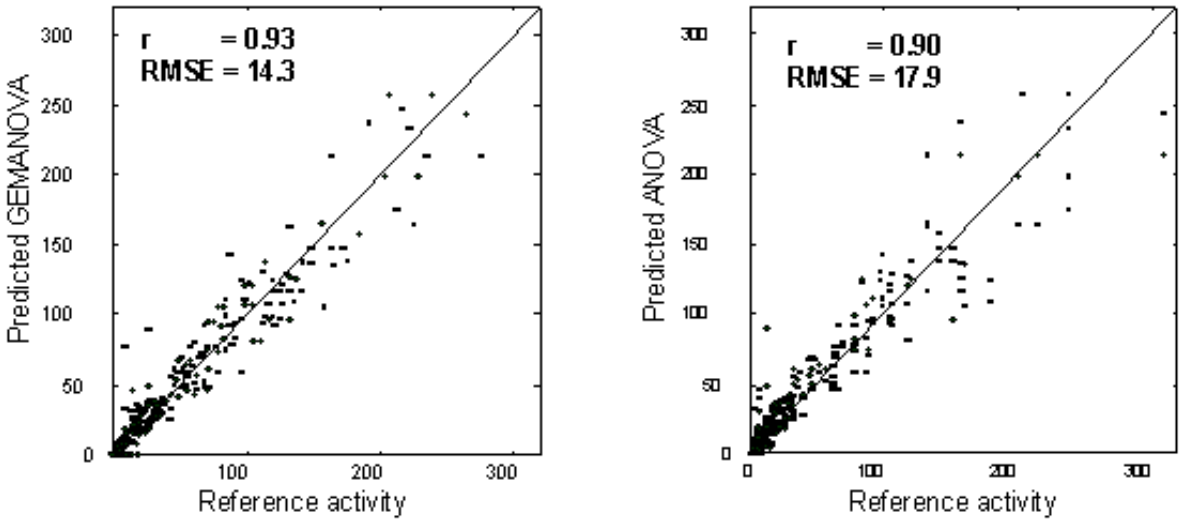
To determine how to obtain low PPO activity, hence low enzymatic browning, it is clear from the figures, that the setting for each factor should be the one with the lowest accompanying loading. As the model is multiplicative this will be the setting yielding the lowest enzymatic activity. The main conclusion derived from the model is therefore that by keeping temperature, oxygen, and pH as low as technologically possible the enzymatic browning will be minimized. The effect of CO_2 is small, but ignoring it leads to a model with significantly poorer predictability. The consequence of this is elaborated on in Heimdal et al. (1997).



Loading vectors of the one-component/one effect GEMANOVA model of the enzymatic activity data.

The multiplicative model could also have been obtained by traditional ANOVA by using the logarithm of activity *and* modeling all factors as qualitative variables. However, even though this would give the same structural model, the loss function would not be the same, and the corresponding model would be poorer with respect to predicting activity. Specifically the root mean squared error of predicting one replicate set from a model built from the other replicate set is 18.0 using this approach whereas it is only 14.3 using

PARAFAC/GEMANOVA. A traditional ANOVA model based on a logarithmic transform of activity and treating all factors except substrate as quantitative factors is also possible. Though such a model give almost as good predictions as the GEMANOVA model, it contains more effects, and is somewhat more difficult to interpret. To compare the two different models both models were used to predict the activities of the test set samples. The resulting predictions are shown in the plot below.



Predictions of independent test set obtained from GEMANOVA (left) and ANOVA (right).

Reference List

1. Bro R, PARAFAC. Tutorial and applications, *Chemom Intell Lab Syst,* 1997, **38**, 149-171.

2. Bro R, Heimdal H, Enzymatic browning of vegetables. Calibration and analysis of variance by multiway methods, *Chemom Intell Lab Syst,* 1996, **34**, 85-102.

3. Fisher RA, MacKenzie WA, Studies in crop variation. II The manurial response of different potato variaties, *J Agri Sci,* 1923, **13**, 311-320.

4. Gollob HF, A statistical model which combines features of factor analytic and analysis of variance techniques, *Psychometrika,* 1968, **33**, 73-115.

5. Hegemann V, Johnson DE, On analyzing two-way anova data with interaction, *Technometrics,* 1976, **18**, 273-281.

6. Heimdal H, Bro R, Larsen LM, Poll L, Prediction of polyphenol oxidase activity in model solutions containing various combinations of chlorogenic acid, (-)-epicatechin, O2, CO2, temperature and pH by multiway analysis, *J Agric Food Chem,* 1997, **45** , 2399-2406.

7. Heiser WJ, Kroonenberg PM, Dimensionwise fitting in PARAFAC-CANDECOMP with missing data and constrained parameters, 1997,

8. Kettenring JR, A case study in data analysis, *Proceedings of Symposia in Applied Mathematics,* 1983, **28**, 105-139.

9. Mandel J, The partitioning of interaction in analysis of variance, *Journal of Research of the National Bureau of Standards B Mathematical Sciences,* 1969, **73B**, 309-328.

10. Mandel J, A new analysis of variance model for non-additive data, *Technometrics,* 1971, **13**, 1-18.

11. Montgomery DC, *Design and analysis of experiments*. John Wiley & Sons, New York, 1991,