Comparison of GEMANOVA and ANOVA model on the basis of principle of parsimony

Jian X. Wu[1], Frans van den Berg[2], Jukka Rantanen[1]

1. Department of Pharmaceutics and Analytical Chemistry, Faculty of Health and Medical Sciences, University of Copenhagen
2. Quality and Technology, Department of Food Science, University of Copenhagen

Abstracts

Present study compares the GEMANOVA with the ANOVA model on modeling complex data set from Design of Experiment (DoE) on the basis of model simplicity and ease of model interpretation. The study suggests that the GEMANOVA model, due to its multiplicative model structure, is easier to understand and interpret in contrast to the ANOVA model with many significant higher order interaction terms. The principle of parsimony states that upon others being equal, a simpler model is preferred over a more complex model. Since both GEMANOVA and ANOVA model showed equal predictability, the study concluded that the GEMANOVA model is the most parsimonious model, and hence can be a valuable tool when developing Quality by Design (QbD) design space.

Keywords: multiway-analysis, GEMANOVA, ANOVA, design of experiments, QbD

1. Introduction

Establishment of a design space, where critical formulation and process parameters are linked to critical quality attributes has become widely recognized as an important tool for building the quality into the product. Often the key critical parameters are identified in the risk assessment phase, and Design of Experiment (DoE) together with analysis of variance (ANOVA) are used in order to mathematically describing the relation between critical factors and responses. Though ANOVA has proved to be useful in describing how factors are related to responses, often the built traditional ANOVA model can be difficult to interpret due to the presence of higher order interaction terms (1). In the ICH Q8

guideline, the importance for the data analyst to understand the design space model has been emphasized (2).

A different model that is capable for analyzing DoE data is the generalized multiplicative analysis of variance (GEMANOVA) model. One of the differences between ANOVA and GEMANOVA models are that ANOVA model is an additive model, and the model starts with main factor terms followed by addition of more complex higher order interaction terms in order to fully explain the underlying data set. In contrast, GEMANOVA due to its multiplicative nature focuses on the higher order interaction term to start with, and additional components are added if a single component is insufficient in fully explaining the underlying data set. The basic model structure for a multilevel four factors one component GEMANOVA model is exemplified in equation 1:

$$y_{ijkl} = a_i b_j c_k d_l + e_{ijkl} \quad (1)$$

where y_{ijkl} is the response element obtained when factor *a*, *b*, *c* and *d* are varied at level *i*, *j*, *k*, and *l* and e_{ijkl} is the residual element respectively. The simplicity of the basic GEMANOVA model structure in this example means that all response elements can be reproduced by multiplication of indices from three sets of loading vectors representing factor *a*, *b*, *c* and *d*. For an in-depth understanding of GEMANOVA model, the reader is referred to the work by Bro (3).

In many cases, it has been shown that same model predictability can be obtained when ANOVA and GEMANOVA models were built on the same data set (1, 4, 5). In such cases, the question often arises is which model should be chosen as the final model for modeling design space? In guiding model selection, the principle of parsimony (also known as Ockham's razor) can be a valuable tool (6). In the context of model selection, the parsimony principle can be interpreted as when others being equal, a simpler model is more preferable over a more complex model.

In the present study, a DoE data set from solid dispersion development (7) is subjected to ANOVA and GEMANOVA modeling. The ANOVA and GEMANOVA models are subsequently compared in terms of ease of interpretability and model simplicity.

2. Materials and Methods

2.1 Data set

Solvent evaporation is one of the main methods for preparing solid dispersions. A previous study highlighted that the evaporation rate of solvent is determinant for drug physical stability upon storage (7). Beside from solvent evaporation rate, other studies have highlighted that other factors such as drug:polymer ratio and polymer molecular weight have significant importance for the physical stability of the drug (8, 9). The degree of crystallinity in solid dispersions can be quantified from polarized light micrograph using image analysis (7, 10), by calculating the percentage area coverage (PAC) according to equation 2:

$$PAC = \frac{A_{crystalline}}{A_{image}} \times 100 \qquad (2)$$

where $A_{crystalline}$ and A_{image} are the area of identified crystalline region and total image area respectively.

The data set originates from a full factorial design with four factors and two replicates at the corner points. The factors polymer:drug ratio, solvent evaporation temperature and polymer molecular weight are varied at two levels. The factors mentioned are related to preparation of solid dispersions, and the formulations are monitored under polarized light micrographs at day 1, 15 and 30 after their preparation see Table 1.

Table 1: Summary of the factors and the levels of variation for DoE. Solvent evaporation temperature (Temp.), polymer molecular weight (P_{mw}).

Factor	Levels
Polymer:drug	1:1 and 3:1
Temp.	30 and 50 °C
P_{mw}	30000 and 1000000
Day	day 1, 15 and 30

2.2 Model development

ANOVA model was built using MODDE (ver. 9.0, Umetrics, Sweden). GEMANOVA model was built using Matlab (ver. 7.10, MathWorks, U.S.) and the PLS_Toolbox support (ver. 5.8 Eigenvector Research, USA).

The data set for the GEMANOVA model is arranged as 5-way array, where the first, second, third, fourth and fifth mode hold the replicate, polymer:drug ratio, solvent evaporation temperature, polymer molecular weight and storage days respectively. A constant constraint is applied on the first mode, since it was assumed that no significant effect exists in the replicate mode. The GEMANOVA model was built using the parallel factorial analysis (PARAFAC) algorithm. The number of significant components for the model is determined using a leave-one-sample out internal cross-validation approach by calculating the root mean square error of cross validation (RMSECV) (11).

The predictability of both models is determined by calculating the root mean square error of prediction (RMSEP) of center points in the DoE, which has been excluded from the modeling.

3. Results and discussion

3.1 The ANOVA model

The ANOVA model suggests that all main factors are significant. Increasing the storage time (day) has the effect of increasing the degree of the crystallinity in the sample, while increase in polymer:drug ratio, solvent evaporation temperature, and polymer molecular weight have the opposite effect (Figure 1). In the built ANOVA model, third order interaction terms were found to be significant. Though the effect of varying the main factors on the degree of crystallinity in the formulations can easily be understood, interpretation of the higher order interaction terms can be difficult

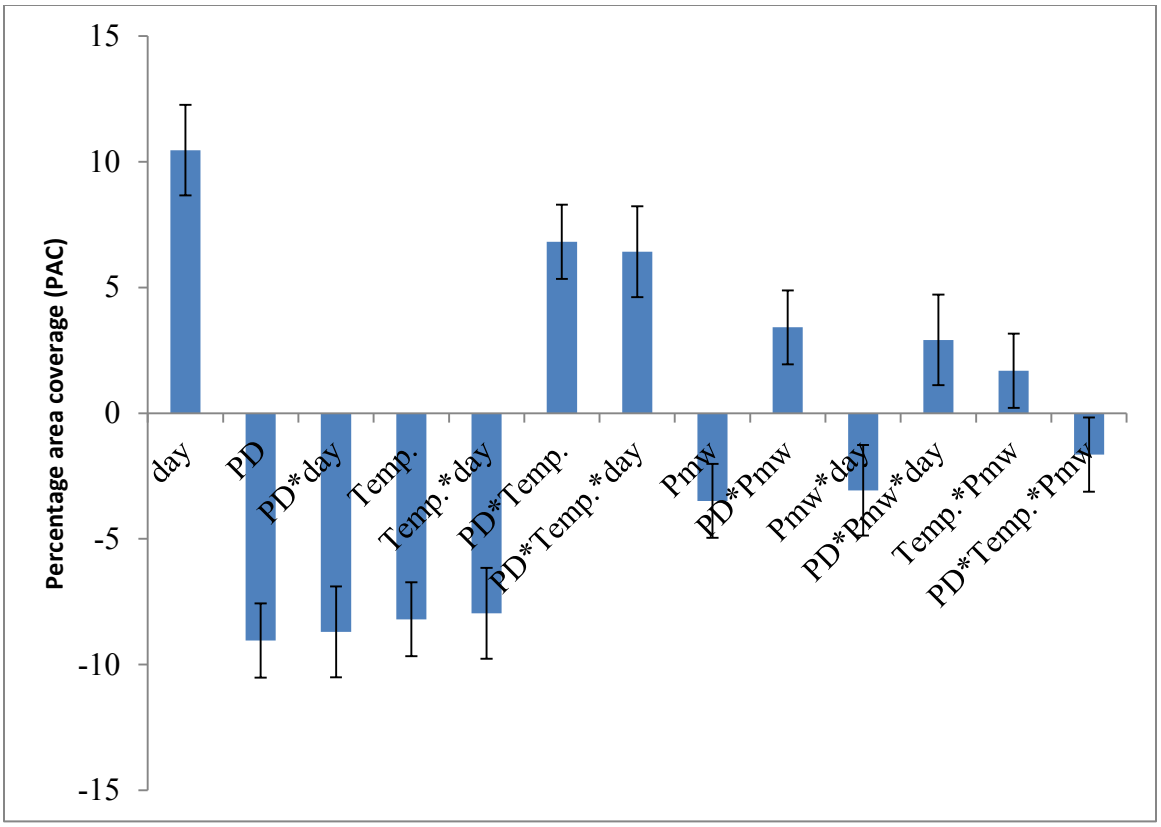

Figure 1: Effect plot from the ANOVA model illustrating the effect of factor increase on the degree of crystalline percentage area coverage (PAC) in the samples. Polymer:drug ratio (PD), solvent evaporation temperature (Temp.), polymer molecular weight (Pmw).

3.2 The GEMANOVA model

Internal cross validation revealed that the one component GEMANOVA model yields the lowest RMSECV (data not shown), hence the basic model structure is the same as in equation 1. The loading plot (Figure 2) illustrates the effect of factors when varied from low to high level. The decrease in loading plots upon increase in polymer:drug ratio, solvent evaporation temperature and polymer molecular weight all suggest that the sample crystallinity will decrease (PAC will decrease) upon increase in the mentioned factors. The day loading plot showed the opposite trend. From the discussion above, it can be inferred that the conclusion from the GEMANOVA model is essentially the same as that for the ANOVA model with regard on the effect of variation of main factors on PAC. However, the GEMANOVA model is much simpler to understand as compared to

the ANOVA model because the model structure focuses on the effect of factor interaction to start with.

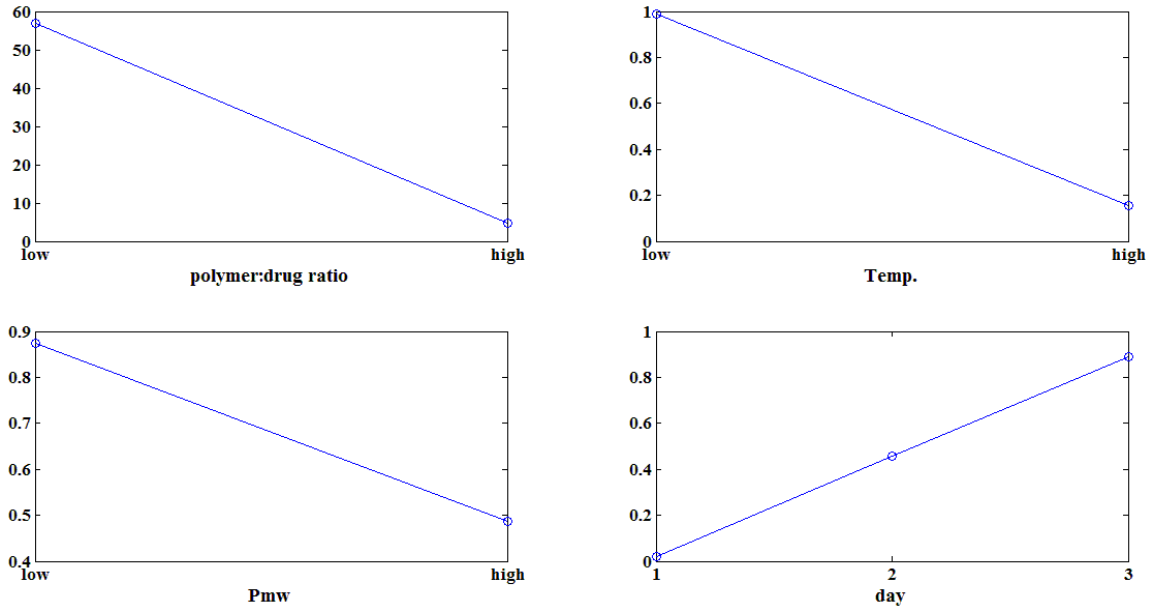

Figure 2: Loadings from GEMANOVA model. Solvent evaporation temperature (Temp.), polymer molecular weight (Pmw).

3.3 Model selection using the parsimony principle

The RMSEP for ANOVA and GEMANOVA has calculated to 3.3 and 4.2 % respectively, and hence the two models are considered having equal predictability. The principle of parsimony then states that a simpler model is more preferable as compared to a more complex model. From a mathematical point of view, a simpler model is regarded as a model using fewer parameters. Since GEMANOVA model was based on 11 against 14 parameters for the ANOVA model, it can be concluded that the GEMANOVA model is simpler from a mathematical and understanding point of view as compared to the ANOVA model.

4. Conclusion

GEMANOVA model when applied on the present DoE data set offers easier model interpretation and led to less complex model as compared to the ANOVA model. The multiplicative GEMANOVA model structure is of particular advantage when modeling complex DoE data set with higher order interaction terms present. The ease of model

interpretation and the simplicity of the GEMANOVA model make it a good candidate for establishing multivariate QbD design space.

Acknowledgement

Jian X. Wu gratefully acknowledges Professor Rasmus Bro for many enlightening discussions on different facets of multiway-analysis. Funding from the Danish Council of Technology and Innovation for the Innovation Consortium NanoMorph (952320/2009) is acknowledged.

References

1. A. Smilde, R. Bro, and P. Geladi. Multi-way Analysis: Applications in the Chemical Sciences, John Wiley & Sons Ltd., Chichester, England, 2004.
2. ICH Harmonised Tripartite Guideline: Pharmaceutical Development Q8 (R2). In I.C.o.H. 2008 (ed.), *International Conference of Harmonisation 2008*, Vol. 71, Federal Register, 2009.
3. B. Rasmus. PARAFAC. Tutorial and applications. Chemom Intell Lab Syst. 38:149-171 (1997).
4. K. Naelapää, M. Allesø, H.G. Kristensen, R. Bro, J. Rantanen, and P. Bertelsen. Increasing process understanding by analyzing complex interactions in experimental data. J Pharm Sci. 98:1852-1861 (2009).
5. K.H. Lilandand E.M. Færgestad. Testing effects of experimental design factors using multi-way analysis. Chemom Intell Lab Syst. 96:172-181 (2009).
6. M.B. Seasholtzand B. Kowalski. The parsimony principle applied to multivariate calibration. Anal Chim Acta. 277:165-177 (1993).
7. J.X. Wu, M. Yang, F.v.d. Berg, J. Pajander, T. Rades, and J. Rantanen. Influence of solvent evaporation rate and formulation factors on solid dispersion physical stability. Eur J Pharm Sci. 44:610-620 (2011).
8. V. Tantishaiyakul, N. Kaewnopparat, and S. Ingkatawornwong. Properties of solid dispersions of piroxicam in polyvinylpyrrolidone K-30. Int J Pharm. 143:59-66 (1996).
9. V. Tantishaiyakul, N. Kaewnopparat, and S. Ingkatawornwong. Properties of solid dispersions of piroxicam in polyvinylpyrrolidone. Int J Pharm. 181:143-151 (1999).

10. D. Xia, J.X. Wu, F. Cui, H. Qu, T. Rades, J. Rantanen, and M. Yang. Solvent-mediated amorphous-to-crystalline transformation of nitrendipine in amorphous particle suspensions containing polymers. Eur J Pharm Sci. (in press)
11. R. Broand M. Jakobsen. Exploring complex interactions in designed data using GEMANOVA. Color changes in fresh beef during storage. J Chemom. 16:294-304 (2002).